Cómo atienden los transformers

Perfiles medidos, modelos condicionales de ley de potencia y auditoría de la evidencia

Carles Marín

Investigador independiente

transformerkmarin@gmail.com

19 de noviembre de 2026 – resultados matemáticos verificados integrados

Edición corregida: resultados formales integrados, 19 de noviembre de 2026

Los errores algebraicos, numéricos, de descripción de implementación y de referencias cruzadas identificados en la auditoría se han corregido en esta edición. Las identificaciones físicas no respaldadas y las garantías de rendimiento han sido retiradas. El texto, tablas y código del controlador obsoletos permanecen en el archivo fuente separado; no se reproducen aquí como un conjunto alternativo de conclusiones. Anexo J registra las resoluciones. Una nueva implementación de referencia corrige la causalidad del prefijo, la ubicación de la proyección de cabeza, el emparejamiento rotatorio y la interfaz entrada-contexto, con comprobaciones sintéticas descritas en la Sección 19.3. Estas comprobaciones no reemplazan los resultados del modelo archivado ni establecen la eficacia del modelo entrenado. Las observaciones históricas inaccesibles se identifican como evidencia faltante; no se han inventado valores para completarlas. El apéndice K agrega ocho grupos de resultados matemáticos junto con sus pruebas verificadas en Isabelle/HOL y hipótesis explícitas.

Resumen

Un perfil medido y una familia matemática condicional. Esta edición corregida estudia perfiles atención-distancia, la familia de potencias finitas $p_\gamma(d) = d^{-\gamma}/Z_N(\gamma)$ y las aplicaciones originalmente propuestas dentro del marco de Atención Termodinámica (TAF). Preserva el tema original y la evidencia detallada, corrigiendo la inferencia desde la geometría rotativa a un exponente entrenado universal. RoPE proporciona una descomposición exacta del puntaje seno–coseno. No impone, sin suposiciones adicionales sobre las consultas y claves aprendidas, una expectativa de distancia logarítmica o el principio de máxima entropía. La fórmula propuesta de configuración $\gamma_P = (2\theta - T\sqrt{2})/(2\theta + T\sqrt{2})$ sigue siendo un candidato algebraico; su derivación por fase estacionaria se retira porque una relación logarítmica esencial es igual -1 en lugar del exponente propuesto.

Dependencia instrumental. El procedimiento inicial E4 cambia la longitud de secuencia con la distancia. Su extractor real devuelve $\widehat{\gamma} = 0.782582$ con $R^2 = 0.988452$ en atención causal uniforme sobre la cuadrícula de siete puntos 30–2000. Por lo tanto, un buen ajuste de ley de potencia no identifica una decaimiento aprendido ni valida

al candidato Padé. En contraste, el estimador QL de octubre auditado utiliza posiciones de consulta fijas y tiene pendiente nula uniforme en su rango ajustado. Estas son comprobaciones de instrumentos sintéticas, no mediciones nuevas del modelo. Los protocolos iniciales y de octubre permanecen distintos.

Controles de coordenadas y campos. Las máscaras B3 disponibles eliminan conjuntos desiguales de coordenadas adyacentes, no planos rotatorios complementarios. El código de amplificación de la GPU también modifica V. El estimador de campo separado utiliza $d + 60$ tokens y proporciona una pendiente nula uniforme 0.755252, cercana a la pendiente inicial del punto de control 0.757234. Los efectos archivados, por lo tanto, no establecen un óptimo geométrico aprendido, un reservorio semántico o una interpretación funcional de la relación espectral.

Registro empírico y su actualización. Los paneles del modelo original, los contrastes de corpus, las ablaciones y los resultados negativos se conservan con su estado histórico. La estadística de precisión en casos seleccionados se retira porque no se pudo reconstruir su pertenencia a fila y el registro de selección (Sección 5.4). Una reproducción con datos guardados reconstruye los ajustes de octubre y los intervalos de bootstrap secuencial para 25 de 42 identificadores históricos del modelo: 15 en contexto 2048 y diez en 1024; 17 no tienen observación recuperada. El exponente del texto histórico cae dentro del intervalo correspondiente en uno de 25 casos. La correlación de Spearman es 0.269 entre los exponentes históricos y los exponentes QL, y -0.445 entre los 19 valores almacenados de Padé y los exponentes QL. Los cambios en el corpus, la precisión y el protocolo impiden atribuir todas las discrepancias a una sola causa. Solo cuatro registros de empleo exitosos contienen una revisión de peso inmutable. Re reproducir arrays no es una replicación end-to-end de inferencia.

Resultados matemáticos con hipótesis explícitas. En un soporte discreto finito, la entropía máxima sujeta a un momento interior prescrito de log-distancia tiene un optimizador único de ley de potencia con exponente real. Para un exponente positivo, ese momento debe ser menor que el de la distribución uniforme. Las sumas finitas exactas determinan el normalizador, la entropía, la información de Fisher y la capacidad calorífica formal. La varianza continua $(\log N)^2/12$ en $\gamma = 1$ no coincide con la discreta: para $N = 2000$ valen 4.81448 y 5.48412, respectivamente. La métrica de Fisher satisface $C_V = \gamma^2 I$; la longitud intrínseca integra $\sqrt{I}$, y ambas cantidades no tienen por qué alcanzar su máximo a la vez. El cuantil continuo de masa de la familia finita está definido a ambos lados de uno y vale N^f en uno; el cuantil de distancia entre tokens se obtiene mediante sumas parciales.

Aplicaciones y límites. Las experimentaciones de compresión, pruebas de extensión del contexto, perfiles por cabeza, comparaciones de alineación, hipótesis de circuito y descripciones arquitectónicas permanecen disponibles en sus secciones originales. Un error de salida en fila fija puede ser acotado a partir de masa removida y diámetro de valor, pero esa cota no certifica la precisión en tareas autoregresivas. Inverse Padé devuelve exactamente su contexto de entrada y por lo tanto no puede validar de forma independiente un horizonte de recuperación. La divergencia de un nulo posicional normalizado no identifica por sí misma un circuito. Los conteos de banda geométrica requieren el subespacio rotatorio real y la convención de emparejamiento; la finalización de un ciclo no establece la pérdida de información. Las analogías restantes y definiciones de puntaje tienen dominios explícitos y no se ha afirmado ninguna identificación física o interpretación predictiva calibrada. La referencia corregida de intervención se distingue de las implementaciones históricas defectuosas y sus puntajes agregados archivados.

La campaña de noviembre agrega 8 064 intervenciones salvadas en 32 prompts

compartidos para tres modelos. Las mejoras comunicadas en predicción de gradiente se refieren a la predicción de error, no a una mejora en la capacidad del modelo. El apéndice K integra ocho grupos formalmente verificados sobre nulos instrumentales, exposición afín, mezclas, cuantiles, sensibilidad de familia finita, truncamiento, softmax y la aproximación condicional RoPE. Sus pruebas establecen las implicaciones matemáticas mencionadas, no la tubería empírica ni la novedad bibliográfica. El Blueprint complementario enlaza los enunciados, las pruebas, las figuras y los cálculos interactivos de esta edición. TAFAgent y la guía, corregidos el 18 de noviembre, ofrecen rutas de aplicación y medición; sus resultados no validan independientemente las afirmaciones empíricas.

Guía de lectura: tres recorridos por este artículo

El manuscrito distingue perfiles de atención medidos, resultados condicionales para una familia finita de leyes de potencia, y hipótesis de intervención. La geometría sola no conecta estos tres niveles en una cadena causal validada.

(T) Pista matemática. Lea las Secciones 4–7 sobre la identidad rotatoria, el exponente candidato, la entropía máxima bajo un momento prescrito y cantidades termodinámicas finitas. La Sección 7.4 separa la cota de entropía finita de sus regímenes asintóticos. Las identificaciones físicas en la Sección 20 son hipótesis históricas conservadas, sujetas a sus correcciones locales, en lugar de consecuencias de estos teoremas.

(E) Ruta empírica y de ingeniería. Comience con las Secciones 2, 8 y 9. Luego lea las Secciones 11 y 12 para cuantiles de masa, ensayos de compresión registrados y álgebra de candidatos inversos. Estos cálculos requieren mediciones específicas del protocolo y evaluación de la tarea antes de su uso como reglas de intervención. El Anexo E proporciona la lista de verificación actual. La Sección 21 preserva la aplicación histórica; su interfaz no es una validación independiente del manuscrito.

(F) Registro completo de investigación. Lea secuencialmente para las observaciones, mecanismos propuestos, resultados negativos y correcciones. Las secciones 18 y 18.1, y el Apéndice F, registran los fracasos y pruebas no resueltas. El Apéndice J registra las correcciones y la ubicación del archivo histórico separado. Las recetas obsoletas están ausentes en esta edición corregida.

Índice

1. Introducción **8**
1.1. Una regularidad empírica 8
1.2. Cuatro preguntas 8
1.3. La Cadena Teórica 9
1.4. El Objeto Central 9
1.5. Las múltiples facetas de γ 10
1.6. Aportaciones de este artículo 10
1.7. Relación con el Trabajo Anterior 12
1.8. Cómo leer este artículo 13

2. Resumen de Resultados **13**

3. γ como coordenada de control candidata **15**
3.1. De una hipótesis observable a una hipótesis de intervención 15

3.2. Tres palancas con diferentes mecanismos 15
3.3. Curvas candidatas y observaciones históricas 16
3.4. Un flujo de trabajo de medición y evaluación 16

4. Contexto y Notación **17**
4.1. Codificación de Posición Rotativa 17
4.2. Ciclos Completados y el Contador del Plano No Ciclado 17
4.3. Tres Escalas de Referencia y la Distancia Exacta del Último Ciclo 18

5. Modelización de γ: geometría, aproximaciones y pruebas empíricas **20**
5.1. El Punto de Partida Empírico 20
5.2. De coordenadas rotativas a una puntuación candidata 21
5.3. El candidato de Padé y su dominio 24
5.4. Comparación del Panel Inicial con Entradas Explícitas 25
5.5. Regresión de Cinco Ejes Exploratoria 28
5.6. Cantidades Derivadas y Consecuencias Verificables 29

6. ¿Por qué la Ley de Potencia: Fundamento de la Entropía Máxima **30**
6.1. El problema de MaxEnt 31
6.2. Interpretación termodinámica 32
6.3. Dualidad con la Ley de Zipf 32
6.4. Funciones Termodinámicas Finitas 33

7. Familias de Potencias Finitas con Ley de Potencia y Límites con Soporte Infinito **33**
7.1. La Función de Partición y Estructura de Fase 33
7.2. Capacidad térmica continua y discreta en uno 34
7.3. Información de Fisher y Capacidad Calorífica 34
7.4. Límite de Entropía Finita y Regímenes Asintóticos 35
7.5. La familia con corte exponencial 36

8. Exponentes Específicos de Protocolo y Coordenadas Derivadas de la Familia **37**
8.1. Panel Inicial E4: Datos y Protocolo de Regresión 37
8.2. Panel Inicial Completo: 21 Entrada/s Modelo/Corpus 38
8.3. Cantidades de Familia Finita desde los Exponentes Ajustados 40
8.4. Observaciones y hipótesis mecanísticas 40
8.5. Protocolo de evidencia y reproducción 45
8.6. Las 25 filas disponibles 45
8.7. Acuerdo histórico y el Padé comparison 47

9. Flujo de trabajo de diagnóstico y controles de instrumentos **47**
9.1. La variable de longitud E4 tiene una nula uniforme no nula 48
9.2. Consultas fijas diferencian a QL de la media diagonal 49
9.3. Informe del observable y su control correspondiente 50
9.4. Cantidades derivadas de la familia y coordenadas inversas 50
9.5. Divergencia respecto a una referencia declarada 50
9.6. Reevaluación de los cuatro casos históricos 51

10. Heterogeneidad de capa y la propuesta de capa crítica **52**
10.1. Una capa coordenada candidata . . . 52
10.2. Una mezcla no tiene el exponente promedio . . . 52
10.3. CKA es una observación separada . . . 53
10.4. Entrenamiento de puntos de control y normalización de recuperación . . . 53

11. Compresión de KV-Cache: Presupuestos Masivos y Pruebas Registradas **54**
11.1. Ventanas continuas y discretas . . . 54
11.2. Una cota condicional de salida y asignación real de caché . . . 56
11.3. EXP-B2: las ventanas realmente evaluadas . . . 56
11.4. Asignación de atención con ley de potencia: presupuestos condicionales de masa . . . 57
11.5. Razón de base desplazada: alcance de la anterior expresión de pérdida de información . . . 58
11.6. Decaimiento de Horizonte Suave: una propuesta de sesgo en puntuación probada . . . 59
11.7. Pruning coordenado y la auditoría de subespacio rotatorio . . . 60

12. Extensión del contexto: Álgebra de Candidatos Inversos y Pruebas Empíricas **62**
12.1. Diseñar una base de candidatos para un exponente especificado . . . 62
12.2. El horizonte de información anterior como coordenada inversa . . . 63
12.3. Dirección futura: prueba de un controlador de retroalimentación . . . 64

13. Aboliciones Coordinadas y los Límites de la Interpretación de Bandas **64**
13.1. Planos rotatorios y nombres históricos de grupos . . . 64
13.2. EXP-B3: todas las ventanas y ambas evaluaciones guardadas . . . 65
13.3. La simulación del puntaje y el registro de amplificación de la GPU . . . 66
13.4. Lo que realmente cambia el código de amplificación . . . 67

14. Perfiles de atención, razones de peso y hipótesis de circuitos **68**
14.1. Una distribución de referencia condicional no es un detector de circuito . . . 68
14.2. Medidas en el espacio de pesos: coordenadas y promedio . . . 68
14.3. El campo de γ: heterogeneidad y su nulo instrumental . . . 72

15. Comparaciones de Base/Instruct bajo una consulta de longitud variable **75**
15.1. Estimando, protocolo y procedencia . . . 75
15.2. Qué las diferencias reportadas respaldan . . . 75

16. Contrastos de Corpus y la Proporción de Arquitectura No Validada **76**
16.1. Una estadística descriptiva con una convención de signo fija . . . 77
16.2. Comparaciones históricas y correcciones de arquitectura . . . 77
16.3. Rondas en la nube: estimador compartido, profundidad diferente de evidencia 78
16.4. Recalcular la propuesta de ratio . . . 78
16.5. ¿Qué falla y qué permanece sin probar . . . 79

17. Manual de Diagnóstico y Enciclopedia del Modelo **80**
17.1. Usando la CLI de diagnóstico archivada 80
17.2. Guía de decisión para medición 80
17.3. Diseño Arquitectónico: Prescripciones Escalares Retiradas 81
17.4. Búsqueda de modelos con dominios de cálculo explícitos 81

18. Limitaciones y Preguntas Abiertas **82**
18.1. Atlas de Fallos: limitaciones observadas e identificaciones no probadas . . . 91

19. Coordenadas Efectivas y Propuestas de Control Dinámico **92**
19.1. Tres superficies de control propuestas 92
19.2. Puntos de anclaje históricos: implementación y validez causal 93
19.3. Referencia de implementación corregida y su alcance verificado 95

20. Coordenadas Racionales y los Límites de la Analogía RG **97**
20.1. Definiciones candidatas de coordenadas y exponentes legados 97
20.2. Traectoria escalar exacta y estabilidad 97
20.3. M̈apa de Mobiüs: dominio correcto y punto fijo algebraico 98
20.4. Propuesta del umbral NIAH: no validada por el panel guardado 99
20.5. La puntuación de la razón de norma no tiene una interpretación establecida de empaquetamiento. 99
20.6. Laplaciano fraccionario: las suposiciones del operador faltante 100
20.7. Analogía conforme: una etiqueta de exponente no es una CFT 100
20.8. Identidades condicionales y sus dominios 101
20.9. Hallucinación como exceso del horizonte: una propuesta no calibrada . . . 102

21. TAF Agente: Registro Histórico de Software y Límites Evidenciales **103**
21.1. Puntos finales registrados y disposición de la fuente 104
21.2. Las cinco recetas históricas y su alcance actual 104
21.3. Catálogo de modos posterior y reglas diagnósticas retiradas 105
21.4. Mapa de fórmulas: cálculo versus interpretación empírica 107
21.5. Rol de aplicación y trabajo restante 108

22. Conclusión **109**

A. Verificaciones Simbólicas y Alcance de la Prueba Formal **110**

B. Auditoría del intento de derivación geométrica **111**

C. Tabla de Fórmulas **112**

D. Zipf y la atención: auditoría de panel guardado **114**

E. Tarjeta de Referencia Rápida **115**

F. Resultados Negativos Productivos **116**

G. Compilado de hallazgos críticos **120**

H. La Alpha-S Histórica: Definición y Fallo **122**
H.1. Definición conservada . . . 123
H.2. Aritmética en los inputs históricos declarados . . . 123
H.3. ¿Por qué se retira la regla de diseño? . . . 123

I. Glosario Bilingüe **124**

J. Registro de Corrección y Archivo Histórico Separado **125**

K. Ocho Grupos de Resultados Verificados Formalmente **127**
K.1. Grupo 1: el nulo instrumental y la calibración logarítmica . . . 127
K.2. Grupo 2: cotas para rejillas finitas en exposición afín . . . 129
K.3. Grupo 3: los perfiles de mezcla cambian la pendiente y la curvatura . . . 129
K.4. Grupo 4: ordenamiento de cuantiles discretos de masa . . . 130
K.5. Grupo 5: cálculo de familia finita y sensibilidad . . . 130
K.6. Grupo 6: qué garantiza la truncación de fila renormalizada . . . 131
K.7. Grupo 7: de errores en puntajes a errores tras el softmax . . . 132
K.8. Grupo 8: una integral coseno con un residuo explícito RoPE . . . 132
K.9. Prueba de procedencia, reproducibilidad y alcance restante . . . 134

1 Introducción

1.1 Una regularidad empírica

Calificación de medición. El patrón empírico inicial debe interpretarse con el control del instrumento en la Sección 9.1 Bajo su regla de muestreo de longitud variable, la atención causal uniforme ya proporciona un exponente ajustado cercano a 0.78 El patrón en sí mismo no establece una decaimiento aprendido, un mecanismo rotatorio o un exponente universal. La QL con consulta fija utiliza una nulidad y un protocolo diferentes.

Los perfiles de distancia de atención resumen cómo un modelo distribuye la atención sobre las separaciones de tokens bajo un corpus, contexto y estimador especificados. La familia

$$A(d) \;\propto\; d^{-\gamma} \tag{1}$$

proporciona una descripción compacta de algunos perfiles de este tipo. Su adecuacidad es una cuestión empírica: un softmax arbitrario de productos punto aprendidos no necesita tener esta forma. Un alto log-log R^2 en un intervalo de distancia seleccionado no establece la misma ley en sumideros de atención, en cada cabeza, o fuera de ese intervalo.

El panel inicial de modelos abarca desde Pythia-14M hasta BLOOM-7b1 y incluye RoPE, ALiBi y codificaciones posicionales aprendidas. En el panel de 21 filas reproducido, P03, P05 y P07 tienen $R^2 < 0.95$; los otros 18 superan este valor. Este panel motivó al candidato configuración-solo estudiado en la Sección 5, cuya precisión debe evaluarse en un conjunto explícitamente identificado de observaciones. Las elecciones sobre corpus, precisión, arquitectura, entrenamiento y medición deben considerarse al comparar los exponentes ajustados.

Las mediciones de octubre guardadas proporcionan una segunda comparación explícitamente definida (Sección 8). Los ajustes y los intervalos de bootstrap secuencial pueden volver a calcularse para 25 de 42 identificadores de modelos: 15 en el contexto 2048 y diez en 1024. Diecisiete identificadores carecen de observaciones recuperadas. Uno de los 25 exponentes iniciales se encuentra en el intervalo correspondiente. Las correlaciones de Spearman son 0.269 entre los exponentes iniciales y los recalculados, y -0.445 entre 19 valores almacenados de Padé y los exponentes recalculados. Estas comparaciones demuestran sensibilidad al protocolo de medición; ya que varias condiciones cambiaron simultáneamente, no identifican una única fuente causal de las diferencias. Solo cuatro registros de trabajo exitosos identifican una revisión de peso inmutable. Recalcular los arrays guardados verifica por tanto los cálculos reportados, no toda la tubería de inferencia.

Resumen en una línea. Medimos perfiles de distancia, analizamos una familia condicional finita de leyes de potencia y probamos intervenciones específicas. La descripción geométrica, la familia matemática y la respuesta empírica están conectadas mediante hipótesis y comparaciones explícitas, en lugar de a través de un exponente universal asumido.

1.2 Cuatro preguntas

Q1. ¿Por qué considerar una ley de potencia? Para un momento log-distancia prescrito y factible en soporte finito, la entropía máxima selecciona una ley de potencia. Si la atención medida cumple con este principio de optimización requiere evidencia.

Q2. ¿Qué determina γ? Un exponente medido depende del modelo y del protocolo. Comparamos con una configuración basada en PadÉl candidato; RoPE en sí mismo no deriva su igualdad.

Q3. ¿Qué es especial sobre $\gamma = 1$? Separa los regímenes de sumabilidad de la familia con soporte infinito. En cualquier contexto finito, el normalizador, la entropía y la varianza permanecen finitos y suaves. Los cuantiles continuos y discretos deben distinguirse.

Q4. ¿Qué podemos hacer con ? γ? Los perfiles y los límites condicionales pueden guiar experimentos de compresión, extensión del contexto y detección de anomalías. Nivel de tarea Los beneficios deben ser medidos bajo la intervención realmente realizada.

1.3 La Cadena Teórica

El argumento comienza con tres objetos separados. Primero, RoPE proporciona un conjunto de frecuencias de rotación y una descomposición exacta de los puntajes de atención en seno–coseno. Segundo, un momento log-distancia especificado define un problema de inferencia restringido. Tercero, las mediciones determinan si un modelo entrenado particular es adecuadamente descrito por la familia resultante.

Para un entero $N > 1$ y un momento factible interior $0 < \lambda < \log N$, maximizar la entropía de Shannon sobre distribuciones en $\{1, \dots, N\}$ con $\mathbb{E}_p[\log d] = \lambda$ selecciona una única ley de potencia. Su exponente es un multiplicador de Lagrange real, positivo solo por debajo del momento logarítmico uniforme $\log(N!)/N$. El argumento basado en la desigualdad de Gibbs y las hipótesis precisas se presentan en la Sección 6. Otras distribuciones pueden compartir el momento prescrito sin maximizar la entropía. La geometría de frecuencia no impone ni el momento ni el principio de optimización en consultas y claves entrenadas.

Esta distinción también limita la predicción geométrica propuesta. Una función racional de variables de configuración puede evaluarse sin inferencia, pero esa conveniencia computacional no la convierte en un exponente derivado entrenado. La sección 5 analiza al candidato, la derivación no exitosa y las comparaciones empíricas por separado. Todos los usos posteriores conservan este límite.

1.4 El Objeto Central

Dentro de la familia normalizada, mayores γ desplazan la masa hacia distancias más cortas. Para un modelo general, un ajustado $\widehat{\gamma}$ describe solo el perfil e intervalo en el que fue estimado. Un valor de configuración γ_P es un candidato separado. La regresión exploratoria de cinco ejes en la Sección 5.5, con reportados $R^2 = 0.44$, describe el panel inicial; sus coeficientes no identifican contribuciones causales separadas.

La cantidad geométrica es el número de planos rotatorios que han completado un ciclo. Para $m = d_{\text{rot}}/2$, $d > 0$ y $\theta > 1$, define

$$\boxed{n_{\text{active}}(d) = m - k_c(d), \qquad k_c(d) = \text{mín}\left\{m, \text{máx}\left\{0, 1 + \left\lfloor m\frac{\log(d/2\pi)}{\log\theta}\right\rfloor\right\}\right\}.} \tag{2}$$

Aquí n_{active} cuenta planos sin un ciclo completo, no dimensiones portadoras de información. El subespacio rotatorio real y el emparejamiento de coordenadas deben utilizarse. El

conteo logarítmico continuo es una aproximación a este conteo entero; no proporciona un exponente de atención entrenado ni un teorema de pérdida de información.

1.5 Las múltiples facetas de γ

La notación debe distinguir cantidades cuya igualdad no está establecida.

Cantidad	Significado y alcance
$\widehat{\gamma}$	Exponente estimado del perfil, con protocolo y incertidumbre.
γ	Parámetro de la familia de probabilidades finitas; las identidades son condicionales a este modelo.
γ_P	Candidato racional calculado a partir de la configuración, requiriendo comparación empírica.
$1/\gamma$	Coordenada formal de temperatura inversa cuando $\gamma \neq 0$, no la temperatura de muestreo de salida.

Las huellas arquitectónicas son estadísticas formadas a partir de estas cantidades. Su utilidad diagnóstica depende del panel de referencia y del protocolo de validación.

1.6 Aportaciones de este artículo

El artículo desarrolla los siguientes análisis conectados. Las afirmaciones matemáticas, cálculos derivados, mediciones del modelo y propuestas llevan diferentes evidencias.

C1. Configuración candidata y sus límites (Sección 5). Basándonos en la estructura medida de distancia de atención [2, 7], analizamos $\gamma_P = (2\theta - T\sqrt{2})/(2\theta + T\sqrt{2})$, incluyendo su dominio, inverso, panel de comparación y argumento fallido de fase estacionaria. La comparación inicial identifica todos los 13 residuos computables; la comparación de octubre utiliza su propio protocolo.

C2. Entropía máxima condicional (Sección 6). Especificamos el soporte, medida de referencia y momento factible, derivamos la familia óptima y distinguimos su entropía $S = \log Z + \gamma \mathbb{E}[\log d]$ de la entropía de atención medida.

C3. Regímenes finitos y asintóticos (Sección 7). Las sumas exactas separan los cálculos de contexto finito de los límites de soporte infinito y las aproximaciones continuas. Las comparaciones de compresibilidad entre modelos permanecen observaciones, no razones universales.

C4. Compresión del caché KV (Sección 11). Derivamos ventanas continuas y discretas de masa y un límite de salida con filas congeladas. Los experimentos conservan las ventanas realmente probadas: EXP-B2 utilizó 445 tokens para Pythia-1B y 1482 para Pythia-2.8B. La comparación de soft-decay favorece políticas diferentes en sus tres modelos; sus porcentajes dependen del denominador indicado. Los casos no exitosos permanecen parte de la evidencia, y ninguno establece garantías generales de precisión.

C5. Extensión del contexto (Sección 12). EXP-B1v5 compara cuatro modelos Pythia $\alpha \in \{1, 4, 8, 16\}$ y longitudes nominales 512–4096. A 4096, $\alpha = 4$ obtiene $86, 100, 100, 100\,\%$ en clasificación de clave restringida. No domina en todos los escenarios y experimenta pequeñas pérdidas en contexto corto registradas en Pythia-70M. El panel de modelos pequeños duplica a v4; la probabilidad nominal no es un control medido. Ni un óptimo ni el rendimiento de ajuste fino se sigue de estos agregados.

C6. Entropía y intervenciones de banda (Secciones 7.4 y 13). El límite de entropía finita es matemático; los cálculos de la familia finita reproducidos se tabulan en la Sección 8. EXP-B3 reportó aumentos de perplexidad de 350 y 9.3 en Pythia-1B para dos conjuntos de filas de pesos ablatadas. La razón obtenida a partir de los registros no redondeados es 37.45 (aproximadamente 37×). Este contraste motiva una hipótesis de bandas, cuya interpretación causal requiere emparejamiento correcto y controles parciales de RoPE.

C7. Anomalías de atención (Sección 14). La divergencia respecto a una referencia posicional normalizada identifica desviaciones de dicha referencia. La identificación de circuitos requiere además intervenciones funcionales.

C8. Flujo de medición y paneles de modelos (Secciones 9 y 17). Las tablas de modelos, el software de medición y los registros del protocolo apoyan comparaciones reproducibles. Las observaciones disponibles y faltantes se reportan explícitamente.

C9. Propuestas de control (Secciones 3 y 19). Separamos los cambios de RoPE, la temperatura del puntaje, el entrenamiento y el muestreo de salida. Las bandas propuestas debajo de 0.65, desde 0.65 hasta 0.85, y por encima de 0.85 son exploratorias Diseños de elección. Un termostato requiere evaluación previa de tareas.

C10. Geometría de la información y analogías físicas (Secciones 7.3 y 20). La identidad de familia finita $C_V = \gamma^2 I$ y las derivadas de mapas racionales tienen dominios matemáticos explícitos. Las interpretaciones de operadores fraccionarios, conformales y de renormalización son evaluadas frente a las estructuras adicionales que requerirían.

C11. Evidencia de intervención (Sección 19.2). Los ganchos históricos adaptativos fallan en la causalidad previa; sus puntajes reportados y la comparación anterior de recuperación Fisher no pueden validar el desempeño causal. La campaña de noviembre proporciona 8 064 intervenciones en 32 prompts compartidos para tres modelos, incluyendo controles y pares de incertidumbre. Las reducciones de errores predictivos no son mejoras en la capacidad del modelo.

C12. Análisis de fallos (Sección 20.9 y Anexo F). Los puntajes propuestos de ilusiones geométricas carecen de calibración en errores factuales. Sus fallos en el dominio y las hipótesis del controlador y físicas no exitosas delimitan lo que el marco puede soportar.

Mapa del lector. Las Secciones 5–7 separan la geometría del candidato de las matemáticas condicionales; Sección 8 presenta los paneles de medición y la reproducción. Las Secciones 11–12 y 19 contienen aplicaciones de intervención. Los diagnósticos, comparaciones de arquitectura y descripciones de software conservan sus métodos detallados y

ejemplos. El catálogo de fórmulas y el compendio organizan los resultados con sus dominios; el documento separado de cambios y evidencia registra el historial de versiones y el alcance exacto de los certificados formales.

1.7 Relación con el Trabajo Anterior

Perfiles de atención y geometría rotatoria. Trabajo sobre la estructura atención-distancia y representaciones rotatorias [2, 7, 40, 50] proporciona contexto para las mediciones de perfil examinadas aquí. Este manuscrito no establece un exponente de primeros principios para la atención aprendida. Su resultado de máxima entropía condicional y su candidato de configuración tienen premisas diferentes y deben evaluarse por separado de los ajustes empíricos.

Extensión del contexto. La literatura sobre extensión del contexto estudia los horarios rotatorios y la extrapolación [8, 24, 27, 30, 35, 52]. Esquemas posicionales híbridos y análisis de dimensiones rotatorias proporcionan comparaciones adicionales [39, 47]. Invertir nuestro candidato no deriva los parámetros óptimos de estos métodos ni identifica qué capas pueden descartar la información rotatoria. La sección 12 da la inversa precisa y la separa de los experimentos de recuperación registrados.

Compresión de KV-cache. SnapKV, H2O, StreamingLLM y PyramidKV utilizan mecanismos diferentes de selección o asignación [4, 23, 45, 51]. Las evaluaciones BLASST y enfocadas en razonamiento ofrecen comparaciones empíricas adicionales [25, 49]. Trabajo relacionado con compresión y salto de capas incluye [10, 31, 37, 53]. Estos métodos no implementan todos una ventana de distancia con cuantil. Nuestro presupuesto de masa condicional no explica un punto de operación universal ni reemplaza su calibración. En particular, el coeficiente BLASST y la coincidencia de retención PyramidKV afirmados previamente eran incorrectos; la álgebra corregida y la comparación numérica se encuentran en la Sección 11.4.

Diagnósticos posicionales y circuitos. Estudios sobre atención posicional y simbólica [2, 13, 43] motivan hipótesis sobre bandas rotatorias. Aquí se mantiene la asimetría observada entre la ablación de bandas, junto con limitaciones instrumentales, en lugar de presentarla como una prueba de un depósito semántico. De igual manera, un contraste de corpus o una divergencia KL respecto a un nulo posicional es un diagnóstico descriptivo. Los circuitos Transformer-circuito y cabeza de inducción trabajan [12, 34], y los métodos de parcheo de activación [15, 32], motivan pruebas causales adicionales; una reversión de signo por sí sola no identifica sus circuitos.

Entropía y analogías físicas. La entropía de atención, diagnósticos en el dominio de frecuencia y la escala de atención son estudiados bajo diferentes definiciones y límites. [7, 16, 46] Enfoques estadístico-mecánicos sobre pesos, entrenamiento, generación o caminos de atención proporcionan objetos relacionados pero distintos. [6, 11, 17, 20, 33, 41, 42]. Our Z_N es una suma finita; su límite con soporte infinito es $\zeta(\gamma)$ Solo para $\gamma > 1$ La coordenada de temperatura inversa de la familia no es la temperatura de muestreo de salida del modelo.

El límite de Shannon finito $S_N \leq \log N$ no identifica una entropía de enredo ni establece una ley de área geométrica para la atención entrenada. La teoría de Ryu–Takayanagi

y el trabajo sobre el aprendizaje de su mapeo inverso [21, 38] no validan esa inferencia. Ninguna afirmación de ser la primera derivación a partir de la geometría posicional es respaldada por el argumento corregido. Las comparaciones históricas y su formulación original permanecen en el archivo separado de predecesores indexado por el Apéndice J.

1.8 Cómo leer este artículo

El diagrama de dependencia de la Figura 3 distingue a la matemática condicional de las pruebas empíricas. Los puntos de entrada en la Tabla 1 Completa el índice del lector al inicio del documento.

Cuadro 1: Rutas de lectura para las conclusiones corregidas y la evidencia conservada.

Lector	Comienza con	Revisa junto a
Investigador en ML	Perfiles y comparaciones de candidatos, Secciones 5, 8	Limitaciones del protocolo y resultados negativos.
Matemático lector	Resultados de familia finita, Secciones 6–7	Soporte, medida de referencia y límites de exponente fijo.
Ingeniero de compresión	Cuantiles e intervenciones, Secciones 11–12	Tarea real y evaluación de recursos.
Historial de auditoría del lector	Apéndice J y el archivo fuente	Registro de corrección y fuente inmediata anterior exacta.

Anexo E presenta una lista de verificación para medición y evaluación; el Anexo G establece el estado actual de F1–F20. No prometen tiempos de ejecución independientes del modelo, seguridad en producción o novedad. Las analogías físicas y las descripciones históricas del software en otros lugares requieren su calificación local, incluso cuando se mantiene el terminología original.

2 Resumen de Resultados

Los resultados respaldados tienen tres formas distintas: identidades dentro de una familia de probabilidad especificada, cálculos reproducidos a partir de arreglos archivados y observaciones empíricas bajo protocolos registrados. Una identidad evaluada en un exponente ajustado no es una medida adicional e independiente de un modelo.

Cuadro 2: Resultados respaldados y la evidencia que los respalda.

Resultado	¿Qué se establece	Evidencia y alcance
Máximo entropía	Optimizador único de ley de potencia en un interior prescrito de momento logarítmico.	Teoremas en la Sección 6; una afirmación de inferencia condicional.
Cálculo de familias finitas	Momentos exactos, entropía, información de Fisher, capacidad térmica y cuantiles de masa discreta.	Secciones 7 y 11; soporte especificado y parámetro.
Instrumentos de control	La variable de longitud E4 tiene una pendiente uniforme no nula; QL con consulta fija tiene pendiente cero.	Sección 9.1; comprobaciones analíticas y sintéticas.
Código de referencia corregido	Contratos causales, rotatorios, de pérdida y numéricos pasan 145 comprobaciones.	Sección 19; pruebas de tensor CPU y aritméticas.
Intervenciones de noviembre	8 064 intervenciones salvadas, 32 prompts compartidos, tres modelos, incluyendo 1 152 controles de fuerza cero.	Sección 19; predicción de efectos de intervención, no mejora del rendimiento del modelo.

Cuatro cantidades y su alcance

Cuadro 3: Alcance actual de las cuatro cantidades anteriormente presentadas como herramientas de despliegue.

Cantidad	Definición	Interpretación
Candidato de Padé	$\gamma_P = (2\theta - T\sqrt{2})/(2\theta + T\sqrt{2})$	Candidato algebraico en $(-1, 1)$ para entradas positivas; no es un exponente medido universal.
Ventana de masa	$D_{f,N} = \min\{k : \sum_{d=1}^{k} d^{-\gamma} \geq fZ_N\}$	El cuantil discreto exacto. El análogo continuo se presenta en la Sección 11; ninguno de ellos representa una garantía de precisión en tareas.
Base inversa	$\theta_{\text{inv}} = T\sqrt{2}(1+g)/(2(1-g))$	Resuelve $\gamma_P = g$ para $-1 < g < 1$. Su relación con una base nominal no es un óptimo demostrado de NTK.
Corpus contrastante	$\Delta\widehat{\gamma} = \widehat{\gamma}_{\text{random}} - \widehat{\gamma}_{\text{text}}$	Diferencia dependiente del protocolo. Su signo solo no identifica la arquitectura o un circuito de inducción.

Puntos clave de la evidencia

La comparación inicial de predictores tiene 18 entradas modelo/corpus, de las cuales 13 tienen una base registrada. En esas 13, el error absoluto medio del exponente es 0.167766 y el error absoluto relativo mediano es 19.2144 % La afirmación sobre la precisión del caso seleccionado se retira por razones de procedencia según la Sección 5.4 La repetición de octubre separada tiene 25 observaciones entre 42 identificadores históricos, incluyendo 17 observaciones no disponibles. Solo cuatro registros de trabajos exitosos identifican

una revisión inmutable de peso. La estimación del texto histórico se encuentra dentro del intervalo de octubre en una de 25 comparaciones; la estimación almacenada Pad Él candidato y el exponente de octubre QL tienen correlación de Spearman -0.445 De 19 pares disponibles. Las diferencias en el protocolo impiden atribuir estas discrepancias a una sola causa.

En soporte finito, la entropía máxima selecciona una distribución de ley de potencia única en un momento log-distancial prescrito en el interior. La entropía, la información de Fisher y la capacidad térmica formal se calculan mediante sumas finitas. Quedan finitas en $\gamma = 1$; la frontera de sumabilidad en soporte infinito no es una transición demostrada en la capacidad de redes finitas. En $N = 2000$ y $\gamma = 1$, la varianza discreta es 5.48412, en comparación con 4.81448 para la densidad continua.

La campaña de noviembre registra 8 064 intervenciones en 32 prompts compartidos para tres modelos, incluyendo 1 152 controles de fuerza nula. Sus mejoras se refieren a la predicción de los efectos de intervención, no a una mejora del rendimiento del modelo. Los experimentos históricos de KV, recuperación, alineación y ablación de banda permanecen en sus secciones originales junto con sus limitaciones en el protocolo.

Cómo utilizar los resultados

Mida un perfil bajo un protocolo declarado, compare familias alternativas ajustadas y propague su incertidumbre en cantidades derivadas. Pruebe cualquier propuesta de caché o intervención rotativa contra las bases de tarea y recursos. El manuscrito no establece ni una relación de compresión segura solo configuración ni una regla general de extensión de contexto o clasificación de arquitectura. El tiempo de medición depende del tamaño del modelo, la hardware y el protocolo; no se respalda una estimación universal de tres minutos para CPU.

3 γ como coordenada de control candidata

3.1 De una hipótesis observable a una hipótesis de intervención

Un exponente ajustado describe un perfil atención-distancia bajo un corpus, soporte y agregación especificados. No determina toda la computación de una red entrenada. Dentro de la familia finita, para una estadística u independent of γ la diferenciación da

$$\partial_\gamma \mathbb{E}_{p_\gamma}[u] = -\operatorname{Cov}_{p_\gamma}(u, \log d). \tag{3}$$

Al aumentar $u(d)$, esta derivada es no positiva: aumentar el exponente de la familia positiva desplaza la masa hacia distancias más cortas. Esta afirmación no identifica una intervención en el modelo que realice un cambio elegido en γ. Las anteriores tareas se limitan a 0.65 y 0.85 y no tienen un estatus universal establecido. Su interpretación original se conserva en el archivo predecesor separado, no se utiliza como clasificación de despliegue aquí.

3.2 Tres palancas con diferentes mecanismos

Elegir un esquema de frecuencia antes del entrenamiento, cambiar las frecuencias rotatorias en la inferencia y modificar las cabezas individuales de atención son diferentes intervenciones. Requieren controles separados. La temperatura del puntaje de atención

también es distinta de la temperatura de muestreo de salida. Una identidad de covarianza para la familia elegida no puede predecir la respuesta completa de las consultas aprendidas, las claves y los estados posteriores. Los perfiles por cabeza en la Sección 14.3 motivan la medición a una resolución más fina, mientras que sus capas faltantes limitan las conclusiones del modelo completo.

Para el candidato en sí mismo

$$\partial_\theta \gamma_P = \frac{4T\sqrt{2}}{(2\theta + T\sqrt{2})^2} > 0.$$

Así, aumentar θ incrementa el exponente candidato a valor fijo de T. Esto corrige la dirección del controlador propuesto anteriormente; no demuestra que un modelo entrenado siga esta derivada.

3.3 Curvas candidatas y observaciones históricas

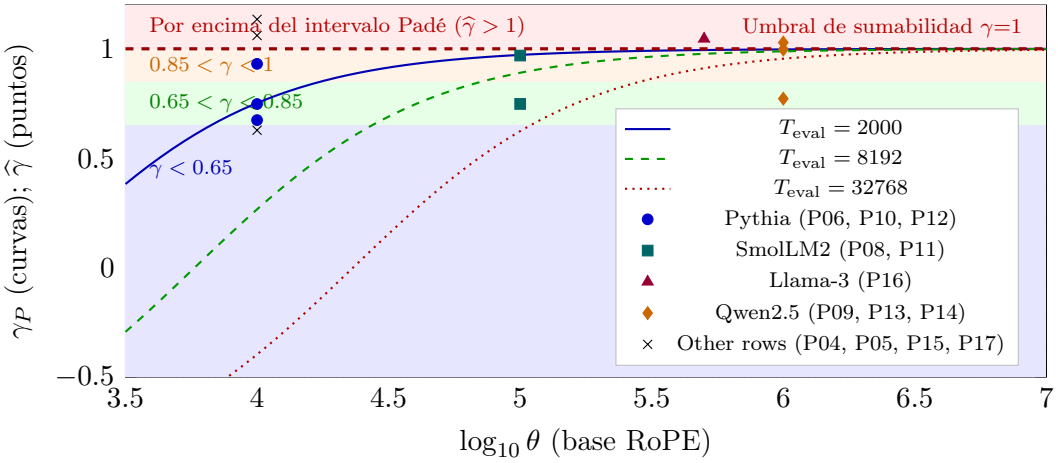


Figura 1: Curvas candidatas y exponentes históricos ajustados conservados. Las curvas evalúan γ_P En los tres contextos mostrados. Los 13 puntos utilizan las bases registradas y los exponentes reproducidos de la Tabla 8 No son mediciones bajo un contexto o corpus común. Las bandas sombreadas preservan los antiguos intervalos numéricos para orientación, sin asignar una capacidad de tarea o un régimen de despliegue seguro. Los puntos por encima de uno no pueden ser predicciones de este candidato. La evidencia del perfil actualizado se encuentra en la Sección 8.

La figura compara una familia algebraica con exponentes ajustados históricamente. Las curvas y puntos no constituyen una escaneo controlado de frecuencia-base. Los valores superiores a uno pueden ocurrir en un perfil medido, aunque no pueden ser salidas de este Pad.Él candidato. Los intervalos coloreados no tienen establecida ninguna Garantía de estabilidad a largo plazo, razonamiento o garantía de estabilidad durante el entrenamiento.

3.4 Un flujo de trabajo de medición y evaluación

1. Define la tarea, el contexto de evaluación y el presupuesto del recurso base.
2. Modelo de record, revisión del tokenizer y corpus, backend de atención, precisión, selección de consultas, soporte y estimador; estimar la incertidumbre del perfil.
3. Calcula cualquier cantidad familiar en el soporte indicado. Trata la inversa Padćomo un escenario experimental propuesto, no como una prescripción óptima para la tarea.

4. Evaluar intervenciones en secuencias de prueba con baselines emparejadas, registrando tanto la calidad de la tarea como los cambios reales en la memoria o latencia.

El termostato de la sección 12.3 sigue siendo una propuesta de control no validada. Es necesario un análisis de respuesta medida y una análisis de estabilidad antes de cualquier afirmación sobre control en tiempo real.

4 Contexto y Notación

Distinguiamos la mapa rotatorio, un conteo de ciclos completados, y las distancias derivadas de ese conteo. Ninguna de estas definiciones por sí sola determina la distribución de atención aprendida.

4.1 Codificación de Posición Rotativa

Sea $h = d_{\text{head}}$ la dimensión de la cabeza de consulta/clave y sea $d_{\text{rot}} = 2m \leq h$, $m \geq 1$ las dimensiones reales rotatorias. La rotación geométrica estándar RoPE [40] rota m planos bidimensionales. Para $\theta > 1$ y el índice de plano cero basado i,

$$\omega_i = \theta^{-i/m}, \qquad i = 0, \ldots, m-1. \tag{4}$$

Las frecuencias van desde 1 hasta $\theta^{-(m-1)/m}$. El período más lento es exactamente $2\pi\theta^{(m-1)/m}$, acercándose a $2\pi\theta$ solo cuando $m \to \infty$. La RoPE parcial deja las coordenadas $h - 2m$ sin rotar; los horarios de frecuencia modificados requieren sus frecuencias reales en lugar de esta fórmula geométrica.

Para vectores de contenido fijos q, k y rotaciones R, la ortogonalidad da $(R(m_0)q)^\top R(n_0)k = q^\top R(n_0 - m_0)k$. El factor posicional explícito depende del desplazamiento relativo firmado. No necesita ser par en ese desplazamiento porque la puntuación total tiene un término seno. Los vectores de contenido pueden depender ellos mismos del contexto circundante, por lo que esta identidad no implica que la atención dependa solo de la distancia absoluta ni garantice la generalización posicional.

Analogía del reloj. Un plano registra una fase módulo 2π en cada distancia, incluyendo antes de su primer giro completo. Completar un giro es un evento umbral en la fase desenrollada; no hace que el plano sea constante ni borre toda la información en el vector conjunto de fases. Diferentes lecturas de reloj pueden distinguir posiciones ideales, pero la discriminación útil también depende de los pesos, precisión y ruido.

4.2 Ciclos Completados y el Contador del Plano No Ciclado

Definición 4.1 (Planos cíclicos y no cíclicos)**.** Para distancia real $d > 0$, se llama plano i ciclado cuando $d\omega_i \geq 2\pi$. Los símbolos legados n_{dead} y n_{active} denotan conteos ciclados y no ciclados, respectivamente; no clasifican la utilidad funcional.

$$n_{\text{dead}}(d) = \text{mín}\left\{m, \text{máx}\left\{0, 1 + \left\lfloor m\frac{\log(d/2\pi)}{\log\theta}\right\rfloor\right\}\right\}. \tag{5}$$

$$n_{\text{active}}(d) = m - n_{\text{dead}}(d). \tag{6}$$

Demostración. Dado que $\log\theta > 0$ el umbral es equivalente a $i \le x = m\log(d/2\pi)/\log\theta$ entre enteros $0,\dots,m-1$ el número que satisface esta desigualdad es cero $x < 0$, $1+\lfloor x\rfloor$ para $0 \le x < m-1$, y m para $x \ge m-1$. El recorte combina estos casos e incluye la igualdad en el umbral. Restando de m da el recuento no ciclado. □

Una aproximación continua conveniente es

$$\widetilde{n}_{\text{active}}(d) = m\left[1 - \frac{\log(d/2\pi)}{\log\theta}\right]_0^1, \qquad [u]_0^1 = \text{mín}(1, \text{máx}(0,u)).$$

Su error absoluto es a lo sumo un plano. Para la fracción, el error es a lo sumo $1/m$; el error relativo frente a una cuenta exacta pequeña o nula no tiene cota uniforme. La escalera y la línea recta recortada son un libro de contabilidad de fases, no una derivación de un puntaje de atención con ley de potencia. En particular, el ciclo no implica que una coordenada contribuya ruido.

Cuadro 4: Conteos exactos para $\theta = 10{,}000$, $d_{\text{rot}} = 64$ ($m = 32$ Todos los valores se derivan de la definición de umbral; no se infiere ningún comportamiento del modelo.

d	n_{dead}	n_{active}	n_{active}/m
10	2	30	93.75 %
628	16	16	50 %
2,000	21	11	34.375 %
7,071	25	7	21.875 %
62,832	32	0	0 %

4.3 Tres Escalas de Referencia y la Distancia Exacta del Último Ciclo

Es útil conservar tres escalas de referencia distintas, mientras se separan sus definiciones. Solo la primera y la tercera surgen del recuento continuo; la segunda es una coordenada de comparación elegida.

Definición 4.2 (Distancias de referencia)**.**

$$T_{\text{cross}} = 2\pi\sqrt{\theta} \qquad \text{(continuous half-count)} \tag{7}$$

$$d_{\text{alias}} = \theta/\sqrt{2} \qquad \text{(comparison scale)} \tag{8}$$

$$T_{\max} = 2\pi\theta \qquad \text{(continuous zero-count)} \tag{9}$$

Para entero $1 \le r < m$, exactamente r planos son cycled

$$2\pi\theta^{(r-1)/m} \le d < 2\pi\theta^{r/m}.$$

Para par. m, tome $r = m/2$ para obtener el intervalo exacto de medio conteo. En su punto extremo superior $2\pi\sqrt{\theta}$, un ciclo adicional de plano. El último ciclo de plano se realiza en $T_{\text{last}} = 2\pi\theta^{(m-1)/m}$, antes de $T_{\max}$. Por tanto, ninguno de los extremos es una prueba de que la posición se vuelva indetectable. También, $\theta/\sqrt{2}$ no es el cero del Pad; el exponente en este artículo: que el candidato desaparece en $T = \sqrt{2}\theta$.

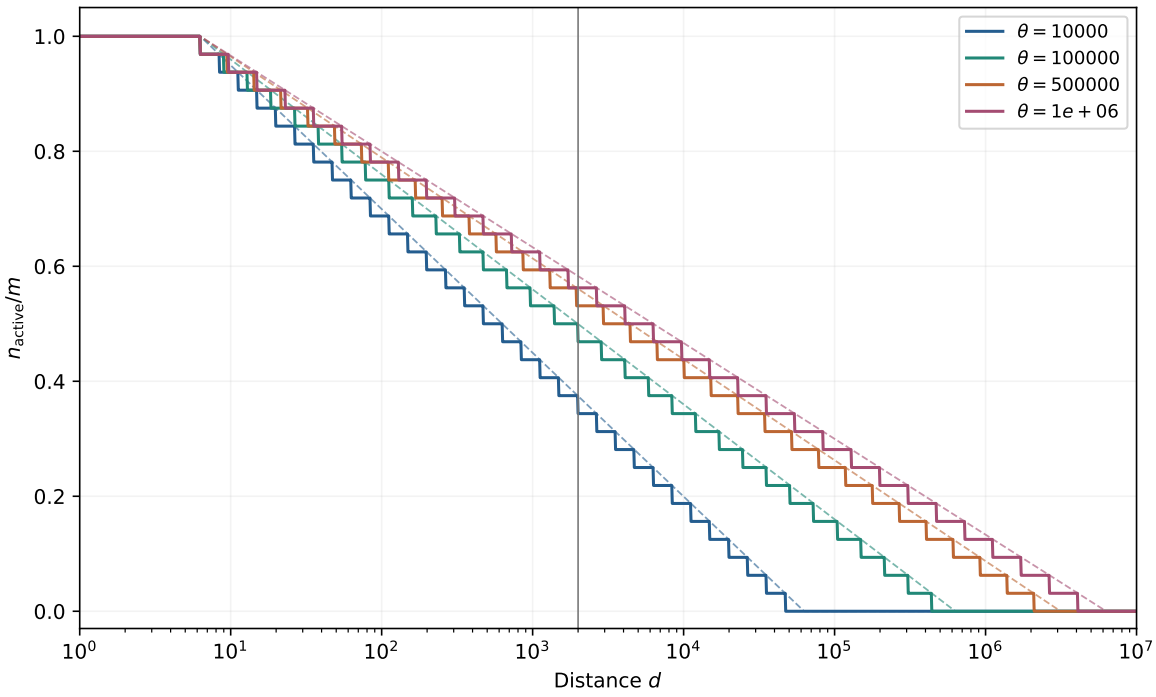


Figura 2: Fracción no ciclada para $m = 32$ y cuatro bases geométricas. Las curvas sólidas utilizan el recuento exacto de enteros; las curvas discontinuas emplean la aproximación continua recortada. La línea vertical marca $d = 2000$. Las magnitudes representadas cuentan umbrales de fase y no miden el decaimiento de la atención o un techo de recuperación.

Cuadro 5: Distancias de referencia, redondeadas al token más cercano; la fila final utiliza $m = 32$ Las columnas especifican bases, no límites empíricamente certificados de los modelos nombrados.

	$\theta = 10^4$	$\theta = 5 \times 10^5$	$\theta = 10^6$
T_{cross}	628	4,443	6,283
d_{alias}	7,071	353,553	707,107
T_{max}	62,832	3,141,593	6,283,185
T_{last}	47,117	2,084,765	4,080,185

Prueba una escala de contexto común. La recuperación, la extensión del contexto y la compresión del caché investigan diferentes observables [23, 26, 35]. Una dependencia común en $\sqrt{\theta}$ requeriría experimentos emparejados variando θ mientras se mantienen otros factores fijos. El conteo de ciclos no establece que estas tareas compartan una pared de contexto.

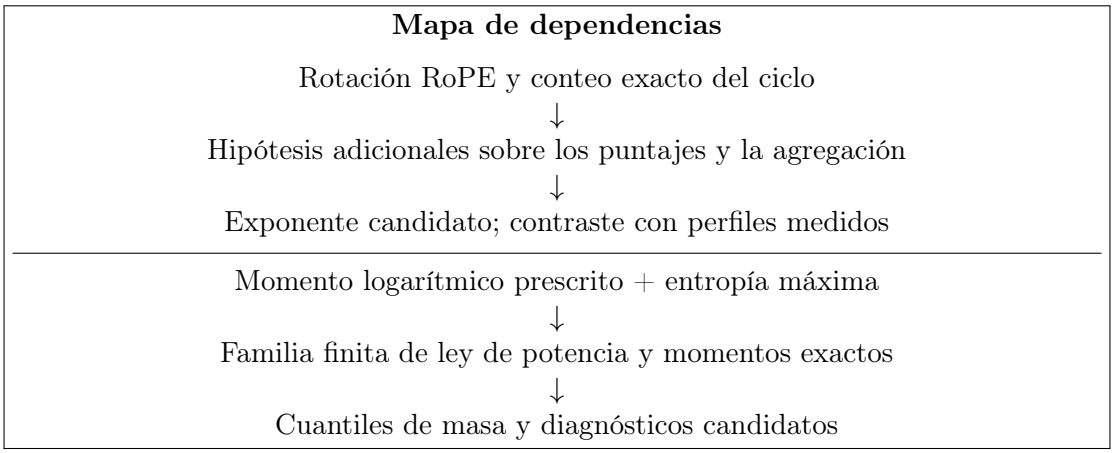


Figura 3: Dependencias lógicas. La rama MaxEnt parte de una restricción prescrita, no solo del recuento de ciclos. Convertir un cuantil de masa o un parámetro inverso en una intervención útil requiere validación empírica separada.

5 Modelización de γ: geometría, aproximaciones y pruebas empíricas

El conteo rotatorio es exacto bajo su programa de frecuencia establecido. Una predicción para un exponente de atención entrenado requiere suposiciones adicionales sobre puntajes de contenido, normalización y agregación. Desarrollamos la construcción candidata explícitamente, identificamos dónde entran esas suposiciones y comparamos el resultado con mediciones guardadas. Los ajustes empíricos no prueban la construcción que los motivó.

5.1 El Punto de Partida Empírico

Una regresión de ley de potencia representa un perfil muestreado por

$$\log A(d) = \log A_0 - \widehat{\gamma} \log d + \epsilon_d. \tag{10}$$

Aquí $A(d)$ es un agregado específico del protocolo y ϵ_d es su residuo logarítmico ajustado. El programa E4 promedia entradas de atención finitas a través de cabezas y capas en la última posición de consulta, luego a través de prompts y semillas. Descarta entradas no finitas sin registrar su fracción final. Los tokens aleatorios cambian la distribución de contenido; no eliminan las interacciones de contenido aprendidas ni los confundos instrumentales. El procedimiento de octubre QL en cambio tiene su propio perfil de contexto fijo y protocolo de bootstrap (Sección 8).

El panel E4 contiene tanto buenos como malos ajustes: por ejemplo, el texto de Pythia-1.4B tiene $R^2 = .841258$, mientras que el texto de Pythia-2.8B tiene $R^2 = .999287$ en

solo tres distancias. Por lo tanto, no hay una afirmación respaldada de que cada modelo probado tenga $R^2 > .95$, ni que cada curva abarque 10–2000. Un ajuste descriptivo de ley de potencia no es una ley exacta y no identifica únicamente su mecanismo generativo.

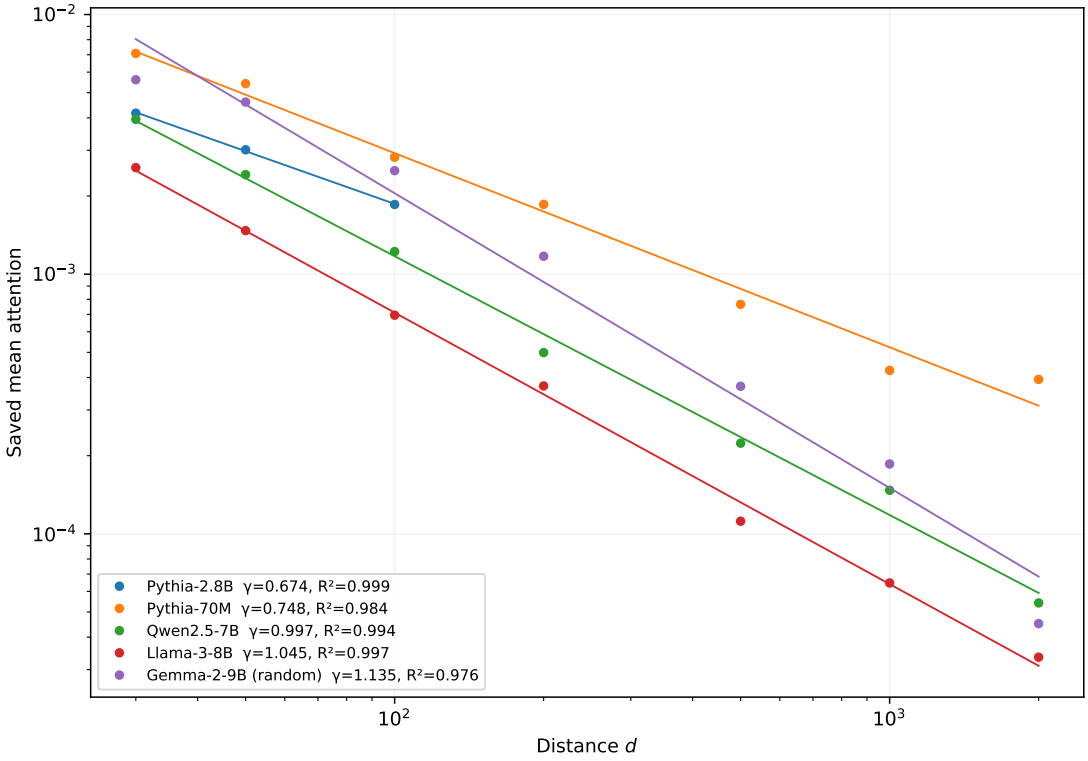


Figura 4: Cinco perfiles E4 guardados representativos. Los puntos son las medias en cada rejilla declarada de ajuste de cada registro; las líneas utilizan tanto la pendiente como la intersección de esa regresión. Pythia-2.8B abarca solo 30–100. La leyenda reporta cada valor real. R^2 No se impone una reclamación de ajuste uniforme de calidad.

5.2 De coordenadas rotativas a una puntuación candidata

Los Grupos 7–8 del Apéndice K proporcionan la cota verificada que relaciona errores en puntajes con errores tras el softmax y el resto de la integral de cosenos, bajo hipótesis explícitas para cada nivel de la fila.

Paso 1: descomposición exacta del puntaje. Sea $\delta = n - m_0$ el desplazamiento clave menos consulta con signo, sea h la dimensión completa de la cabeza, y sea $J = R'(0)$. Para los pares rotados $m = d_{\text{rot}}/2$ y el resto no rotado $q_\perp, k_\perp$, el puntaje ordinario de producto punto escalado es

$$s(m_0, n) = \frac{1}{\sqrt{h}} \Bigg\{ q_\perp^\top k_\perp + \sum_{i=0}^{m-1} \big[(q_i^\top k_i) \cos(\delta\omega_i) + (q_i^\top J k_i) \sin(\delta\omega_i) \big] \Bigg\}. \tag{11}$$

Esto se sigue al expandir $R(\delta\omega_i) = I\cos(\delta\omega_i) + J\sin(\delta\omega_i)$ en cada plano. Deben agregarse adicionales sesgos y máscaras de atención por separado. En implementaciones de

mitad-parte, el emparejamiento de coordenadas físicas debe coincidir con la implementación antes de aplicar esta identidad para ponderar las filas.

Paso 2: un modelo de puntaje condicional. Un coseno ya puede ser negativo antes de completar un ciclo: en fase π es igual a -1. Por el contrario, completar un ciclo no hace que su contribución promedie a cero sobre el contenido. Una pequeña covarianza consulta/clave no establece tampoco esa afirmación. Una hipótesis de modelado explícita sería

$$\mathbb{E}[s(d)] \approx b_0 + a\,\widetilde{n}_{\text{active}}(d). \tag{12}$$

El coeficiente a lleva las amplitudes y la escala del puntaje; establecerlo en uno es una elección adicional. En la parte no recortada del recuento continuo, este ansatz tiene un término de distancia logarítmica con coeficiente $-am/\log\theta$ e intercepto $b_1 = b_0 + am(1 + \log(2\pi)/\log\theta)$.

Paso 3: un modelo de normalización condicional. Si una fila idealizada tiene puntajes determinísticos $b_1 - \kappa \log d$ en un soporte fijo, entonces su softmax es exactamente una ley de potencia:

$$p(d) = \frac{e^{b_1 - \kappa \log d}}{\sum_{u=1}^{N} e^{b_1 - \kappa \log u}} = \frac{d^{-\kappa}}{\sum_{u=1}^{N} u^{-\kappa}}, \qquad \kappa = \frac{am}{\log\theta}. \tag{13}$$

Reemplazar los puntajes aleatorios por su media antes del softmax no es una identidad. Para vectores de puntajes equiprobables $(2, 0)$ y $(0, 0)$, la esperanza del primer componente del softmax es 0.690399, mientras que el softmax de la media da 0.731059. Promediar a través de cabezas, posiciones de consulta o contextos agrega operaciones no conmutativas adicionales. En E4 la longitud del contexto también cambia con d, por lo que la normalización no necesita ser la misma a lo largo de la curva representada.

Pythia-70M tiene dimensión de cabeza 64 pero dimensión rotatoria 16: su `rotary_pct` $= 0.25$ deja solo $m = 8$ pares rotados. Con $a = 1$ y $\theta = 10^4$, el exponente condicional es 0.868589, no 3.474356. En relación con la estimación del texto E4 0.747602, la discrepancia es 16.18 %; la nula uniforme coincidente E4 es 0.782582. Corregir la dimensión por tanto no valida la hipótesis ni sus suposiciones de promedio. Más directamente, establecer todas las proyecciones de consulta a cero da atención uniforme sobre las claves no mascaradas para cada base RoPE, siempre que no haya un posible sesgo de puntaje dependiente de la clave adicional. Ese contraejemplo muestra por qué la geometría RoPE sola no puede forzar un exponente de ley de potencia no nulo.

Una verificación de configuración contra perfiles de consulta fija. Table 6 evalúa solo lo estipulado $a = 1$ enfoque, usando $m = d_{\text{rot}}/2$ De configuraciones guardadas. Las entradas QL son los ajustes de texto de octubre existentes en $T = 2048$ en FP32. Cada candidato se encuentra fuera de su correspondiente intervalo de bootstrap guardado. En particular, los dos subespacios rotatorios más pequeños tienen las mayores pendientes QL en este conjunto de seis filas; las pendientes no aumentan de manera monótona con m como este fijo-a El candidato lo hace. Esto es una comprobación descriptiva negativa de esa especificación, no una intervención controlada en la dimensión rotatoria ni la rechazo de cada modelo de puntuación aprendido. Las configuraciones 14M y 31M se recuperaron desde los repositorios oficiales de modelos con sus revisiones resueltas; las otras cuatro son

configuraciones previamente archivadas. Estos archivos no certifican retroactivamente las revisiones de peso de los empleos de octubre.

Cuadro 6: Conteo de pares rotatorios versus el panel de consulta fija existente. El ansatz fija $a = 1$ y $\theta = 10^4$.

Pythia	d_{rot}	m	$\kappa = m/\log\theta$	QL [95 % CI]
14m	8	4	0.434294	1.244245 [1.212552, 1.281474]
31m	8	4	0.434294	1.232606 [1.202567, 1.264370]
70m	16	8	0.868589	1.173706 [1.147172, 1.199493]
1b	64	32	3.474356	0.974251 [0.943375, 1.006482]
1.4b	32	16	1.737178	0.930250 [0.896514, 0.966649]
2.8b	20	10	1.085736	0.866809 [0.830561, 0.904920]

El ansatz de puntaje y el candidato racional siguiente son construcciones diferentes. Con un coeficiente de alineación fijo, $\kappa = am/\log\theta$ depende del conteo de pares rotatorios y no tiene parámetro contextual; $\gamma_P(\theta, T)$ depende del contexto y omite ese conteo. No se identifica relación entre a, m, T que los conecte. La expresión racional es una función de configuración elegida, no una consecuencia de los pasos 2–3.

Una alternativa condicional de la integral coseno. Escribiendo $S_m(d) = \sum_{i=0}^{m-1} \cos(d\theta^{-i/m})$, su aproximación de Riemann izquierda en un valor fijo de $d, \theta > 1$ es

$$S_m(d) \approx \frac{m}{\log\theta} \int_{1/\theta}^{1} \frac{\cos(d\omega)}{\omega}\, \mathrm{d}\omega = \frac{m}{\log\theta}[\mathrm{Ci}(d) - \mathrm{Ci}(d/\theta)].$$

Para $1 \ll d \ll \theta$, el paréntesis es $\log\theta - \gamma_E - \log d + O(d^{-1} + (d/\theta)^2)$. Estas son dos aproximaciones distintas. El error del sumatorio de Riemann absoluto está acotado por la variación total, por lo tanto por $d(1 - 1/\theta)$; esta cota no necesita ser pequeña para un subespacio rotatorio finito. El error normalizado tiende a cero cuando $m \to \infty$ a distancia fija d, θ, no uniformemente sobre rangos de distancia arbitrariamente crecientes. En 400 distancias espaciadas geométricamente desde 10 hasta 2000, $m = 32$, $\theta = 10^4$, la pendiente del sumatorio ajustado es -3.480436 versus $-m/\log\theta = -3.474356$, mientras que el error puntual máximo frente a la integral es 6.868580. El acuerdo solo de las pendientes ajustadas no controla las oscilaciones individuales.

Si los productos punto entre planos pre-RoPE tienen una media común e independiente de la distancia a, las contribuciones seno tienen media nula, y otros componentes del puntaje tienen media constante, entonces la pendiente del puntaje medio es condicionalmente $-am/(\sqrt{d_{\text{head}}}\log\theta)$. Aquí a es un producto punto entre planos no escalado, a diferencia del coeficiente escalado por puntaje en el Paso 2. Medir estas amplitudes y la pendiente pre-softmax con consultas fijas es una prueba de las suposiciones. No conmuta la expectativa con softmax ni deriva el candidato racional.

Candidato adimensional. Definir $z = T\sqrt{2}/\theta$, usando un parámetro contextual elegido $T > 0$. El candidato lineal es

$$\gamma_{\text{linear}} = 1 - z = 1 - T\sqrt{2}/\theta. \tag{14}$$

El factor $\sqrt{2}$ es parte de esta parametrización. No se deriva por el recuento de ciclos ni por un cero afirmado de una suma de cosenos en $\theta/\sqrt{2}$. La escala real de medio recuento se discute en la Sección 4.3. Un exponente negativo es matemáticamente admisible en soporte finito; si se ajusta a una observación particular es una cuestión separada.

Proposición 5.1 (Obstáculo al razonamiento de suma de cosenos propuesto)**.** *Una suma coseno finita de la forma*

$$S_K(d) = \frac{1}{m}\sum_{i=0}^{K-1}\cos(d\omega_i), \qquad 1 \le K \le m, \tag{15}$$

no es automáticamente una distribución de atención no negativa normalizada. En particular, la propuesta de razón de logaritmos no puede suministrar el exponente de Padé:

$$\frac{\log r}{\log(1/r)} = -1 \neq \frac{2-z}{2+z}, \quad r > 0,\ r \neq 1,\ z > 0. \tag{16}$$

Demostración. Para $r > 0$, $\log(1/r) = -\log r$ y $r \neq 1$, el denominador no es cero. También $(2-z)/(2+z) > -1$ para $z > 0$. Una suma de un término coseno es negativa cuando su fase es π. Estas observaciones invalidan la normalización y la inferencia del exponente establecidas. Un argumento de fase estacionaria requeriría un parámetro asintótico explícito, hipótesis sobre la amplitud y un resto controlado; ninguna de estas se suministra por la suma finita sola. □

5.3 El candidato de Padé y su dominio

El candidato racional

$$\gamma_P(\theta, T) = \frac{2\theta - T\sqrt{2}}{2\theta + T\sqrt{2}} = \frac{2-z}{2+z}, \qquad z = \frac{T\sqrt{2}}{\theta}, \tag{17}$$

Es el $[1,1]$ Padé aproximante a e^{-z} en cero, ya que ambos tienen expansión $1 - z + z^2/2 + O(z^3)$. Este teorema de aproximación local no identifica un exponente observado con e^{-z}. Para $z > 0$, $-1 < \gamma_P < 1$ solo es positivo para $z < 2$, se anula en $T = \sqrt{2}\theta$ y enfoques -1 as $z \to \infty$. Por lo tanto, no puede predecir los exponentes observados superiores a uno. Sus derivadas describen la función candidata, no la respuesta de los pesos entrenados a una intervención.

Comparación de aproximaciones. Los siguientes son fórmulas candidatas distintas, no secuencias sucesivas de un orden validado de expansión de atención:

$$1, \qquad 1 - z, \qquad e^{-z}, \qquad \frac{2-z}{2+z}, \qquad \frac{1}{1+z}. \tag{18}$$

La última función es $[0,1]$, con expansión $1 - z + z^2 + O(z^3)$; no es una aproximación de orden superior que $[1,1]$. Para $z > 0$, su exceso sobre γ_P es $z^2/[(1+z)(2+z)] > 0$. Seleccionarlo después de observar que un exponente excede γ_P utiliza la clave para elegir el predictor y no establece una precisión prospectiva. Ni SWA ni GQA implican esta corrección. Una hipótesis de ventana de contexto debe utilizar el esquema de máscara real y la agregación de cabeza/layer.

Hipótesis de implementación. Una expresión racional puede ser económica de evaluar, pero ningún gate está probado como cercano a óptimo por ese hecho. La calidad de ruteo, la normalización, el costo de kernel sparso y el error de tarea requieren un benchmarking directo. El parámetro de exponente $z = T\sqrt{2}/\theta$ no debe silenciosamente convertirse en un argumento del kernel por distancia.

Una coordenada algebraica de base desplazada es

$$\theta_{\text{eff}} = \theta + T/\sqrt{2}. \tag{19}$$

Proporciona la identidad $\gamma_P = 1 - T\sqrt{2}/\theta_{\text{eff}}$ para esta definición. No es la base que produce el mismo exponente en $T = 0$, ya que cada base positiva da $\gamma_P = 1$ allí. Tampoco es la coordenada de base ajustada inversa del apartado 9; son definiciones diferentes y no pueden conjuntamente diagnosticar la historia de entrenamiento.

Cuadro 7: Fórmulas candidatas y su alcance matemático.

Candidato	Expresión	Relación con e^{-z}
Linear	$1 - z$	$[1, 0]$; $O(z^2)$
Referencia exponencial	e^{-z}	
Padé	$(2 - z)/(2 + z)$	$[1, 1]$; $O(z^3)$
Rational	$1/(1 + z)$	$[0, 1]$; $O(z^2)$
Abreviatura	$\log(10^4)/\log\theta$	no T dependencia

5.4 Comparación del Panel Inicial con Entradas Explícitas

Tabla 8 conserva los 18 entradas modelo/corpus de la comparación inicial del predictor, utilizando los exponentes y bases reproducidos realmente registrados en sus archivos JSON. Su conjunto es P01–P21 excluyendo P02, P03 y P21; incluye P19 y P20 a pesar de que faltan sus campos de base. Esto es diferente de las 18 filas almacenadas de ajuste en la Sección 8, que excluyen P02, P19 y P20. Cinco entradas no tienen registrada base y por lo tanto no tienen valor candidato calculado aquí. Para las restantes 13, $T = 2000$ es un parámetro de comparación establecido; la adquisición E4 en sí misma utilizó longitudes de secuencia dependientes de la distancia. Los campos de base almacenados no son una reconstrucción de configuraciones inmutables del modelo. Estas limitaciones impiden tratar esta tabla como un experimento arquitectónico controlado.

Estadística de caso seleccionado no reconstruida. La mediana anterior 4.3 % se atribuía a nueve casos "no anómalos" seleccionados. Ni una lista clara de membresía de nueve filas ni el proceso de selección y cálculo del denominador fue recuperada. El JSON de comparación que sobrevive Padé contiene cinco entradas y utiliza un denominador de predicción; no establece la estadística de nueve casos reclamada. Por lo tanto, retiramos ese número como resultado de precisión, en lugar de elegir nueve filas después de observar sus errores. Las 13 comparaciones identificables que aquí se presentan permanecen disponibles en su totalidad, con el denominador del exponente observado indicado anteriormente.

En estos 13 registros, el error absoluto medio es 0.167766, el error absoluto mediano es 0.124555, y el error relativo absoluto mediano es 19.2144 %, utilizando $|\widehat{\gamma}|$ en el denominador. Estos son errores descriptivos en un panel inicial mixto, no en un benchmark de muestra externa. Los porcentajes de casos seleccionados anteriormente no se transfieren

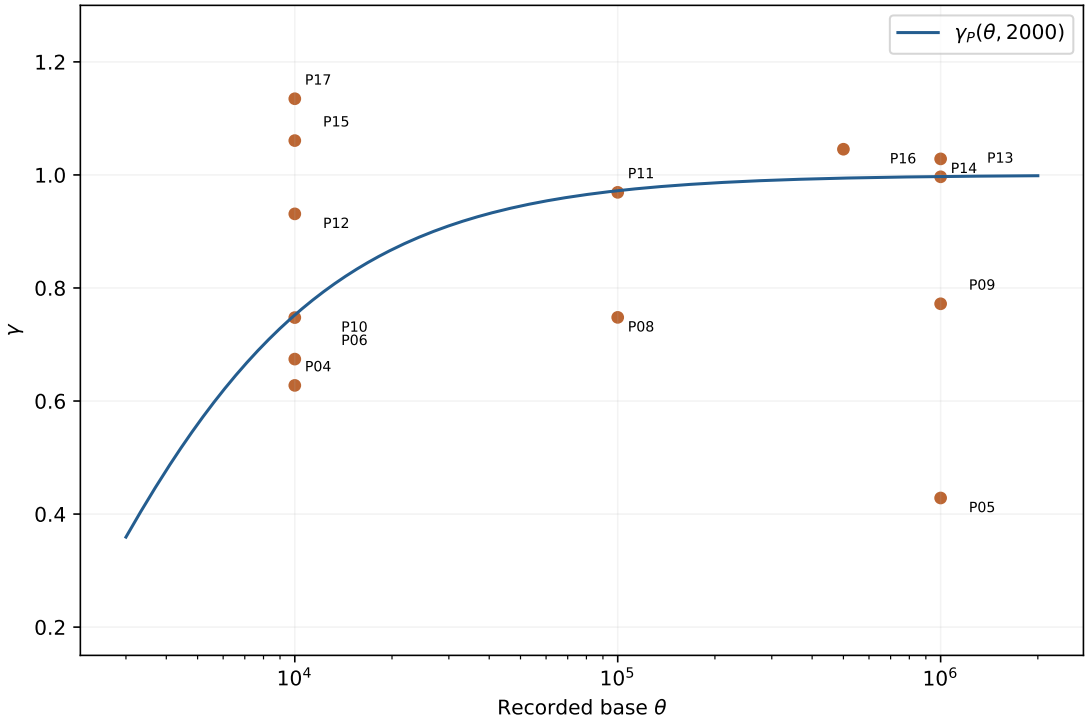


Figura 5: La curva candidata en especificado $T = 2000$ y 13 observaciones con campos base registrados. Los puntos utilizan los mismos IDs que la Tabla 8 Una tendencia en diferentes modelos no es una intervención θ.

Cuadro 8: Comparación inicial de predictores. Los IDs se refieren a la tabla de perfil completo. Un guion denota ausencia de registro de base, no una predicción cero. Residual es $\widehat{\gamma} - \gamma_P$ y no se le asigna una etiqueta causal.

ID	Modelo	θ	γ_P	$\widehat{\gamma}$	R^2	residual
P01	phi-3-mini (text)	—	—	0.6296	0.9853	—
P04	Gemma-2-9B (text)	10000	0.7522	0.6276	0.9773	-0.1246
P05	Mistral-Nemo-Instruct	1e+06	0.9972	0.4284	0.9224	-0.5687
P06	Pythia-2.8B	10000	0.7522	0.6742	0.9993	-0.0780
P07	Pythia-1.4B	—	—	0.7051	0.8413	—
P08	SmolLM2-135M	100000	0.9721	0.7480	0.9889	-0.2241
P09	Qwen2.5-3B	1e+06	0.9972	0.7720	0.9958	-0.2251
P10	Pythia-70M	10000	0.7522	0.7476	0.9843	-0.0046
P11	SmolLM2-360M	100000	0.9721	0.9692	0.9981	-0.0029
P12	Pythia-1B	10000	0.7522	0.9311	0.9831	+0.1789
P13	Qwen2.5-7B	1e+06	0.9972	0.9967	0.9939	-0.0005
P14	Qwen2.5-0.5B	1e+06	0.9972	1.0284	0.9968	+0.0312
P15	Mistral-7B-v0.1	10000	0.7522	1.0608	0.9987	+0.3085
P16	Llama-3-8B	500000	0.9944	1.0455	0.9975	+0.0511
P17	Gemma-2-9B (aleatorio)	10000	0.7522	1.1348	0.9765	+0.3826
P18	phi-3-mini (aleatorio)	—	—	1.0366	0.9779	—
P19	Llama-2-7B (Kaggle)	—	—	1.2108	0.9877	—
P20	Llama-2-13B (Kaggle)	—	—	1.0545	0.9943	—

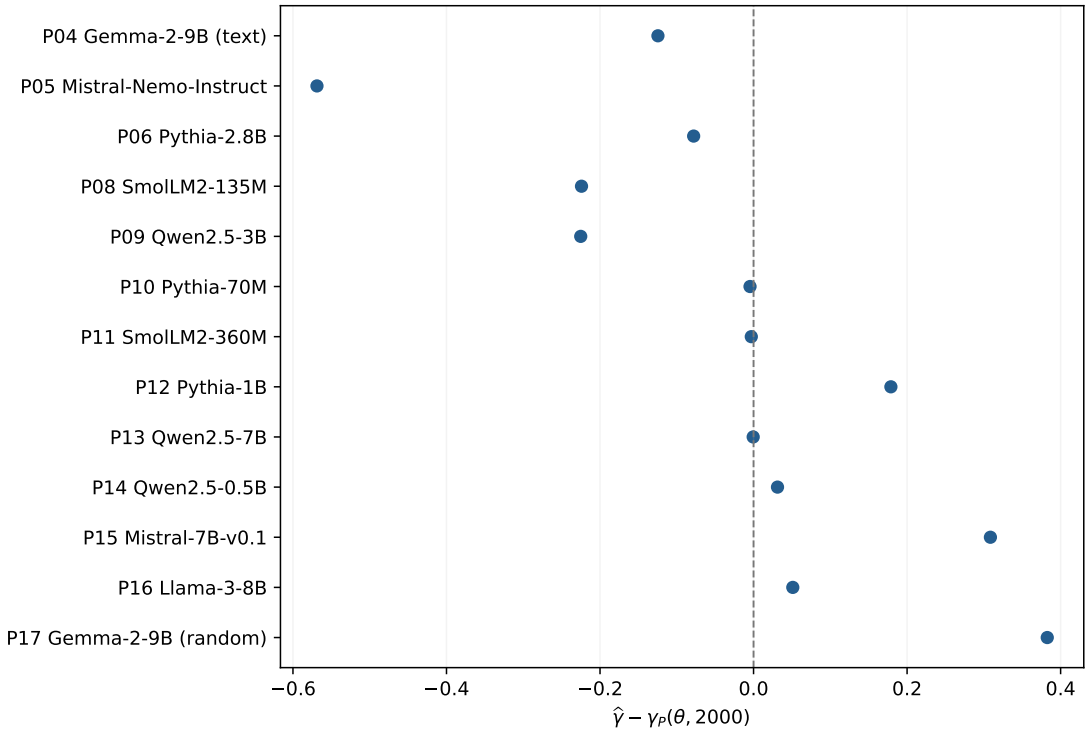


Figura 6: Todos los 13 residuos computables, sin excluir observaciones mal predichas. La falta de ajuste es una limitación del candidato; el residuo solo no identifica su causa.

a este denominador. En el reenvío de octubre aparte, solo un exponente histórico textual cae en su intervalo correspondiente entre 25 observaciones, y la comparación almacenada Padé tiene correlación de rango -0.445 en 19 filas (Sección 8).

Predicciones triviales en las mismas 13 filas. Los errores absolutos medios son 0.182656 para el constante uno, 0.177669 para la media del panel y 0.172106 para la mediana del panel, en comparación con 0.167766 para γ_P. El último mejora sobre la mediana solo en 0.004340 (alrededor de 2.52 %). La media y la mediana utilizan este mismo panel, por lo que esta es una comparación descriptiva en muestra, no un experimento de base con datos mantenidos fuera o una prueba de información predictiva nula. La utilidad prospectiva no ha sido demostrada. Añadir bases públicas actuales a los cinco campos faltantes sería un análisis de imputación separado; no recuperaría las configuraciones de esas ejecuciones históricas.

Comparación entre modelos base. Las cinco entradas siguientes conservan la comparación original en tres escalas nominales base. Varían en tamaño de modelo, entrenamiento, manejo de corpus y soporte; no son un barrido θ controlado. Una prueba genuina modificaría únicamente la programación de frecuencia en el mismo modelo bajo condiciones de medición coincidentes, y distinguiría los cambios en tiempo de inferencia de los reentrenamientos.

ID	θ	γ_P	$\widehat{\gamma}$
P10	10000	0.752201	0.747602
P06	10000	0.752201	0.674162
P11	100000	0.972110	0.969173
P14	1e+06	0.997176	1.028374
P13	1e+06	0.997176	0.996695

5.5 Regresión de Cinco Ejes Exploratoria

La especificación exploratoria original regresa el exponente del texto sobre cinco covariables:

$$\widehat{\gamma}_t = \beta_0 + \beta_P \gamma_P + \beta_c(\widehat{\gamma}_t - \widehat{\gamma}_r) + \beta_G \mathbf{1}_{\text{GQA}} + \beta_S \mathbf{1}_{\text{SWA}} + \beta_{\text{IH}} \mathbf{1}_{\text{post-IH}} + \varepsilon. \tag{20}$$

El script de origen proporciona banderas manuales de arquitectura, un parámetro $T = 100$ candidato, y una bandera "post-IH" derivada del tamaño; la última no es una etiqueta circuital medida. Reproducir su matriz de diseño de 23 filas reproduce los siguientes coeficientes. Esto es una comprobación de la especificación original de regresión, no un respaldo de sus metadatos o suposiciones de identificación.

Coeficiente	Estimación
β_0	+4.578763
β_P	-3.686409
β_c	+0.597427
β_G	+0.107691
β_S	+0.001117
β_{IH}	-0.151244

El ajuste tiene $R^2 = 0.437821$, ajustado $R^2 = 0.272474$, y desviación estándar residual 0.164644. La covariada del corpus contiene la respuesta $\widehat{\gamma}_t$ misma, por lo que la regresión no es un predictor solo de configuración y su aparente poder explicativo puede incluir filtraje hacia el objetivo. La columna de Padé casi constante también está casi colineal con la intersección. La razón de la varianza del predictor sobre la varianza observada no es la varianza explicada; la varianza no explicada no puede asignarse automáticamente al otro eje.

Los residuos más grandes según la especificación original son aproximadamente +.47 (Pythia-31M), −.33 (SmolLM2-135M), −.26 (OLMo-7B), −.19 (Pythia-2.8B) y −.18 (Llama-2-7B). Estos motivan pruebas sobre tokenización, datos, optimización y soporte de medición, pero no identifican estos factores. En particular, el corpus Pythia-2.8B con soporte desigual requiere la corrección en la Sección 8.4.

Una hipótesis útil sigue siendo una notación relevante

$$\gamma_{\text{obs}} = \gamma_P + \delta_{\text{data}} + \delta_{\text{arch}} + \delta_{\text{circuit}} + \delta_{\text{train}} + \varepsilon. \tag{21}$$

Sin intervenciones ni restricciones identificables, infinitas asignaciones de los términos δ dan el mismo residuo. La descomposición no establece una cota geométrica, una mediana fija, ni que las perturbaciones expliquen 99 % de varianza. Los exponentes medidos por encima de uno ya exceden el rango positivo-z de γ_P.

5.6 Cantidades Derivadas y Consecuencias Verificables

P1: entropía condicional sobre la familia. Para la familia finita, la identidad es

$$H_\gamma(N) = \log Z_N + \gamma U_N, \qquad U_N = -\partial_\gamma \log Z_N. \tag{22}$$

Las columnas de entropía en el conjunto de datos termodinámico inicial se derivaron de exponentes ajustados; no son observaciones independientes de entropía. Una prueba debe medir, en su lugar, una distribución empírica normalizada en el mismo soporte y comparar su entropía con el valor de la familia ajustada, teniendo en cuenta la incertidumbre de estimación. Normalizar los valores en siete distancias muestreadas da la entropía de una variable de siete puntos, la cual no puede ser directamente igualada con la entropía sobre 2000 distancias de tokens.

P2: conteo efectivo basado en la entropía. La exponencial de la entropía de Shannon es

$$L_{\text{eff}} := N_{\text{eff}} := e^{H_\gamma(N)} = \prod_{d=1}^{N} p_\gamma(d)^{-p_\gamma(d)} = Z_N \prod_{d=1}^{N} d^{\gamma p_\gamma(d)}. \tag{23}$$

El producto utiliza pesos de probabilidad $p_\gamma(d)$, no pesos uniformes $1/N$. Para $N = 2000$, sumas finitas exactas dan $N_{\text{eff}} = 1160.379$ en $\gamma = .65$ y 362.107 en $.95$. Para $0 < \gamma < 1$ fijo, $N_{\text{eff}}/N \to e^{-\gamma/(1-\gamma)}/(1-\gamma)$; para $\gamma > 1$ fijo, N_{eff} tiende a una constante finita y $N_{\text{eff}}/N \to 0$. Estas son medidas de diversidad de la distribución de distancia elegida. No limitan los espacios de razonamiento multi-hop, la facticidad o el número de tokens útiles en una red entrenada.

P3: variación de capa con ajustes faltantes reportados. Recalculamos las medias de capa a partir de las entradas finitas de las matrices de exponentes por cabeza guardadas, excluimos efectivamente las primeras y últimas capas de la arquitectura y reportamos la desviación estándar poblacional sobre las medias de capa observadas restantes. Las capas internas completamente faltantes permanecen como faltantes; no se reclasifican como capas de frontera. Algunas capas observadas también carecen de ajustes individuales por cabeza, como se registra en el manifiesto de evidencia.

Modelo	observó	ausente	media	desviación estándar	coeficiente de variación
pythia-70m	3	1	1.424844	0.248423	0.174351
pythia-410m	14	8	1.309708	0.175708	0.134158
pythia-1b	14	0	0.915502	0.137759	0.150474
pythia-1.4b	22	0	0.900704	0.231879	0.257442
pythia-2.8b	30	0	1.326845	0.257319	0.193933

Cuatro de cinco resúmenes de casos disponibles tienen CV $< .20$, y solo uno tiene CV $< .15$. Pythia-1.4B tiene CV $= .257442$. Los datos no establecen una cota universal de estabilidad de capas ni una causa geométrica. En particular, las matrices parciales no pueden certificar la estabilidad del modelo completo. El procedimiento anterior eliminó la última capa observada después de filtrar las capas faltantes, lo cual no necesariamente elimina la frontera arquitectónica.

P4: residuos a corto y largo alcance. La planificación a corto alcance y la decaimiento más rápido a largo alcance son patrones candidatos verificables. No son consecuencias de "todos los planos no ciclados por debajo de $2\pi\sqrt{\theta}$", lo cual es falso: el primer umbral es 2π. La longitud del entrenamiento tampoco fuerza que cada fase aliasée uniformemente. El registro extendido a largo alcance guardado de BLOOM-1b7 contiene un cero numérico en $d = 5000$, con valores positivos no monótonos en 2048 y 3000. Un cero numérico en un experimento ALiBi no es un teorema sobre RoPE ni una prueba de un corte universal duro. Probar los signos de los residuos requiere un intervalo de ajuste predeclarado, el contexto real de entrenamiento y observaciones positivas adecuadas en cada región.

Líneas de base nulas en los ajustes guardados disponibles. Comparamos regresiones de ley de potencia y exponencial en el espacio log-atención en cada rejilla de distancia declarada por cada registro, utilizando textos canónicos/ archivos aleatorios con al menos tres puntos positivos y un ajuste almacenado. El conjunto resultante tiene 74 perfiles, de los cuales 72 favorecen la ley de potencia según la diferencia en muestra R^2; la diferencia mediana es $+.203803$. La lista manifiesta incluye cada registro incluido. Este es un modelo comparativo de forma descriptivo, con rangos desiguales y sin evaluación de validación cruzada. No descarta mezclas, mecanismos dependientes del contenido o explicaciones basadas únicamente en aprendizaje, y no demuestra una restricción de MaxEnt. La correlación entre arquitecturas no descarta el entrenamiento o los datos como causa.

6 ¿Por qué la Ley de Potencia: Fundamento de la Entropía Máxima

La familia finita de ley de potencia se deriva de un problema especificado de maximización de entropía. La restricción y la medida de referencia son premisas de dicho problema. La geometría frecuencial de RoPE no establece que una distribución de atención entrenada resuelva ese problema. Esta sección proporciona el resultado matemático condicional y su conexión con perfiles medidos.

Teorema 6.1 (Selección MaxEnt en un momento prescrito de la familia)**.** *Sean $N \geq 1$ un entero y $\gamma \in \mathbb{R}$. Definamos*

$$p_\gamma(d) = \frac{d^{-\gamma}}{Z_N(\gamma)}, \qquad Z_N(\gamma) = \sum_{j=1}^{N} j^{-\gamma}, \qquad d \in \{1, \ldots, N\}.$$

Entre todos los vectores de probabilidad p en este soporte que satisfacen $\mathbb{E}_p[\log d] = \mathbb{E}_{p_\gamma}[\log d]$, el único maximizador de entropía es p_γ. Utilizamos medida de conteo, logaritmos naturales y $0 \log 0 = 0$.

Demostración. El vector de referencia es estrictamente positivo. Dado que $\log p_\gamma(d) = -\gamma \log d - \log Z_N$, las restricciones de momento y normalización dan $\sum p \log p_\gamma = \sum p_\gamma \log p_\gamma$. Consecuentemente,

$$H(p_\gamma) - H(p) = \sum_{d=1}^{N} p(d) \log \frac{p(d)}{p_\gamma(d)} \geq 0.$$

La desigualdad de Gibbs también se aplica cuando algunos elementos de p se anulan. La igualdad se cumple solo para $p = p_\gamma$. Esto demuestra la optimalidad global y la unicidad, en lugar de solo un punto estacionario de un lagrangiano. □

Qué selecciona la restricción y qué queda abierto

Las distribuciones exponenciales, gaussianas o con corte sobre un conjunto finito pueden tener el mismo momento de distancia logarítmica. No están excluidas como modelos empíricos de atención. El teorema destaca al maximizador de entropía entre las distribuciones factibles. Las familias de perfiles en competencia deben compararse en las mismas observaciones, soporte y pérdida; no se requiere diferenciabilidad en distancia sobre un soporte discreto.

Ejemplo: un dado de seis caras. Si la única estadística prescrita es la media de log(face), elegir la entropía máxima asigna $p(k) \propto k^{-\gamma}$ en $\{1, \dots, 6\}$. Esto es una regla de inferencia bajo información limitada. No establece que un dado cargado físico, o un transformador entrenado, deba realizar esa regla.

6.1 El problema de MaxEnt

Teorema 6.2 (Entropía máxima con soporte finito [9, 18])**.** *Para enteros $N > 1$ y $0 < \lambda < \log N$, consider*

$$\max_{p \geq 0} \; H(p) = -\sum_{d=1}^{N} p(d) \log p(d), \tag{24}$$

$$\sum_{d=1}^{N} p(d) = 1, \qquad \sum_{d=1}^{N} p(d) \log d = \lambda. \tag{25}$$

Existe un único exponente real y una única distribución que maximiza:

$$p^*(d) = \frac{d^{-\gamma}}{Z_N(\gamma)}, \qquad -\partial_\gamma \log Z_N(\gamma) = \lambda. \tag{26}$$

El exponente es positivo, cero o negativo según sea que λ esté por debajo, igual o por encima de $\log(N!)/N$. En los puntos extremos $\lambda = 0$ y $\lambda = \log N$, la distribución factible es la masa puntual en 1 o N, respectivamente, obtenida como límite de un exponente. Para $N = 1$ la distribución es única pero el exponente no es identificable.

Demostración. La función lagrangiana $\mathcal{L} = H(p) - \mu(\sum p - 1) - \gamma(\sum p \log d - \lambda)$ tiene ecuación estacionaria $-\log p(d) - 1 - \mu - \gamma \log d = 0$. La normalización da lugar a la familia mostrada. Para establecer la existencia del exponente requerido, defínase $U_N(\gamma) = \mathbb{E}_{p_\gamma}[\log d]$. La diferenciación de la suma finita da

$$U_N'(\gamma) = -\operatorname{Var}_{p_\gamma}(\log d) < 0, \qquad U_N(0) = \frac{\log(N!)}{N}.$$

La negatividad estricta sigue porque todas las probabilidades son positivas y $\log d$ es no constante para $N > 1$. Como $\gamma \to +\infty$, la masa se concentra en 1 y $U_N \to 0$; como $\gamma \to -\infty$, se concentra en N y $U_N \to \log N$. La continuidad y la monotonicidad estricta establecen la existencia, la unicidad y la condición de signo. El argumento de Gibbs anterior entonces prueba la optimalidad global sobre todos los vectores factibles, incluyendo sus fronteras. □

Interpretation El teorema es un resultado de inferencia restringida, no un resultado de dinámicas de entrenamiento. Su premisa debe justificarse por separado cuando la familia se aplica a la atención. Los modelos con configuración idéntica de RoPE pueden aprender diferentes consultas y claves y producir diferentes perfiles. El panel exploratorio de cinco ejes en la Sección 5.5 reporta una desviación estándar residual de la población 0.164644 (approximately 0.16 En su regresión múltiple exploratoria, no una varianza medida en configuración fija; ni esa regresión ni un centroide asumido convierten la premisa de MaxEnt en una necesidad geométrica.

6.2 Interpretación termodinámica

La familia puede escribirse en forma de Gibbs:

$$p_\gamma(d) = \frac{e^{-\gamma \log d}}{Z_N(\gamma)}, \qquad H(d) = \log d, \qquad T_{\text{attn}} = 1/\gamma \quad (\gamma \neq 0). \tag{27}$$

Este es un diccionario matemático para una familia elegida de probabilidades. La coordenada de temperatura inversa no es la temperatura de muestreo de salida, y el diccionario en sí no identifica un proceso de equilibrio físico o la respuesta de un modelo entrenado a una intervención.

Cuadro 9: Coordenadas termodinámicas de la familia finita y sus dominios.

Cantidad	Fórmula	Alcance
Inversa de temperatura	γ	Parámetro de familia, no una temperatura física medida.
Temperatura	$1/\gamma$	definida para $\gamma \neq 0$; un aumento positivo γ favorece distancias más cortas.
Función de partición	$\sum_{d=1}^{N} d^{-\gamma}$	Finita para cada N; la suma infinita converge solo para $\gamma > 1$.
Capacidad calorífica	$-\gamma^2 U_N'(\gamma)$	Una susceptibilidad de familia finita, no sensibilidad de tarea.

Para $S = \log Z_N + \gamma U_N$, la diferenciación cancela los dos términos que involucran U_N solamente. Dado que $U_N' < 0$ para $N > 1$, la parametrización da

$$\frac{\mathrm{d}S}{\mathrm{d}\gamma} = \gamma \frac{\mathrm{d}U_N}{\mathrm{d}\gamma}, \qquad \frac{\mathrm{d}S}{\mathrm{d}U_N} = \gamma \quad (N > 1).$$

6.3 Dualidad con la Ley de Zipf

La ley del vocabulario de Zipf [54] y una ley de potencia a distancia pueden representarse ambos mediante familias de entropía restringida. Para dos miembros de la misma familia finita de ley de potencia con exponente entero $N > 1$ La estricta monotonía de su logaritmo de momento identifica exponentes iguales. Esto requiere el mismo soporte y medida de referencia, así como la suposición de familia.

$$\mathbb{E}[\log \text{rank}] = \mathbb{E}[\log d] \quad \Longrightarrow \quad \alpha = \gamma \quad \text{same family, support and measure; } N > 1. \tag{28}$$

Los rangos de vocabulario y las distancias de token generalmente tienen soportes diferentes. Las siete parejas guardadas con éxito dan $r(\alpha_{\text{Zipf}}, \gamma) = -0.505314$. No se establece una correlación positiva corregida. Las tres filas de Pythia tienen exponente constante de rango, por lo tanto su correlación de Pearson dentro del grupo es indefinida (Apéndice D).

6.4 Funciones Termodinámicas Finitas

Group 5 en el Apéndice K Proporciona las afirmaciones completas de suavidad, entropía y divergencia para esta familia finita, junto con sus demostraciones.

Las cantidades discretas se definen mediante sumas exactas:

$$Z_N(\gamma) = \sum_{d=1}^{N} d^{-\gamma}, \qquad Z_N(1) = H_N, \tag{29}$$

$$U_N(\gamma) = \frac{\sum_{d=1}^{N} d^{-\gamma} \log d}{Z_N(\gamma)}, \tag{30}$$

$$C_V(\gamma, N) = -\gamma^2 \partial_\gamma U_N = \gamma^2 \operatorname{Var}_{p_\gamma}(\log d). \tag{31}$$

Por separado, con medida de Lebesgue en $[1, N]$ el normalizador continuo es

$$Z_c(\gamma, N) = \begin{cases} (N^{1-\gamma} - 1)/(1-\gamma), & \gamma \neq 1, \\ \log N, & \gamma = 1. \end{cases}$$

Los momentos continuos y discretos no deben mezclarse. Un ajuste de cola en $[16, T/2]$ no establece esta familia en todo $[1, N]$. El soporte y si una cantidad es medida o calculada a partir de un exponente acompañan cada aplicación.

Ejemplos numéricos derivados Para $N = 2000$, con los exponentes reproducidos en alta precisión para P10, P06, P13, P16 y P17, respectivamente, las capacidades caloríficas de la familia son 2.5396, 1.8335, 5.4488, 5.9361 y 6.5951. Son las mismas entradas y la misma convención de redondeo de la Sección 8. El máximo discreto se halla cerca de $\gamma = 1.256885$, con $C_V = 6.919812$. Son cálculos de la familia prescrita, no momentos medidos en filas de atención del modelo.

7 Familias de Potencias Finitas con Ley de Potencia y Límites con Soporte Infinito

7.1 La Función de Partición y Estructura de Fase

Para un exponente real fijo, el límite de soporte infinito del normalizador discreto distingue tres regímenes de sumabilidad:

$$\gamma < 1: \quad Z_N(\gamma) \sim N^{1-\gamma}/(1-\gamma) \to \infty, \tag{32}$$

$$\gamma = 1: \quad Z_N(1) = H_N \sim \log N \to \infty, \tag{33}$$

$$\gamma > 1: \quad Z_N(\gamma) \to \zeta(\gamma) < \infty. \tag{34}$$

Cada distribución finita-N permanece normalizada. La divergencia de la suma infinita no normalizada significa que los mismos pesos no pueden normalizarse en todos los enteros positivos; no significa que un modelo tenga masa de atención infinita o capacidad de recuperación no acotada. Para $\gamma > 1$ la distribución límite aún tiene una cola infinita, no un horizonte finito rígido. El término Hagedorn [14] denota una analogía a una frontera de sumabilidad, no una identificación con una transición de fase física.

Analogía y su límite. Un cambio de régimen puede representarse como una frontera de fase. A diferencia de un sistema de ebullición o congelación, sin embargo, las sumas finitas aquí son funciones analíticas de γ. El límite matemático $N \to \infty$ debe distinguirse de una intervención en una red de contexto finito.

7.2 Capacidad térmica continua y discreta en uno

Teorema 7.1 (Varianza log-distancia continua)**.** *Sea $N > 1$ y sea $p_c(x) = 1/(x \log N)$ una densidad de probabilidad sobre $[1, N]$ con respecto a la medida de Lebesgue. Entonces*

$$C_{V,c}(1, N) = \mathrm{Var}_{p_c}(\log x) = \frac{(\log N)^2}{12}. \tag{35}$$

Demostración. Escribe $L = \log N$. El cambio de variable $y = \log x$, con $\mathrm{d}y = \mathrm{d}x/x$, mapea la densidad a la distribución uniforme en $[0, L]$. Integración directa da

$$\mathbb{E}_{p_c}[\log x] = \frac{\log N}{2}, \qquad \mathbb{E}_{p_c}[(\log x)^2] = \frac{(\log N)^2}{3}.$$

Restando el cuadrado del primer momento del segundo se obtiene $L^2/3 - L^2/4 = L^2/12$. Este cálculo se refiere a la densidad continua, no a probabilidades armónicas en distancias enteras. □

Expansión continua equivalente Para $\epsilon = 1 - \gamma$ y L fijo, $Z_c = (e^{\epsilon L} - 1)/\epsilon$ tiene una singularidad removible en cero. Su expansión analítica da

$$\log Z_c = \log L + \frac{\epsilon L}{2} + \frac{\epsilon^2 L^2}{24} + O(\epsilon^3), \qquad U_c = \frac{L}{2} + \frac{\epsilon L^2}{12} + O(\epsilon^2).$$

Tomar la derivada en $\epsilon = 0$ recupera la misma varianza. Los términos restantes son esenciales: una expansión truncada no es una expresión exacta para U_c lejos de uno, y un normalizador de primer orden es insuficiente para su segunda derivada.

Para distancias de tokens discretas, el valor exacto en su lugar es

$$C_V(1, N) = \frac{\sum_{d=1}^{N} (\log d)^2/d}{H_N} - \left(\frac{\sum_{d=1}^{N} \log d/d}{H_N} \right)^2.$$

En $N = 2000$, el valor continuo es 4.81448, mientras que el valor discreto es 5.48412. El cálculo de suma finita coloca el máximo discreto cerca de $\gamma = 1.256885$, con $C_V = 6.919812$. Ambos permanecen finitos. Estos números son cálculos derivados de una familia, no mediciones independientes sobre modelos de lenguaje.

7.3 Información de Fisher y Capacidad Calorífica

En soporte finito fijo $\partial_\gamma \log p_\gamma(d) = U_N - \log d$ Por lo tanto, elevar al cuadrado y promediar da

$$I(\gamma) = \mathbb{E}_{p_\gamma}[(\partial_\gamma \log p_\gamma)^2] = \mathrm{Var}_{p_\gamma}(\log d) = \partial_\gamma^2 \log Z_N(\gamma). \tag{36}$$

$$C_V(\gamma) = \gamma^2 I(\gamma), \qquad I(\gamma) = C_V(\gamma)/\gamma^2 \quad (\gamma \neq 0). \tag{37}$$

En $\gamma = 0$, I se evalúa directamente como una varianza; la división por γ^2 no está definida. Ninguna de estas cantidades es singular en N finito. Sus extremos no necesitan coincidir debido a $C_V' = 2\gamma I + \gamma^2 I'$. Para $N > 1$, la longitud de Fisher intrínseca dentro de esta familia de un parámetro es

$$d_{\mathrm{FR}}(\gamma_1, \gamma_2) = \left| \int_{\gamma_1}^{\gamma_2} \sqrt{I(\gamma)}\, \mathrm{d}\gamma \right|. \tag{38}$$

La integral utiliza $\sqrt{I}$, no I. Este es una distancia a lo largo de la familia estadística especificada; no es automáticamente la geodésica del simplex ambiental ni una distancia entre redes entrenadas con protocolos diferentes.

7.4 Límite de Entropía Finita y Regímenes Asintóticos

Teorema 7.2 (Entropía en soporte de distancia finita)**.** *Para entero $N > 1$ y la familia discreta p_γ, define $S_\gamma(N) = -\sum p_\gamma \log p_\gamma$.*

(I) Para cada real γ, $0 \le S_\gamma(N) \le \log N$ con igualdad en el límite superior exactamente en $\gamma = 0$. For $N = 1$ La entropía es cero para cada exponente.

(II) Para fijado $0 < \gamma < 1$, as $N \to \infty$,

$$S_\gamma(N) = \log N - \log(1-\gamma) - \frac{\gamma}{1-\gamma} + O(N^{\gamma-1} \log N). \tag{39}$$

(III) A $\gamma = 1$, $S_1(N) = \frac{1}{2}\log N + \log\log N - \gamma_E/2 + o(1)$ Este es un régimen asintótico separado, no una sustitución en la fórmula anterior con exponente fijo.

(IV) Para fijado $\gamma > 1$, $S_\gamma(N) \to \log \zeta(\gamma) - \gamma\zeta'(\gamma)/\zeta(\gamma) < \infty$.

Demostración. La primera parte es la cota de Shannon para alfabeto finito, con igualdad solo en el vector uniforme. Para $N > 1$, el vector de ley de potencia es uniforme exactamente cuando $\gamma = 0$. Para la segunda parte, la comparación integral (o Euler–Maclaurin con exponente fijo) da $Z_N = N^{1-\gamma}/(1-\gamma) + O(1)$ y $U_N = \log N - 1/(1-\gamma) + O(N^{\gamma-1}\log N)$. Sustituye en $S = \log Z_N + \gamma U_N$. A uno, $H_N = \log N + \gamma_E + O(N^{-1})$ y $\sum_{d \le N} \log d/d = \frac{1}{2}(\log N)^2 + O(1)$ dan la expansión separada indicada. Por encima de uno, ambos $\sum d^{-\gamma}$ y $\sum d^{-\gamma}\log d$ convergen, dando la expresión zeta. Las cotas del resto con exponente fijo no son uniformes como $\gamma \to 1$. □

Cantidades del modelo-panel calculadas Para las 21 entradas del modelo/corpus del panel inicial en la Sección 8 Entropías exactas de familias finitas $N = 2000$ varían desde 4.377455 to 7.477491 Cada valor se calcula a partir del exponente correspondiente reproducido, no midido independientemente de la atención. Su asociación con γ No se puede validar el teorema del alfabeto finito o un mecanismo físico. La tabla reporta los exponentes de entrada y los valores derivados juntos.

Ámbito de la analogía holográfica La relación Ryu–Takayanagi [38] se refiere a la entropía de entrelazamiento y áreas geométricas en su contexto físico establecido. Una cota logarítmica para una variable de distancia establece ni ese contexto ni una carga central $c = 3$. Igualar su coeficiente principal a una fórmula de CFT es insuficiente. Del mismo

modo, la cota no genera presión termodinámica que redistribuya coordenadas rotatorias al contenido semántico. Tal interpretación de canal requiere evidencia funcional separada. El trabajo sobre el aprendizaje de mapas holográficos inversos [21] se refiere a un problema de aprendizaje diferente y no identifica el operador o la geometría de los modelos estudiados aquí.

Exponente de umbral de conteo $1/\gamma$ Para la familia normalizada, un umbral de probabilidad tiene el conteo de soporte finito

$$\#\{d \in \{1,\dots,N\} : p_\gamma(d) \geq \epsilon\} = \min\{N, \lfloor (\epsilon Z_N)^{-1/\gamma} \rfloor\}, \qquad \gamma > 0, \quad \epsilon > 0.$$

$$d_{\text{eff}} := 1/\gamma \quad (\gamma > 0). \tag{40}$$

La notación d_{eff} registra un exponente de escalado de umbral, no una dimensión de Hausdorff o de conteo de cajas de los estados del modelo. Los pesos no normalizados $d^{-\gamma}$ dan $\lfloor \epsilon^{-1/\gamma} \rfloor$ antes de recortarlos a N. Los ejemplos $\gamma = .674$ y 1.045 producen $1/\gamma \approx 1.48$ y 0.96, respectivamente; en uno de ellos el recíproco es uno. Estos valores aritméticos no demuestran un soporte fractal, una addressabilidad de medida cero o la exclusión de arquitecturas hiperbólicas como Hypformer [48]. Medir una dimensión geométrica del soporte por cabeza es un experimento propuesto por separado.

Medidas de ablación de banda en EXP-B3 La ventana completa guardada y el análisis exacto de la máscara están en la Sección 13, Tabla 24. Los grupos históricos son máscaras de coordenadas adyacentes desiguales en lugar de dos conjuntos de planos rotatorios intactos. La máscara más grande tiene un PPL mayor en cada ventana guardada. Ni la cota de entropía finita ni esas ablaciones establecen un reservorio semántico ni una disociación doble semántica/posicional. El esquema ha sido reemplazado por las curvas registradas de PPL en la Figura 13.

7.5 La familia con corte exponencial

Una segunda familia añade un coeficiente de corte $\lambda \geq 0$; aquí λ es distinto del momento logarítmico prescrito en la Sección 6. Su normalizador de soporte infinito, cuando converge, es

$$Z_\infty(\gamma, \lambda) = \text{Li}_\gamma(e^{-\lambda}) = \sum_{d=1}^{\infty} d^{-\gamma} e^{-\lambda d}. \tag{41}$$

Definición 7.3 (Regímenes de sumabilidad)**.** At $\lambda = 0, \gamma = 1$, $Z_N = H_N \sim \log N$ Los demás casos son

$$\lambda = 0,\ \gamma < 1: \quad Z_N \sim N^{1-\gamma}/(1-\gamma) \to \infty, \tag{42}$$

$$\lambda = 0,\ \gamma > 1: \quad Z_N \to \zeta(\gamma) < \infty, \tag{43}$$

$$\lambda > 0: \quad Z_\infty < \infty \quad (\gamma \in \mathbb{R}). \tag{44}$$

Estos son casos distribucionales, no una partición automática de arquitecturas. Una bias de ALiBi puede motivar un factor posicional exponencial, pero también contribuyen los puntajes de contenido aprendidos. Las distancias medias reportadas de 19 tokens para BLOOM y 772 para Falcon permanecen observaciones de panel; no son impuestas solo por el normalizador.

Regresión de tamaño exploratoria La expresión anteriormente impresa se mantiene solo para identificar la reclamación retirada:

$$\gamma_{\text{old}}(N_{\text{params}}) = 11.57\, N_{\text{params}}^{-0.079} \quad \text{(withdrawn fit)}. \tag{45}$$

Con N_{params} en parámetros individuales, se obtiene 3.153474 en 14M y 1.930059 en 7B, incompatibles con el panel que se dijo describir. Mantener el coeficiente fijo mientras se expresan las cuentas en millones o billones hace que estas predicciones sean mayores, no menores. No se recuperó ninguna normalización, ya sea explícita o alternativa declarada, para corregir el coeficiente o el exponente; el acompañante $R^2 = 0.41$ se retira por lo tanto también. El script E9 disponible se refiere a ocho tamaños de Pythia, y su informe de log guardado reporta cero observaciones parseadas, no un ajuste exitoso de 16 modelos. Un nuevo ajuste a un panel explícito sería un nuevo análisis y no se sustituye silenciosamente aquí.

8 Exponentes Específicos de Protocolo y Coordenadas Derivadas de la Familia

Esta sección distingue perfiles de distancia medida, estimaciones de regresión y cantidades calculadas a partir de una distribución ajustada. Los registros iniciales E4 y el atlas QL de octubre utilizan protocolos diferentes de adquisición y agregación. Cada uno se analiza en su propio soporte. El acuerdo entre ellos es una cuestión empírica, no una consecuencia de las identidades de la familia finita.

8.1 Panel Inicial E4: Datos y Protocolo de Regresión

La nula uniforme correspondiente a este extractor no es plana. Las 21 filas originales que aparecen a continuación permanecen como observables registrados. Tabla 10 ahora muestra la nula uniforme y la pendiente excesiva en cada una de las grillas propias de la fila, también proporcionadas en el archivo CSV adjunto. Estos son diagnósticos del instrumento, no mediciones nuevas o una descomposición no sesgada en efectos aprendidos y efectos geométricos. Véase la Sección 9.1 para la prueba del instrumento y sus limitaciones.

Los registros guardados contienen la atención media a distancias muestreadas, generalmente con tres medias de semilla. Para una cuadrícula declarada de ajuste D, regresamos $\log A(d)$ sobre $\log d$ con una intersección y peso igual por distancia. La pendiente con signo invertido es $\widehat{\gamma}$. Utilizamos la cuadrícula guardada exacta cuando está disponible. Para el registro parcial Falcon y los dos registros Kaggle Llama-2, que no contienen cuadrícula de ajuste, el presente análisis utiliza explícitamente todas las distancias guardadas $d \geq 30$. Estas son estimaciones recalculadas a partir de esos agregados; no reconstruyen un ajuste original no documentado.

El programa de medición de origen utiliza secuencias de longitud $d + 50$ en cada distancia, en lugar de un contexto fijo para toda la curva. Su modo de texto se basa en el corpus de episodios de MongoDB y su modo aleatorio genera secuencias de tokens. Los registros completos suelen listar semillas 42, 123 y 7 y un conteo nominal de prompts de 150. El programa puede utilizar un conteo diferente de prompts a grandes distancias, pero esa opción no se registra en el JSON final. En consecuencia, 150 no debe afirmarse como el conteo verificado en cada distancia ni como 150 prompts independientes agrupados en

todos los semillas. Los agregados guardados no establecen la instantánea exacta del corpus, la revisión inmutable de pesos, la precisión de carga o la correspondencia de prompts entre corpora.

Para las 18 filas con una estimación de ley de potencia almacenada, el exponente reproducido coincide dentro de 8×10^{-16}. Esto verifica la aritmética de regresión en las entradas guardadas; no valida la instrumentación de atención ni la adquisición de esas entradas. No se realiza ninguna nueva inferencia del modelo. La evidencia acompañante manifiesta mapas P01–P21 a rutas de entrada y hashes SHA-256, preserva los valores impresos previos y registra la sensibilidad a cortes de distancia inferiores 10, 30 y 50.

8.2 Panel Inicial Completo: 21 Entrada/s Modelo/Corpus

Cuadro 10: Los exponentes se reproducen a partir de los agregados guardados E4. La columna rango proporciona $[d_{\text{mín}}, d_{\text{máx}}]$ y el número n de distancias ajustadas; no es una estimación del tamaño de muestra para prompts independientes. P02, P19 y P20 utilizan la cuadrícula de auditoría explícita $d \geq 30$. Las entradas de Kaggle están etiquetadas según su plataforma registrada, sin inferir precisión del nombre del archivo. Todas las entradas son texto excepto las dos marcadas como aleatorias.

ID	Modelo	$\widehat{\gamma}$	R^2	Rango / n	$\widehat{\gamma}_0$	exceso
P01	phi-3-mini (text)	0.629563	0.985314	30–1000 / 6	0.741726	-0.112163
P02	Falcon-7B	0.924685	0.989728	30–1000 / 6	0.741726	+0.182958
P03	CodeLlama-13B	0.382302	0.928789	30–2000 / 7	0.782582	-0.400280
P04	Gemma-2-9B (text)	0.627646	0.977314	30–2000 / 7	0.782582	-0.154936
P05	Mistral-Nemo-Instruct	0.428448	0.922354	20–1000 / 7	0.698299	-0.269851
P06	Pythia-2.8B	0.674162	0.999287	30–100 / 3	0.525737	+0.148425
P07	Pythia-1.4B	0.705073	0.841258	30–2000 / 7	0.782582	-0.077510
P08	SmolLM2-135M	0.747996	0.988937	30–2000 / 7	0.782582	-0.034586
P09	Qwen2.5-3B	0.772033	0.995841	30–2000 / 7	0.782582	-0.010549
P10	Pythia-70M	0.747602	0.984269	30–2000 / 7	0.782582	-0.034981
P11	SmolLM2-360M	0.969173	0.998139	30–2000 / 7	0.782582	+0.186590
P12	Pythia-1B	0.931108	0.983104	30–2000 / 7	0.782582	+0.148526
P13	Qwen2.5-7B	0.996695	0.993942	30–2000 / 7	0.782582	+0.214113
P14	Qwen2.5-0.5B	1.028374	0.996775	30–2000 / 7	0.782582	+0.245792
P15	Mistral-7B-v0.1	1.060750	0.998690	30–2000 / 7	0.782582	+0.278168
P16	Llama-3-8B	1.045476	0.997461	30–2000 / 7	0.782582	+0.262894
P17	Gemma-2-9B (aleatorio)	1.134796	0.976472	30–2000 / 7	0.782582	+0.352214
P18	phi-3-mini (aleatorio)	1.036602	0.977923	30–1000 / 6	0.741726	+0.294876
P19	Llama-2-7B (Kaggle)	1.210761	0.987709	30–1000 / 6	0.741726	+0.469035
P20	Llama-2-13B (Kaggle)	1.054530	0.994348	30–1000 / 6	0.741726	+0.312804
P21	BLOOM-7b1	1.217824	0.997330	30–2000 / 7	0.782582	+0.435242

La última columna es $\widehat{\gamma} - \widehat{\gamma}_0$ con $A_0(d) = 1/(d+50)$ en cada fila exacta de la cuadrícula. Para P10, el candidato en $T = 2000$ es 0.752201, el ajuste es 0.747602, y la pendiente nula uniforme es 0.782582: un pequeño residuo del candidato no distingue la estructura aprendida de este instrumento. El exceso negativo no es una atención negativa ni una estimación causal del aprendizaje.

Pythia-2.8B tiene solo tres distancias de texto (30, 50, 100); su alta R^2 no puede establecer una cola hasta 2000. Mistral-Nemo utiliza 20–1000. Phi-3 y Falcon terminan en 1000. Estas diferencias de alcance son relevantes tanto para el ajuste como para las comparaciones de corpus. Un único puntaje típico de incertidumbre como ± 0.02 no está justificado por estos agregados. Los intervalos de resampling de distancia guardados y los intervalos de bootstrap secuencial de octubre tienen diferentes unidades de muestreo y no deben ser combinados.

El agregado de Kaggle para Llama-2-13B da $\widehat{\gamma} = 1.054530$ en $d \geq 30$ y 0.994652 en $d \geq 50$. Para Llama-2-7B los valores correspondientes son 1.210761 y 1.265264. Las notas incorporadas no son suficientemente consistentes con estos arrays para certificar la comparación anteriormente reportada en fp16. Estas filas permanecen disponibles como ajustes agregados, con procedencia de adquisición no resuelta; se excluyen de las afirmaciones sobre efectos de cuantización.

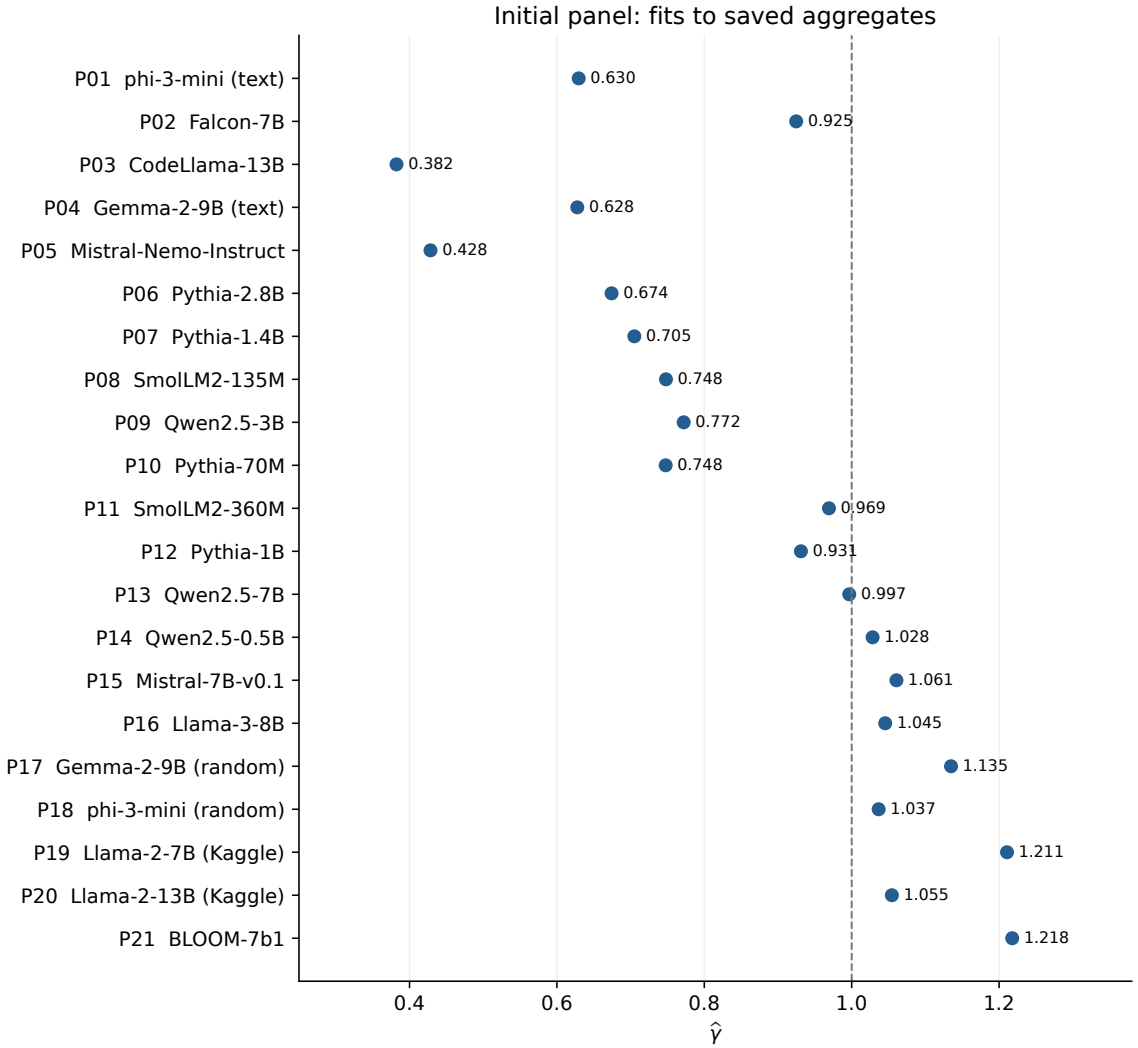


Figura 7: Las 21 entradas iniciales del panel, utilizando las estimaciones exactas de la Tabla 10 La línea discontinua es el límite de sumabilidad con soporte infinito matemático. $\gamma = 1$ No es una transición de fase medida en una red. No se reemplaza ningún exponente faltante o aproximado por cero.

8.3 Cantidades de Familia Finita desde los Exponentes Ajustados

Para cada exponente ajustado a plena precisión calculamos el normalizador discreto y los momentos en $d = 1, \ldots, 2000$. Así obtenemos $T_{\text{attn}} = 1/\widehat{\gamma}$, $U = \mathbb{E} \log d$, $S = \log Z_N + \widehat{\gamma} U$, $I = \text{Var}(\log d)$ y $C_V = \widehat{\gamma}^2 I$. Extender una cola de regresión a este soporte implica una suposición adicional distribucional. Estos valores no son entropías directas de atención ni mediciones nuevas. En particular, la correlación de S con su exponente de entrada no valida de forma independiente al modelo. La coordenada $\chi = 1/|\widehat{\gamma} - 1|$ se incluye para retener el diagnóstico propuesto, pero no es la susceptibilidad de Fisher de la familia finita: I permanece finito en uno.

Cuadro 11: Cálculos exactos de suma finita en $N = 2000$, claveada a P01–P21. Se produce el redondeo después del cálculo.

ID	T_{attn}	U	S	C_V	I	χ
P01	1.5884	5.3149	7.1060	1.4730	3.7163	2.6995
P02	1.0814	3.9348	6.0208	4.6217	5.4052	13.2775
P03	2.6157	6.0337	7.4775	0.3204	2.1922	1.6189
P04	1.5933	5.3220	7.1105	1.4586	3.7027	2.6856
P05	2.3340	5.9271	7.4343	0.4461	2.4302	1.7496
P06	1.4833	5.1421	6.9933	1.8335	4.0342	3.0690
P07	1.4183	5.0140	6.9050	2.1142	4.2528	3.3907
P08	1.3369	4.8251	6.7677	2.5437	4.5465	3.9682
P09	1.2953	4.7139	6.6832	2.8027	4.7023	4.3866
P10	1.3376	4.8269	6.7690	2.5396	4.5439	3.9620
P11	1.0318	3.6925	5.7914	5.1444	5.4768	32.4387
P12	1.0740	3.9000	5.9886	4.6989	5.4199	14.5154
P13	1.0033	3.5416	5.6430	5.4488	5.4850	302.6061
P14	0.9724	3.3681	5.4674	5.7740	5.4598	35.2435
P15	0.9427	3.1923	5.2837	6.0720	5.3964	16.4608
P16	0.9565	3.2750	5.3708	5.9361	5.4310	21.9895
P17	0.8812	2.8018	4.8552	6.5951	5.1213	7.4186
P18	0.9647	3.3232	5.4211	5.8533	5.4472	27.3206
P19	0.8259	2.4284	4.4175	6.8753	4.6900	4.7447
P20	0.9483	3.2259	5.3193	6.0178	5.4115	18.3384
P21	0.8211	2.3954	4.3775	6.8881	4.6444	4.5909

8.4 Observaciones y hipótesis mecanísticas

C1: el máximo finito de C_V. Para $N = 2000$, la maximización numérica de la suma discreta exacta da $\gamma \simeq 1.256885$ y $C_V \simeq 6.919812$. En uno, $C_V = 5.484118$, mientras que la aproximación continua da 4.814477. Esta es una propiedad de la familia estadística elegida: no sitúa una transición empírica de la red ni permite tratar la fórmula continua como identidad discreta.

C2: las grandes bases nominales no determinan el exponente ajustado. Los ajustes iniciales del texto Qwen son 1.028374 (0.5B), 0.772033 (3B) y 0.996695 (7B).

Esta secuencia no monótona no se explica por un candidato que dependa solo de la compartida θ y longitud de evaluación. Para $T = 2000$ y $\theta = 10^6$, la declaración Pad, la expresión da aproximadamente 0.997. El ajuste inicial del texto Llama-3 es 1.045476. Estas comparaciones son pruebas de un predictor propuesto, no confirmaciones derivadas únicamente de la configuración.

Sustituyendo $\widehat{\gamma} = .772033$ en el inverso propuesto da

$$\theta^{\star}_{\text{eff}} = \frac{T\sqrt{2}(1+\widehat{\gamma})}{2(1-\widehat{\gamma})} \simeq 11{,}000.$$

Este es un coordenada del modelo Padé, no una base física recuperada RoPE. La ajuste aleatorio Qwen-3B es sobre .964, por lo tanto su contraste de corpus es sobre +.192. Esa observación no demuestra geometría intacta, ausencia de cabezas de inducción, o un mecanismo causado por dos cabezas KV compartidas. Tales alternativas requieren activación controlada o intervenciones de peso. De igual manera, un gran θ en sí no cuenta con planos rotatorios activos: la dimensión rotatoria real y el umbral de fase entran en el conteo entero.

GQA y la analogía propuesta de presión. La relación entre las cabezas de consulta y las de clave y valor puede motivar una hipótesis arquitectónica sobre representaciones compartidas. El contraste actual en el corpus no identifica una presión, una ley de potencia o umbral de entrenamiento. Se necesitaría una intervención de GQA emparejada, con datos fijos, soporte de distancia y precisión de carga fijas, para probar dicha hipótesis.

C3: el contraste del corpus es descriptivo. Define $\Delta\gamma = \widehat{\gamma}_{\text{random}} - \widehat{\gamma}_{\text{text}}$. Los contrastes positivos, negativos y cercanos a cero ocurren todos. Seis pares representativos aparecen en Figure 8; la evidencia manifestada contiene 34 pares disponibles. Cada par mostrado se refit en la intersección de sus rejillas de ajuste guardadas. Las distancias coincidentes no establecen tokenizadores, prompts, semillas, precisiones o corridas de adquisición idénticas. Un signo positivo no es un detector suficiente del atención en ventana deslizante, y un signo negativo no es un detector suficiente de GQA o cabezas de inducción. No se establece ninguna estimación de precisión para tal clasificador aquí.

C4: verificación de cuantización y adquisición. El registro de texto inicial Llama-2 contiene $\widehat{\gamma} = .287057$ Al comparar ese número con un registro de una plataforma diferente no se puede aislar un efecto de 4 bits: el corpus, el rango de ajuste, la implementación y la precisión deben coincidir, y los metadatos de Kaggle discutidos anteriormente son incompletos. El propuesto inverso también tiene un dominio estricto: $\widehat{\gamma} = 1$ es singular y $\widehat{\gamma} > 1$ produce un valor negativo para positivo T No es un estimador físicamente válido en general para modelos en fp16. Una revisión controlada de precisión sigue siendo el experimento adecuado.

C5: ALiBi y ajustes de ley de potencia con alcance finito. BLOOM-7b1 tiene un exponente de texto 1.217824 con $R^2 = .997330$; la estimación inicial aleatoria es aproximadamente .882. Estos ajustes describen distancias muestreadas, no una cola infinita. Una mezcla de puntajes de contenido y sesgos posicionales no necesita tener la misma forma funcional que su sesgo solo. Se requieren datos de distancias mayores y una comparación con modelos competidores en datos de validación para distinguir entre colas de potencia,

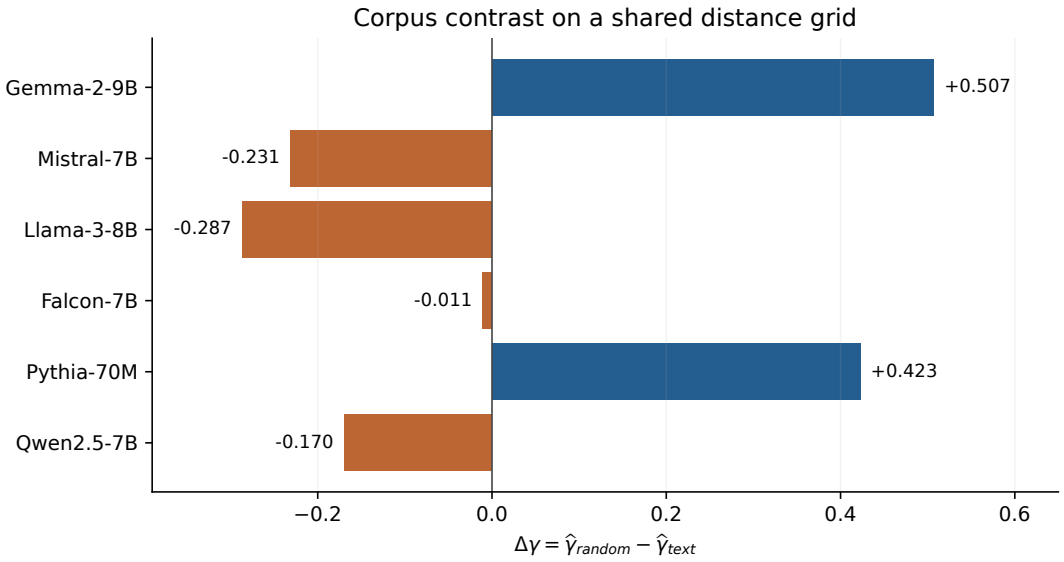


Figura 8: Seis contrastes de corpus recalculados a partir de medias guardadas en una rejilla común dentro del par. Falcon utiliza 30–1000; los otros cinco pares utilizan 30–2000. Los dos puntos anteriores ausentes de corpus aleatorio tienen datos guardados: Pythia-70M da +0.422912 y Qwen2.5-7B da −0.169680 Las barras representan estimaciones puntuales, sin un mecanismo inferido ni un intervalo de confianza no calculado.

exponenciales y mixtas. El teorema de MaxEnt identifica un optimizador bajo una restricción de momento prescrita; no demuestra que el entrenamiento impone esa restricción o que una ley de potencia es un atractor universal.

C6: Serie de tamaños de Pythia en rejillas de distancia emparejadas. Los ocho tamaños de Pythia [3] conservan sus observaciones individuales. La siguiente tabla ajusta ambos corpus en las mismas distancias dentro de cada par; la última fila tiene solo tres distancias. Esta restricción cambia la estimación aleatoria de Pythia-2.8B desde .551 en su rango más largo registrado a .790851 en 30–100, produciendo $\Delta\gamma = +.116690$. Comparar .551 con la estimación de texto de tres puntos mezclaría soportes.

Pythia	$\widehat{\gamma}_r$	R^2_r	$\widehat{\gamma}_t$	$\Delta\gamma$	n
14M	1.003714	0.977698	0.685288	+0.318427	7
31M	1.539824	0.964259	1.235001	+0.304823	7
70M	1.170514	0.994076	0.747602	+0.422912	7
160M	1.017145	0.981723	0.510892	+0.506254	7
410M	0.936235	0.987529	1.021853	-0.085618	7
1B	0.712754	0.956201	0.931108	-0.218354	7
1.4B	0.687585	0.948837	0.705073	-0.017488	7
2.8B	0.790851	0.998392	0.674162	+0.116690	3

Entre las siete parejas que abarcan 30–2000, el signo cambia entre 160M y 410M, pero esta asociación no identifica umbral de formación de cabeza inductiva. El resultado de 2.8B de soporte común también impide una afirmación universal sobre el signo posterior a 410M. El conteo de parámetros, la evolución del entrenamiento y la identidad de la cabeza mecanística son variables diferentes.

Conexión a dinámicas de cabeza de inducción. Investigación sobre cabeza de inducción [34] motiva la prueba de patrones de activación específicos de copia y su contribución causal durante el entrenamiento. Una sola comparación de pendientes de corpus no puede identificar esos circuitos, establecer un mínimo de parámetros para ellos, ni equiparar una comparación en serie de tamaños con una transición en el tiempo de entrenamiento. La propuesta conexión sigue siendo una hipótesis mecanística verificable.

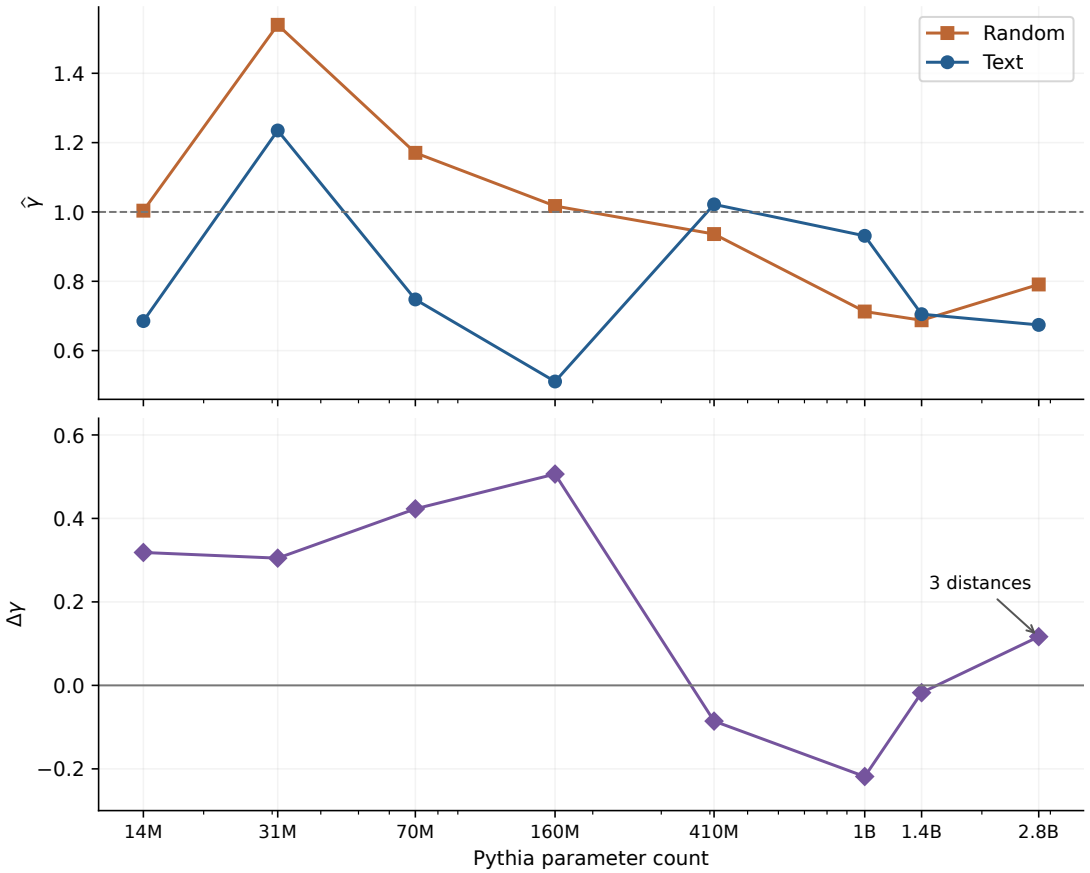


Figura 9: Estimaciones de Pythia y contrastes de corpus en rejillas de distancia dentro del par emparejado. El último punto utiliza tres distancias, todos los demás siete. Las líneas guían la mirada; no se infieren etiquetas de cabeza de inducción ni umbrales mecanicistas a partir del signo.

EXP-T1: puntos de entrenamiento y normalización H3. El experimento guardado H3 contiene ocho puntos de control, incluyendo el paso 3000. Registra un denominador $\mu_{d=10} - \mu_{\text{baseline}}$ y estadísticas de recuperación con 50 prompts. Un denominador negativo invierte la dirección intencional de normalización; un denominador positivo pequeño es numéricamente inestable. Estos hechos se refieren a la definición del puntaje, no a una transición Hagedorn. Los valores almacenados L_{crit} son salidas de ese protocolo, no una verdad base independiente del circuito.

Paso	denom	$L_{\text{crit},90}$
512	+0.3927	—
1000	+0.0024	1
2000	-1.0740	—
3000	-0.7385	—
8000	-0.9241	—
16000	-1.0837	—
32000	-0.6498	—
143000	+0.7085	4

Un archivo separado de perfiles de atención contiene nueve puntos de control y ocho distancias $\{10, 20, 50, 100, 200, 500, 1000, 2000\}$. Regresar los promedios guardados reproduce sus exponentes. Su conjunto de puntos de control difiere de H3: incluye los pasos 4000 y 64000, pero no el 3000. La tabla siguiente retiene ese experimento por separado. En particular, $\gamma > 1$ describe una decrecimiento más pronunciado dentro de la familia de leyes de potencia, no una atención maximalmente extendida. Alinear los dos protocolos por punto de control no establece un vínculo causal entre sus salidas.

Step	$\widehat{\gamma}$	R^2
512	0.855025	0.916052
1000	0.856545	0.975162
2000	0.904754	0.988525
4000	0.983180	0.988667
8000	1.063573	0.989672
16000	1.001999	0.991789
32000	0.886514	0.991220
64000	0.794771	0.990799
143000	0.780527	0.988175

C7: Falcon y la propuesta de cola mixta. La familia mixta $A(d) = C_0 d^{-\gamma} e^{-\lambda d}$ sigue siendo un candidato útil. Un resumen de ajuste guardado reporta los siguientes parámetros, pero no incluye su rejilla de ajuste, convención de pérdida o hashes de entrada. Se conservan como un informe de ajuste no verificado, no como parámetros reproducidos por el presente resumen agregado.

Corpus	C_0	γ	λ	$1/\lambda$	R^2
mongo	0.03857	0.743944	0.00103866	962.780	0.998985
random	0.00981	0.297567	0.00510298	195.964	0.994250

En los 30–1000 agregados actualmente guardados, el ajuste log-log puro da $R^2 = .989728$ para texto y .967807 para aleatorio. Un ajuste mixto log-lineal no restringido sobre esos mismos puntos da negativo λ en ambos corpus, y por lo tanto no reproduce el resumen con corte positivo anterior. La discrepancia debe resolverse desde las entradas originales y el objetivo antes de que esos parámetros mixtos respalden un mecanismo. Incluso un mayor entrenamiento R^2 para un modelo de tres parámetros no establecería por sí solo la copia de inducción, la dominancia exponencial o la superioridad predictiva; la comparación debe tener en cuenta la complejidad del modelo y el error en datos de prueba.

8.5 Protocolo de evidencia y reproducción

El atlas de octubre archivado proporciona 25 mediciones entre 42 identificadores históricos. Recalcularon el binning, la regresión y el bootstrap a partir de los perfiles guardados de NumPy, y asociaron cada fila a su registro de checkpoint exitoso. La discrepancia máxima en el exponente fue inferior a 3×10^{-15}; la discrepancia máxima en el punto final del intervalo fue inferior a 1.4×10^{-9}. El último refleja la almacenamiento de perfiles de secuencia en flotantes de 32 bits. Estas verificaciones reproducen los cálculos sobre los datos guardados; no reinician el modelo ni certifican que su implementación fue correcta.

El protocolo de origen utiliza texto del snapshot de pile-10k registrado, seleccionando documentos suficientemente largos para el tokenizer y el contexto del modelo, y tomando sus primeros T tokens. Esto no garantiza documentos o secuencias de tokens idénticas entre modelos. Todos los 25 registros exitosos reportan una fracción nula de atención no finita. Ese campo es una comprobación de instrumento registrada, no una atención completa reconstruida de forma independiente. Cuatro registros de trabajo identifican una revisión inmutable del modelo en su propia metadatos; los demás usan nombres mutables o rutas locales. Los hashes de perfiles guardados no reparan esa limitación.

8.6 Las 25 filas disponibles

Diecisiete modelos históricos no tienen una medición correspondiente en octubre en esta reconstrucción. Son observaciones faltantes, no fallos con exponente nulo, y son excluidos de los siguientes denominadores. La tabla también mezcla 15 contextos de 2048 y 10 de 1024, diferentes precisiones y tokenizadores específicos del modelo. Sus filas no deben tratarse como una intervención controlada sobre la arquitectura.

Un modelo, varios instrumentos. Pythia-2.8B muestra hasta qué punto el resultado depende del protocolo: E4 con texto da 0.674162 en tres distancias; E4 con tokens aleatorios da cerca de 0.551 en su rango más largo y 0.790851 en el soporte corto emparejado; QL con $T = 2048$ da 0.866809; el ajuste del campo, alrededor de 0.969; la media de los ajustes por celda, cerca de 1.309; y el registro de resumen por capa, cerca de 1.327. Son estimandos y registros distintos. En particular, una curva E4 muestrea diferentes filas y longitudes: no es un vector de probabilidades sobre distancias. Sustituir su pendiente en $p_g(d)$ construye otro miembro de la familia. Su entropía y sus cuantiles no deben presentarse como masa de atención medida ni como retención KV.

Controles descriptivos en el panel de consulta fija. Las ocho estimaciones punto de Pythia QL disminuyen estrictamente con el tamaño: 1.244, 1.233, 1.174, 1.073, 0.984, 0.974, 0.930, 0.867 (Spearman -1). En su nivel nominal compartido $\theta = 10000, T = 2048$, el candidato es constante (0.747006). Esto es una tendencia descriptiva en ajustes guardados; intervalos superpuestos, un punto de control por tamaño, arquitectura cambiante y revisiones inmutables ausentes limitan la atribución causal a escala. Las variantes de GPT-2, que utilizan posiciones absolutas aprendidas, abarcan aproximadamente 0.791–0.899 y superponen el panel RoPE. Una pendiente de distancia ajustada por lo tanto no identifica únicamente a RoPE. Estos controles motivan mediciones prospectivas emparejadas.

Nombres de puntos de control y identidad. Los archivos JSON iniciales identifican `google/gemma-2-9b-it`, `codellama/CodeLlama-13b-Instruct-hf` y `microsoft/phi-3-`

Modelo	QL exponent [95 % CI]	T	Precision
Llama-2-7b-hf	0.782 [0.750, 0.816]	2048	bf16
CodeLlama-13b-Instruct-hf	0.756 [0.716, 0.804]	1024	bf16
Mistral-Nemo-Instruct-2407	0.776 [0.742, 0.811]	1024	bf16
pythia-160m	1.073 [1.047, 1.101]	2048	fp32
gemma-2-9b-it	0.842 [0.793, 0.899]	2048	bf16
phi-3-mini-4k-instruct	0.385 [0.358, 0.414]	2048	bf16
pythia-2.8b	0.867 [0.831, 0.905]	2048	fp32
pythia-14m	1.244 [1.213, 1.281]	2048	fp32
pythia-1.4b	0.930 [0.897, 0.967]	2048	fp32
pythia-70m	1.174 [1.147, 1.199]	2048	fp32
gpt2-large	0.899 [0.873, 0.924]	1024	fp32
Qwen2.5-3B	0.895 [0.841, 0.949]	2048	bf16
gpt2-medium	0.880 [0.853, 0.906]	1024	fp32
OLMo-7B-hf	0.494 [0.482, 0.506]	1024	bf16
falcon-7b	0.838 [0.811, 0.867]	1024	bf16
pythia-1b	0.974 [0.943, 1.006]	2048	fp32
Qwen2.5-7B	0.844 [0.804, 0.884]	1024	bf16
gpt2-xl	0.791 [0.766, 0.815]	1024	fp32
pythia-410m	0.984 [0.953, 1.016]	2048	fp32
gpt2	0.873 [0.849, 0.898]	1024	fp32
phi-2	0.782 [0.750, 0.811]	2048	bf16
Meta-Llama-3-8B	0.854 [0.811, 0.902]	2048	bf16
Mistral-7B-v0.1	0.855 [0.815, 0.893]	2048	bf16
bloom-7b1	1.095 [1.069, 1.127]	1024	bf16
pythia-31m	1.233 [1.203, 1.264]	2048	fp32

Cuadro 12: Medidas de octubre guardadas, reproducidas desde perfiles archivados. Estas son ajustes específicos del protocolo, no observaciones independientes sobre el rendimiento del modelo.

`mini-4k-instruct`; etiquetas abreviadas como “Gemma-2-9B” no denotan puntos de control base verificados. La tabla QL nombra las mismas variantes. Coincidir con nombres de repositorios no establece revisiones de pesos idénticas, y el atlas legado de 42 filas no debe confundirse con el panel inicial de 21 filas E4. El acuerdo histórico/QL sigue siendo una comparación de identificadores archivados con procedencia de ejecución incompleta.

8.7 Acuerdo histórico y el Padé comparison

El exponente del texto histórico se encuentra dentro del intervalo de bootstrap de octubre para uno de los 25 modelos. La correlación de Spearman entre los exponentes históricos y QL es 0.269 Las 19 filas con un Pad almacenadoÉl valor da la correlación de Spearman -0.445 Contra el exponente QL. Empates en la PadLos valores reciben rangos promedio. Estos son cálculos descriptivos, no pruebas de significancia corregidas o evaluaciones prospectivas predictivas.

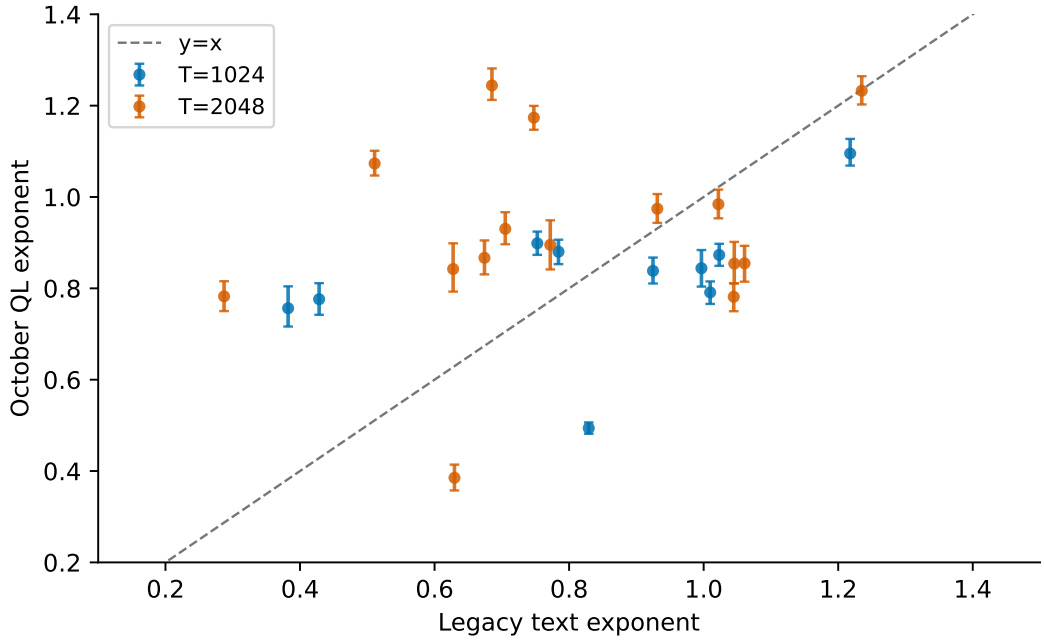


Figura 10: Texto histórico de exponentes frente a ajustes de octubre QL. Las barras de error son los intervalos de bootstrap archivados, reproducidos desde perfiles de secuencias. La diagonal denota igualdad; los colores diferencian la longitud del contexto. Los cambios en el corpus y el protocolo impiden interpretar el desacuerdo como una estimación causal de un sesgo instrumental nulo.

9 Flujo de trabajo de diagnóstico y controles de instrumentos

Un exponente ajustado describe la combinación de un mecanismo de atención y un protocolo de medición. La primera pregunta diagnóstica es si ese protocolo produce una pendiente en una distribución nula. Un R^2 alto dentro de la muestra no responde a esa pregunta: el control uniforme a continuación tiene $R^2 = 0.988452$ En la red principal E4. Las afirmaciones anteriores sobre la velocidad y los umbrales diagnósticos universales

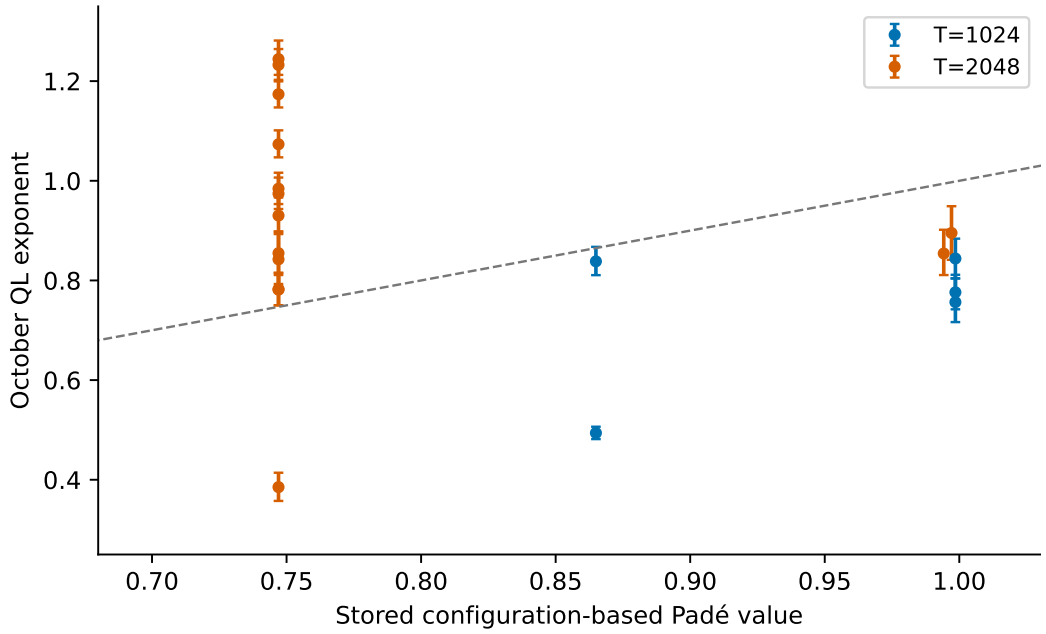


Figura 11: Nineteen valores almacenados basados en configuraciones Padé frente a ajustes QL. Este análisis verifica la comparación almacenada, no la configuración inmutable original de cada modelo. La gráfica no valida a Padé como predictor.

no son establecidas por este manuscrito. Los ejemplos de comandos en otros lugares son interfaces históricas.

9.1 La variable de longitud E4 tiene una nula uniforme no nula

Grupos 1–2 en el Apéndice K generalizan este control a probadores ponderados y demuestran cotas estrictas y monotonicidad para la exposición afín.

El extractor E4 disponible establece la longitud de la secuencia en $n(d) = d + 50$, lee la última consulta $i = d + 49$ y muestrea la clave $j = 49$. Así, modifica la consulta, el contexto disponible y el normalizador de fila con distancia, manteniendo fijo el índice de la clave. Para una atención causal uniforme, incluyendo la clave propia de la consulta

$$A_0(d) = \frac{1}{d+50}, \qquad \widehat{\gamma}_0 = \frac{\sum_k (x_k - \bar{x})(\log(d_k + 50) - \overline{\log(d+50)})}{\sum_k (x_k - \bar{x})^2}, \quad x_k = \log d_k. \tag{46}$$

Este es la pendiente OLS causada por el instrumento; no requiere pesos entrenados, codificación posicional o semántica de tokens. Su pendiente logarítmica local es $d/(d + 50)$, la cual aumenta con la distancia. Por lo tanto, el contraste entre pendiente corta y larga puede ocurrir incluso bajo una uniforme nula instrumental.

Estos valores se verificaron tanto analíticamente como al pasar tensores sintéticos de atención causal a través de las funciones guardadas de extracción y ajuste E4. No se ejecutó ningún modelo lingüístico. Para la mezcla $(1 - s)U_n + s\delta_0$, con constante $s < 1$ y U_n uniforme en todas las $n = d + 50$ claves elegibles, las entradas fuera del sumidero son $(1 - s)/(d + 50)$ y la pendiente permanece inalterada. Si en cambio la clave cero recibe exactamente s y las otras $n - 1$ claves comparten $1 - s$, las entradas son $(1 - s)/(d + 49)$: el nulo instrumental de siete puntos es 0.785491, en lugar de 0.782582. Una masa del sumidero dependiente del contexto introduce otra dependencia. Las cabezas fallidas,

Cuadro 13: Estimador E4 mediante nulo causal uniforme. No se utilizan pesos de modelo.

Grid	$\widehat{\gamma}_0$	R^2
30–2000 (7 points)	0.782582	0.988452
30–1000 (6 points)	0.741726	0.987211
10–2000 (8 points)	0.681834	0.964115
30–100 (3 points)	0.525737	0.993818
10–100 (4 points)	0.396753	0.968315
200–2000 (4 points)	0.913768	0.999228

máscaras y exclusiones de valores no finitos requieren controles correspondientemente ajustados.

El valor de configuración $\gamma_P(10^4, 2000) = 0.752201$ se encuentra cercano a esta pendiente instrumental, pero la cercanía en sí misma ni demuestra ni cuantifica la fracción de un ajuste empírico causada por el confundidor. La anterior comparación de precisión no puede identificar la geometría rotativa como causa. De manera similar, respuestas débiles a la dirección o un exponente cercano a 0.75 en la inicialización no establecen independientemente una ley geométrica aprendida.

9.2 Consultas fijas diferencian a QL de la media diagonal

La implementación de QL en octubre utiliza el mismo conjunto de consultas $Q = \{T - \lfloor T/4 \rfloor, \ldots, T - 1\}$ en cada distancia ajustada. Bajo atención causal uniforme su perfil es

$$A_{0,\mathrm{QL}}(d) = |Q|^{-1} \sum_{i \in Q} \frac{1}{i+1}, \qquad 16 \le d \le T/2,$$

que es constante en d. En contraste, el perfil ALL promedia sobre consultas $i \ge d$ y tiene nulo

$$A_{0,\mathrm{ALL}}(d) = \frac{1}{T-d} \sum_{i=d}^{T-1} \frac{1}{i+1}.$$

Ejecutar el acumulador QL real y el ajustador de binos logarítmicos en matrices sintéticas da $|\widehat{\gamma}_0| < 10^{-10}$ en $T = 1024$ y 2048, ambos con y sin una masa añadida como sumidero 0.2. Para TODOS incluyendo la clave cero pero sin masa adicional como sumidero, las pendientes son 0.317033 en $T = 1024$ y 0.298064 en $T = 2048$. Los dos números indexan longitudes de contexto, no la presencia versus ausencia de un sumidero. Estos números usan 16 bins sobre $[16, T/2]$; otras rangos o reglas de ponderación producen otras pendientes de TODOS. Deben atribuirse necesariamente a QL.

El perfil completo almacenado QL alcanza $d = 3T/4$. En ese último punto final, una consulta muestrea la clave cero, en contra de un comentario en el instrumento. El rango ajustado $[16, T/2]$ excluye ese punto final y todos los términos de clave cero. Esta distinción en la frontera es parte del auditorio del instrumento, no una corrección de los 25 ajustes QL guardados. Un control sintético plano verifica la reducción en ese nulo; no certifica la adquisición del modelo real, mascarada, precisión o procedencia.

9.3 Informe del observable y su control correspondiente

Para las mismas distancias, exclusiones y pesos de regresión, la linealidad de OLS da

$$\widehat{\gamma}_{\text{excess}} = \widehat{\gamma} - \widehat{\gamma}_0 = -\operatorname{slope}_{\log d} \log\left(\frac{A(d)}{A_0(d)}\right). \tag{47}$$

El paquete de reproducibilidad registra este diagnóstico para todas las 21 filas del panel inicial, conservando cada pendiente original, cuadrícula y fuente. Asume que la fila registrada fue adquirida con el procedimiento declarado $d + 50$; la procedencia incompleta no es reparada por esa asunción. El exceso de pendiente no es automáticamente un exponente causal o aprendido, y ajustar una proporción no lo convierte en una distribución de probabilidad normalizada. Las comparaciones de corpus deben utilizar también soportes comunes dentro del par.

Una medición prospectiva debe fijar la longitud del contexto y las posiciones de las consultas, indicar si las claves de autoatención y de sumidero están excluidas, y archivar perfiles por secuencia, exclusiones, revisiones del modelo/tokenizador, implementación de atención, precisión, corpus y semilla. Los controles de causalidad uniforme, causalidad uniforme más sumidero, pesos aleatorizados y tokens mezclados responden a diferentes preguntas. Solo los primeros dos controles sintéticos se ejecutan en esta revisión. Las familias alternativas deben compararse en secuencias de prueba bajo las mismas convenciones de observación y pérdida. Resamplear secuencias, no distancias individuales, para la incertidumbre cuando existan registros a nivel de secuencia.

9.4 Cantidades derivadas de la familia y coordenadas inversas

Para una familia finita de ley de potencia declarada, calcule Z_N, la entropía, la información de Fisher y C_V mediante sumas finitas. Estas cantidades describen esa familia; no establecen la sensibilidad del modelo ante cambios de contexto. $1/\gamma$ es una temperatura formal solo cuando el denominador no es cero; interpretarla como temperatura positiva exige $\gamma > 0$. Los intervalos alrededor de 0.95 y 1.05 eran etiquetas heurísticas, no clasificadores arquitectónicos validados.

La inversa correcta del candidato racional en valor fijo $T > 0$ es

$$\theta_{\text{inv}}(g, T) = \frac{T\sqrt{2}(1+g)}{2(1-g)}, \qquad -1 < g < 1. \tag{48}$$

Es una coordenada algebraica positiva en ese dominio. No recupera una base de entrenamiento física a partir de una pendiente observada. La expresión anterior $T\sqrt{2}/(1-g)$ es la inversa de una aproximación lineal; sustituyendo $g = \gamma_P(\theta, T)$ se obtiene $\theta + T/\sqrt{2}$, no θ. Los exponentes iguales o superiores a uno no pueden invertirse en una base positiva por el candidato.

9.5 Divergencia respecto a una referencia declarada

Para una referencia normalizada positiva q y una probabilidad observada p en el mismo soporte, una discrepancia descriptiva es

$$\mathrm{KL}_{\text{grammar}} = \sum_d p(d) \log \frac{p(d)}{q(d)}, \qquad q(d) = d^{-g} / \sum_{r \in \mathcal{D}} r^{-g}. \tag{49}$$

El script diagnóstico disponible normaliza las distancias muestreadas sparsas con pesos iguales en cada punto, en lugar de reconstruir todas las distancias de token o masas de bin. Su KL se refiere entonces a ese soporte muestreado. Ni un pequeño KL ni el antiguo umbral 0.05 identifican un circuito posicional, un sumidero de atención o una cabeza de inducción. La atribución de circuitos requiere evidencia funcional más allá de un residual.

Cuadro 14: Panel histórico preservado de KL. Corpus mixtos y soportes escasos; no se conservan diagnósticos de circuito ni atribuciones de efecto de precisión.

Entrada histórica	$\widehat{\gamma}$	KL
Gemma-2-9B (aleatorio)	1.135	0.075
Mistral-7B-v0.1	1.061	0.041
LLaMA-3-8B	1.046	0.018
LLaMA-2-7B (reportado fp16)	1.026	0.002
Qwen2.5-7B	0.997	0.011
Pythia-1B	0.931	0.004
Pythia-70M	0.748	0.008
Pythia-2.8B	0.674	0.001
SmolLM2-135M (resumen de par antiguo)	0.902	0.009

El SmolLM2-135M input 0.902 es el redondeado 0.9019 en el resumen separado del 21 de abril base/instruct, etiquetado como aleatorio en el resumen maestro. No es el resultado de texto E4 P08 (0.747996); la matriz aleatoria disponible de E4 reproduce en lugar a 0.626655. El resumen de la pareja más antigua no tiene medias por distancia para esta fila, por lo que su adquisición no puede reconstruirse aquí. Su valor KL permanece como un informe histórico. El mismo exponente más antiguo se mantiene como entrada histórica explícita en la Tabla 16, únicamente para auditar la aritmética de cuantiles; no se sustituye por P08.

9.6 Reevaluación de los cuatro casos históricos

LLaMA-2-7B. Las pendientes comunicadas 0.287 y 1.026 se conservan como registros históricos, pero los arrays recuperados y las notas discrepan (Sección 8). No certifican un efecto emparejado de fp16 frente a 4 bits, invariancia con el tamaño ni una recomendación NTK. Además, el valor superior a uno queda fuera del dominio de la inversa con base positiva.

Mistral-7B-v0.1. El antiguo par texto/aleatorio 0.830/1.061 es un informe específico del protocolo. La configuración oficial declara una ventana deslizante de 4096 tokens; una secuencia de hasta 2050 tokens no excluye claves mediante esa ventana. Por ello, una máscara SWA directa no explica ese contraste de contexto corto. La fuente de configuración es `https://huggingface.co/mistralai/Mistral-7B-v0.1/blob/main/config.json`. Pueden existir otras diferencias arquitectónicas y de entrenamiento. confunden, y $\Delta\gamma$ no es un detector de GQA.

Qwen2.5-7B. Un valor ajustado 0.997 no implica atención puramente posicional. El cálculo continuo límite es $2000^{0.9} = 935.248$, mientras que utilizar $g = 0.997$ da una

diferente cuantil. Ninguno es un resultado medido de retención de tokens o precisión; los valores discretos se tabulan a continuación.

Pythia-2.8B. Su ajuste inicial de texto utiliza solo distancias 30, 50 y 100. La nula uniforme en esa cuadrícula es 0.525737, y la comparación de soporte común cambia el contraste aleatorio/texto. No se sigue ninguna política de seguridad, estabilidad o cache óptima de largo contexto de ese exponente de tres puntos.

10 Heterogeneidad de capa y la propuesta de capa crítica

Perfiles resueltos por capa, intervenciones de recuperación y medida de similitud representativa son observables diferentes. Su acuerdo debería ser probado bajo un protocolo común, en lugar de incorporarse en la definición de una frontera. Esta sección preserva los informes originales de capa mientras separa una fórmula propuesta de configuración de umbrales medidos.

10.1 Una capa coordenada candidata

Definición 10.1 (Propuesta histórica de coordenada de capa)**.** Para $\theta > 1$, $T > 0$, N_L capas y dimensión de cabeza $d_{\text{head}} > 0$, define

$$L_{\text{pred}} = N_L \log_\theta(T/(2\pi))(d_{\text{head}}/64)^{1/\sqrt{12}}. \tag{50}$$

Esta es la configuración histórica propuesta anteriormente denotada L_{crit}. Su factor de cabeza completa también ignora la RoPE parcial: Pythia-1B tiene $d_{\text{head}} = 256$ pero $d_{\text{rot}} = 64$. Reemplazar uno por el otro definiría un nuevo candidato, no validaría la fórmula antigua. Su valor no necesita ser un entero ni estar en $[0, N_L]$; el redondeo y cualquier clipado deben establecerse antes de compararlo con un índice de capa medido.

La identidad $\text{sd}(\text{Unif}[0,1]) = 1/\sqrt{12}$ no deriva el exponente empírico de la arquitectura. GPT-2 ha aprendido posiciones absolutas y no tiene una base RoPE para sustituir en esta ecuación. Sus informes de capa histórica permanecen observaciones bajo otro protocolo, no evaluaciones de la fórmula RoPE.

Cuadro 15: Resumen de la capa histórica conservado para comparación. Las dos últimas columnas son informes, no confirmaciones independientes del candidato.

Modelo	N_L	reportó L_{crit}	α	predicción antigua	ajuste antiguo
Pythia-70M	6	4	.667	.748	.748
Pythia-1B	16	15	.938	.931	.931
GPT2-medium	24	23	.958	—	—
GPT2-large	36	35	.972	—	—

10.2 Una mezcla no tiene el exponente promedio

Grupo 3 en el Apéndice K prueba las afirmaciones sobre la pendiente local, la curvatura y la regresión global para cualquier mezcla finita positiva de potencias.

Para coeficientes positivos y dos perfiles de componentes, la pendiente local exacta es

$$A(d) = ad^{-g_1} + bd^{-g_2}, \quad a, b > 0,$$
$$-\frac{d \log A(d)}{d \log d} = \frac{ag_1 d^{-g_1} + bg_2 d^{-g_2}}{ad^{-g_1} + bd^{-g_2}}. \tag{51}$$

Los pesos dependen de la distancia a menos que los exponentes componentes coincidan o uno de los componentes se anule. Una pendiente OLS global de un perfil promedio de atención tiene por tanto que no necesariamente igualar una media ponderada por fracción de capa de exponentes ajustados. Esa igualdad también no tiene en cuenta normalizadores finitos y ajustes fallidos.

La lista $1.35, 1.41, 1.12, 1.10$ para Pythia-70M sobrevive como un informe manuscrito redondeado; su registro original de perfiles de capas no fue recuperado. No debe identificarse con la matriz por cabeza utilizada para P3 en la Sección 5.6. Esa matriz proporciona medias por cabeza finitas $1.918196, 1.663184, 1.529207, 1.082143$ para las capas L0–L3, con las capas L4–L5 ausentes; la media de L1–L3 es 1.424844, mientras que la lista antigua proporciona 1.210000 allí. Para la antigua lista de cuatro entradas, la media aritmética es 1.245; multiplicando por $4/6$ se obtiene 0.830, no 0.748. Incluso la media aritmética corregida no es una derivación de la pendiente global. Las entradas marcadas como NaN en las capas 4–5 denotan ajustes no disponibles o rechazados, no un exponente nulo o una prueba de procesamiento semántico. Los valores reportados y predichos duplicados en la antigua tabla no son validaciones independientes sin el ajuste original y registro de selección de parámetros.

10.3 CKA es una observación separada

El análisis complementario [29] fue citado para valores de CKA en capas adyacentes cercanos a uno antes de una propuesta frontera y por debajo 0.5 posteriormente. Esas afirmaciones históricas se conservan en el registro de auditoría. Un cambio en CKA puede motivar un análisis compartido, pero por sí mismo no identifica cálculo posicional versus semántico ni valida la ecuación (50). Una prueba de confirmación debe especificar la misma clave, ejemplos, representaciones, indexación y selección de umbral, y evaluar el acuerdo en observaciones mantenidas fuera de entrenamiento.

10.4 Entrenamiento de puntos de control y normalización de recuperación

El experimento H3 guardado tiene ocho puntos de control, incluyendo el paso 3000, y utiliza 50 promt. Su normalización de recuperación contiene $\mu_{d=10} - \mu_{\text{baseline}}$ Un denominador negativo invierte la dirección intencional; uno cercano a cero amplifica la variación numérica. La ausencia o inestabilidad de umbral no implica que la frontera de una capa física deje de existir. Los valores de umbral almacenados y los denominadores se conservan en la Sección 8 y en el paquete de reproducibilidad.

La trayectoria de atención tiene nueve puntos de control, incluyendo 4000 y 64000 pero no 3000. Estos son protocolos separados. Su exponente ajustado excede uno en los pasos 8000 y 16000, no a lo largo de los pasos 2000–32000. Dentro de la familia de leyes de potencia, un exponente positivo más grande significa decaimiento más pronunciado, no atención maximamente extendida. El acuerdo en el último punto de control no establece

un atractor estable, una formación del tiempo para cabezas de inducción o un mecanismo de evaporación. Esas interpretaciones requieren intervenciones circuiteras independientes.

11 Compresión de KV-Cache: Presupuestos Masivos y Pruebas Registradas

Grupos 4 y 6 en el Anexo K prueban el orden cuantil discreto y las identidades exactas de truncamiento de fila congelada, incluyendo sus dominios y cotas afiladas.

Una masa cuantil responde a una pregunta distribucional sobre un soporte declarado. No identifica por sí misma una política de caché que preserve la calidad. En particular, los ajustes E4 de longitud variable tienen la confusión instrumental demostrada en la Sección 9.1. Los cálculos de familia que se presentan a continuación permanecen válidos matemáticamente, mientras que utilizar la pendiente del modelo ajustado como entrada requiere un protocolo de medición y evaluación emparejado.

11.1 Ventanas continuas y discretas

Teorema 11.1 (Ventana continua finita de masa)**.** *Para densidad proporcional a d^{-g} en $[1, N]$, $N > 1$ y $0 < f < 1$, la única cuantil de masa f es*

$$D_{f,c} = \begin{cases} [(1-f) + fN^{1-g}]^{1/(1-g)}, & g \neq 1, \\ N^f, & g = 1. \end{cases} \tag{52}$$

Para distancias de tokens $\{1, \ldots, N\}$ con medida de conteo, el correspondiente cuantil exacto es en su lugar $D_{f,N} = \min\{k : \sum_{d=1}^{k} d^{-g} \geq f \sum_{d=1}^{N} d^{-g}\}$. Para el caso discreto N es un entero. Ambos cuantiles están definidos para cada real finito g. No incluyen una clave propia en $d = 0$ o un sumidero de atención modelado por separado.

Demostración. Para $g \neq 1$, dividir $\int_1^D d^{-g}\,\mathrm{d}d$ entre $\int_1^N d^{-g}\,\mathrm{d}d$ da $(D^{1-g} - 1)/(N^{1-g} - 1)$. Ambas integrales contienen el factor $1/(1-g)$. Igualar esta razón a f da lugar a la fórmula indicada. En $g = 1$ la proporción es $\log D / \log N$. En cualquier caso, la densidad acumulativa es estrictamente creciente. En el soporte discreto, las sumas parciales aumentan estrictamente y el índice más pequeño que cumple la condición de masa es único. □

Para fijado $g < 1$ y f, el gran-N approximation es $D_{f,c} \sim Nf^{1/(1-g)}$. No es uniforme como $g \to 1$ en un valor fijo de N. Para un valor fijo de $g > 1$, el cuantil continua tiende a $(1-f)^{-1/(g-1)}$ cuando $N \to \infty$; esto no es un argumento de serie geométrica y su precisión en un valor particular finito de N debe verificarse.

La tabla conserva los exponentes originales redondeados para que el cambio aritmético pueda ser aislado. Los nombres de los modelos identifican esos inputs históricos, no mediciones actualizadas no validadas. Por ejemplo, en $g = 0.674$, la ventana de masa media es aproximadamente 305.477 continuamente y 293 discretamente; el anterior 987 no era la cota de esta familia. En $g = 0.75$, $N = 2000$ y $f = 0.9$, los valores esquemáticos corregidos son 1401.606 (continuos) y 1389 (discretos).

La masa empírica real debe medirse en un soporte fijo y adecuado, en lugar de sustituirse por cálculo sin verificar el ajuste.

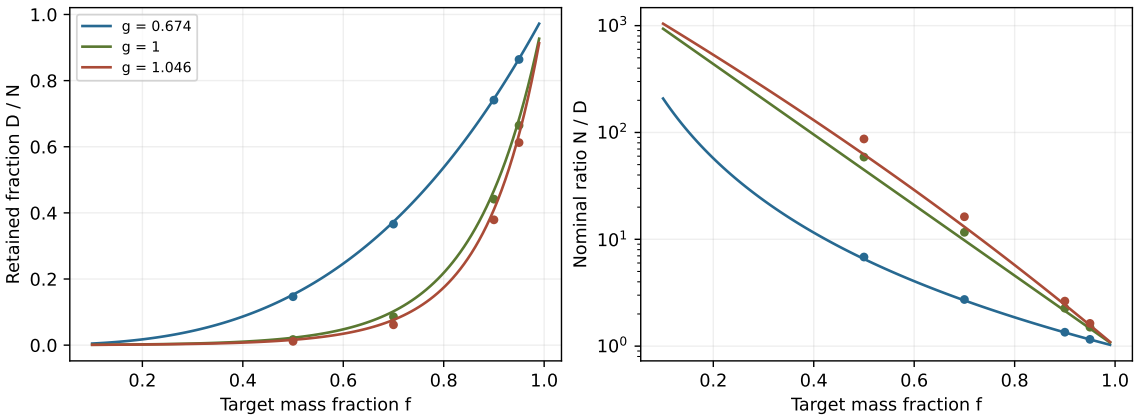


Figura 12: Familia de cálculos en $N = 2000$: cuantiles continuos (líneas) y cuantiles discretos de distancia de token (marcadores). Los inputs son exponentes ilustrativos, no políticas calibradas del modelo. El cuantil continuo es suave a través de $g = 1$; el cuantil discreto es de valor entero y puede tener saltos. Ninguno de ellos tiene umbral especial de precisión de tarea allí.

Cuadro 16: Cuantiles familiares recalculados $N = 2000$ Los valores continuos y discretos no son intercambiables.

Historial de entrada	g	$D_{.9,c}$	$D_{.9,N}$	$D_{.5,c}$	$D_{.5,N}$
LLaMA-2-7B	1.026	870.9	813	37.1	27
Pythia-2.8B	0.674	1489.5	1482	305.5	293
Pythia-70M	0.748	1404.2	1392	220.4	205
SmolLM2-135M (input antiguo)	0.902	1153.3	1121	89.3	74
Pythia-1B	0.931	1093.4	1056	73.2	59
Qwen2.5-7B	0.997	942.5	891	45.7	34
LLaMA-3-8B	1.046	820.2	758	32.1	23

11.2 Una cota condicional de salida y asignación real de caché

Sea una fila de atención congelada tenga pesos p_j, valores v_j y masa removida $0 \le \epsilon < 1$. Después de remover esas entradas y renormalizar los pesos retentos, escribe las medias condicionales conservadas y removidas como o_K, o_R. Para $\epsilon > 0$, $o = (1-\epsilon)o_K + \epsilon o_R$, por lo tanto

$$\|o - o_K\| = \epsilon\|o_R - o_K\| \le \epsilon \operatorname{diam}\{v_j\} \le 2\epsilon \max_j \|v_j\|.$$

El caso de masa nula es inmediato. La primera desigualdad se sigue porque ambos promedios condicionales son combinaciones convexas de los valores. Una proyección lineal posterior suma su norma operador inducida. Esto es una cota de fila fija; no abarca estados ocultos alterados, pasos autoregresivos repetidos o la precisión de la tarea después de la expulsión.

Para un caché denso no expandido convencional con L capas, H_{KV} cabezas clave/valor, dimensión de cabeza d_{head} y b bytes por escalar, la expulsión física de $N - W$ tokens ahorra $2bLH_{\mathrm{KV}}d_{\mathrm{head}}(N - W)$ bytes por secuencia, antes de los gastos del administrador de memoria. GQA utiliza el número de cabezas KV, no el número de cabezas de consulta. Inclinación de puntajes o establecimiento de pesos en cero no reduce por sí misma una asignación densa. Para la entrada ilustrativa antigua $g = 1.046$, $N = 8192$, $f = 0.9$, la ventana continua es 2757.511, una reducción nominal de 66.34 %; la ventana discreta es 2537, una reducción de 69.03 %.

Estas son cálculos de presupuesto de masa; no establecen a detalle el perfil de atención de longitud 8192 ni certifican un contexto más largo utilizable.

11.3 EXP-B2: las ventanas realmente evaluadas

El registro conservado de dos modelos especifica $N_{\mathrm{full}} = 2048$, 120 muestras y 64 objetivos evaluados por muestra en el conjunto de prueba de WikiText-2. El código disponible utiliza $W = \mathrm{round}(Nf^{1/(1-g)})$, con un mínimo de 64 y una ventana completa cuando $g \ge 1$. Por tanto, los ensayos con $W = 445$ y 1482 prueban esa regla asintótica, no el cuantil finito exacto del teorema.

Cuadro 17: Registro original de dos modelos EXP-B2. Los deltas positivos indican un PPL registrado mayor.

Modelo	g	probado W	PPL completo	Δ PPL(W)	Δ PPL(1024)
pythia-1b	0.931	445	11.7956	+1.4443	-0.1433
pythia-2.8b	0.674	1482	9.2421	-0.5141	-0.0126

Para los mismos inputs redondeados y $N = 2048$, las ventanas exactas 90 % son 1117.959 continuamente y 1079 discretamente para Pythia-1B, y 1524.892 y 1518 para Pythia-2.8B. Estas son nuevas cantidades familiares recalculadas, no ventanas que fueron evaluadas en las pruebas guardadas. La coincidencia numérica entre 1482 y un cuantil discreto en $N = 2000$ no cambia el registro experimental de $N = 2048$.

El script muestrea cada objetivo desde un intervalo cuyo punto final inferior es W. Reutilizar la semilla 42, por tanto, no mantiene las posiciones de los objetivos idénticas entre ventanas. Además, durante las 64 predicciones forzadas por el maestro, el prefijo disponible crece desde W hasta $W + 63$; esto no es un benchmark de expulsión de caché con contexto fijo en cada objetivo. Los agregados guardados solos no pueden recuperar

intervalos de confianza pareados ni diferenciar un efecto de ventana de cambios en las posiciones muestreadas. La muestra negativa ΔPPL no demuestra que los tokens descartados fueran ruido.

Extensión de siete modelos separados. El registro extendido conserva siete modelos y utiliza la misma regla asintótica. Su script disponible especifica 120 muestras y 32 objetivos por muestra, mientras que el JSON omite esos dos ajustes. Esta evidencia respalda una descripción del código fuente, no un certificado de la exacta ejecución histórica.

Cuadro 18: Panel extendido separado EXP-B2, conservado sin reetiquetar sus ventanas asintóticas como cuantiles finitos exactos.

Modelo	g	probado W	W/N	Δ PPL	almacenado rodilla
pythia-14m	0.685	1466	71.6 %	+4.3765	1280
pythia-70m	0.748	1348	65.8 %	-1.1329	1024
pythia-160m	0.511	1651	80.6 %	+1.4443	1024
pythia-410m	1.022	2048	100.0 %	+0.0000	768
pythia-1b	0.931	445	21.7 %	+0.9997	768
pythia-1.4b	0.705	1433	70.0 %	-0.1715	768
pythia-2.8b	0.674	1482	72.4 %	-0.3408	768

Las dos entradas mostradas +1.4443 son una coincidencia de redondeo: el archivo JSON original de Pythia-1B proporciona $13.2399089962 - 11.7955767485 = 1.4443322477$, mientras que el archivo JSON extendido de Pythia-160M proporciona $23.9412332628 - 22.4969311427 = 1.4443021201$. Son diferencias guardadas distintas, ambas reproducidas a partir de sus PPLs registrados.

Los diferentes deltas de Pythia-1B y 2.8B en los dos paneles pertenecen a ejecuciones separadas y no deben ser fusionados. Pythia-14M tiene $g = 0.685$ pero presenta un valor positivo de ΔPPL en el panel extendido, por lo que las observaciones no establecen un intervalo favorable universal como $[0.65, 0.85]$. La entrada de 410M utiliza la ventana completa y no es una prueba de compresión exitosa. El puntaje heurístico registrado depende de la escaneo muestreado y su criterio; no se sigue una sustitución óptima o general $f = 0.93$–0.95 de este registro.

11.4 Asignación de atención con ley de potencia: presupuestos condicionales de masa

Para la familia discreta con $\gamma > 0$, los pesos disminuyen con la distancia. Entre conjuntos de cardinalidad fija D, las primeras D distancias maximizan la masa retentiva de la familia. Su masa es

$$M_N(D) = \frac{\sum_{d=1}^{D} d^{-\gamma}}{Z_N(\gamma)}.$$

El presupuesto entero mínimo que alcanza masa al menos f es el cuantil discreto $D_{f,N}$ del Teorema 11.1. Este resultado se refiere a la familia ideal, no a rankings de tokens dependientes del contenido, precisión de tarea o un kernel sparso óptimo. En $\gamma = 0$ todas las familias de esa cardinalidad empatan; en negativo γ los mayores pesos de la familia están a las mayores distancias. Una interpretación de ventana reciente requiere por tanto la condición de exponente positivo.

Aproximación continua y álgebra de umbrales. Para valores fijos de $0 < \gamma < 1$ y $0 < f < 1$, el cuantil continua satisface

$$D_{f,c}(L) \sim L f^{1/(1-\gamma)}, \qquad \frac{f}{D_{f,c}(L)} \sim \frac{f^{-\gamma/(1-\gamma)}}{L} \quad (L \to \infty).$$

La segunda expresión corresponde a la densidad promedio retentiva, no a su umbral de frontera. La densidad continua en la frontera satisface en cambio $p_c(D_{f,c}) \sim (1-\gamma) f^{-\gamma/(1-\gamma)}/L$. Una dependencia inversa de longitud en un método empírico como BLASST [49] no identifica su coeficiente calibrado con ninguna de estas expresiones. El exponente anterior $f^{1/(1-\gamma)}$ en el cálculo de densidad promedio fue un error algebraico. La consulta de configuración no puede reemplazar la calibración del método basada en este cálculo.

Comparación con PyramidKV. PyramidKV [4] reporta experimentos que retienen 12 % del caché KV. Su selección dependiente de la capa es diferente a retener una ventana fija de distancia. Para examinar la comparación numérica realizada previamente aquí, tome $\gamma = 1.05$ como un ejemplo de *familia de entrada*, no una predicción Padé: el último siempre es inferior a uno para entradas positivas. En $N = 8192$, la fórmula continua exacta da

$$D_{0.9,c}/N = 0.330509, \qquad D_{0.8,c}/N = 0.115769, \qquad D_{0.78,c}/N = 0.094466.$$

Estos reemplazan a los anteriores 0.27, 0.13 y la asociación afirmada 12 % con $f = 0.78$. Los cuantiles de distancia de token usan sumas finitas en su lugar. Ni el acuerdo numérico ni la elección de f después de ver un presupuesto externo predice la precisión del método externo.

Razonamiento y rendimiento en contexto largo. SnapKV, PyramidKV, H2O y estudios de compresión orientados a la razón [4, 23, 25, 49, 51] Proporcionen comparaciones empíricas específicas de la tarea. No establecen que un cuantil masivo preserve la razonamiento. Los valores de referencia reportados anteriormente se conservan en el archivo separado del predecesor indexado por el Anexo. J No se volvieron a ejecutar para esta revisión. Del mismo modo, $1/\gamma$ es un exponente de umbral de conteo bajo la familia indicada, no una dimensión geométrica demostrada que prohíba el razonamiento multisalto cuando $\gamma > 1$ Calidad del test, posiciones conservadas, pozos de atención, memoria y latencia directamente para cada intervención propuesta.

11.5 Razón de base desplazada: alcance de la anterior expresión de pérdida de información

Definir la coordenada algebraica de base desplazada $b(x) = \theta + x/\sqrt{2}$. La cantidad anteriormente denotada como pérdida de información es la razón logarítmica signada

$$\Delta H_{\text{coord}}(D, N) = \log\left(\frac{\theta + D/\sqrt{2}}{\theta + N/\sqrt{2}}\right), \qquad \theta > 0, \quad 0 \le D \le N. \tag{53}$$

Se anula en $D = N$ y su extensión algebraica a $D \to 0$ tiende a $\log\theta - \log(\theta + N/\sqrt{2})$. Esto es una diferencia de logaritmos, no la proporción previamente impresa de logaritmos. La expresión es no positiva; no se establece como pérdida de entropía de Shannon, divergencia

KL o error de tarea. El punto final $D = 0$ es una extensión de esta coordenada, no un cuantil del soporte continuo $[1, N]$. Dentro de la definición de coordenada, la fórmula es exacta y no requiere ninguna estimación de error de primer orden.

11.6 Decaimiento de Horizonte Suave: una propuesta de sesgo en puntuación probada

La propuesta de peso de distancia puede estudiarse como una intervención elegida, sin interpretar su umbral como un horizonte físico. Para $h > 0$, $g > 0$ y distancia causal $d \geq 0$, define

$$w(d) = \begin{cases} 1, & d = 0, \\ (1 - d/h)^g, & 0 < d < h, \\ 0, & d \geq h. \end{cases} \tag{54}$$

Añadir $\log w(d)$ a los puntajes pre-softmax produce pesos proporcionales a $w(d)e^{s_{ij}}$, con pesos nulos representados por una máscara. Esto cambia su normalización y presupuesto de masa; no es equivalente a escalar ambos K y V. La suave decaimiento de w no establece un mecanismo de aliasing ni predicciones superiores. La implementación disponible redondea el corte de coordenada inversa a un entero y utiliza una máscara densa de cuatro dimensiones con atención inmediata. Su consumo de memoria no representa un ahorro medido en la caché KV.

El test archivado utilizó tres modelos Pythia, longitud 4096, 30 bloques, WikiText-2 y bf16 en el script disponible. El contexto de entrenamiento de Pythia es de 2048 tokens, por lo tanto, esto es una prueba de 2× extrapolación. La base de atención completa elevada y las mejoras bajo algunas máscaras pueden reflejar una mitigación del fracaso en la extrapolación; no establecen una compresión en contexto de KV. Una prueba igualada en otro aspecto a 2048, junto con mediciones reales de expulsión, memoria y latencia, son necesarias para esa afirmación. Las cotas para ambos 1.4B ($h = 2447$) y 2.8B ($h = 2754$) superan a 2048; la posición de corte en sí sola no explica sus diferentes signos. Las cuatro políticas registradas se conservan a continuación. La columna etiquetada como hard D_f registra la ventana seleccionada por ese script, no un cuantil empírico de la matriz de atención real.

Cuadro 19: Todos los valores PPL guardados para la comparación de decaimiento suave. Valores más pequeños son mejores.

Modelo	g	h	baseline	corte rígido	suave	rígido D_f
pythia-2.8b	0.674	2754	113.533	120.232	118.843	120.910
pythia-1.4b	0.705	2447	53.657	31.655	25.161	48.386
pythia-1b	0.931	505	79.087	25.261	26.721	19.955

Para Pythia-2.8B, la descomposición suave reduce el PPL en 1.155 % en comparación con el corte duro, pero lo incrementa en 4.678 % en relación con la base de atención completa. La antigua "21 %" solo puede recuperarse como una 20.729 % reducción del exceso de PPL sobre esa base, un denominador diferente. Pythia-1.4B tiene una 20.514 % reducción relativa de PPL frente al corte duro. Pythia-1B favorece la política de corte duro probada D_f. El menor PPL registrado entre todas las cuatro políticas es por tanto la base para 2.8B, la descomposición suave para 1.4B y el corte duro D_f para 1B.

El script disponible reinicia la semilla de muestreo por política, lo cual solicita que los mismos puntos de inicio de fragmento se usen para un corpus tokenizado fijo. No almacena las pérdidas por fragmento ni los conteos reales de ejemplos finitos, y salta las pérdidas no finitas. Sus desviaciones estándar agregadas no son intervalos de confianza emparejados. La regla $h \gtrsim T_{\text{train}}/2$ fue inferida a partir de tres casos, no validada como una frontera operativa. La truncación dura específicamente en T_{train} no fue uno de estos cuatro columnas de política. Ni esta ni la explicación de que todas las políticas retienen masa de atención igual son justificadas por esta prueba.

11.7 Pruning coordenado y la auditoría de subespacio rotatorio

La hipótesis original propuso que los componentes de frecuencia de alto índice se vuelven redundantes más allá L_{crit} Sus requisitos previos no están establecidos: un ajuste fallido a la ley de potencia no identifica una capa semántica, y la cota de entropía de Shannon no prescribe un conjunto suficiente de coordenadas posicionales. Las ablationes guardadas se conservan como intervenciones coordenadas con las limitaciones mencionadas a continuación.

Qué cambia el script disponible. Para la proyección fusionada de NeoX, representada como $W[h, t, r, c]$, donde $t \in \{Q, K, V\}$, r es la coordenada de salida y c la de entrada, la operación archivada es

$$W^{(\ell)}[h, Q, r, :] = W^{(\ell)}[h, K, r, :] = 0, \quad \ell \geq L_{\text{prune}}, \quad r \in \{2k, 2k+1 : k_0 \leq k < d_{\text{head}}/2\}. \tag{55}$$

Estas son filas de salida en la convención lineal-peso almacenada, no columnas de entrada. El código calcula k_0 a partir de la dimensión completa de cabeza y umbral π, sin consultar la dimensión parcial rotatoria o emparejamiento.

Cuadro 20: Snapshots de configuración local: conteos de ciclo completo en $T = 2048$, $\theta = 10000$ Utilizando solo el subespacio rotatorio. Un conteo de ciclos no es una clasificación de utilidad.

Pythia	Q heads	d_{head}	rotary dims	rotary pairs	cycled pairs
70m	8	64	16	8	6
1b	8	256	64	32	21
1.4b	16	128	32	16	11
2.8b	32	80	20	10	7

Las cuatro configuraciones congeladas de Pythia establecen `rotary_pct=0.25`. En la implementación NeoX inspeccionada, la rotación actúa sobre las primeras $r = d_{\text{head}}/4$ coordenadas Q/K, emparejadas como $(k, k + r/2)$; el resto bypassa la rotación. Para el umbral registrado de 1B $k_0 = 91$, el script mascara las coordenadas 182–255, completamente fuera del subespacio rotatorio de 64 dimensiones. Para 2.8B, $k_0 = 29$ mascara 58–79, fuera del subespacio de 20 dimensiones. Por lo tanto, estas máscaras no pueden denominarse eliminación de planos RoPE no ciclados bajo las configuraciones auditadas. El código exacto instalado y la configuración de la ejecución histórica no están registrados; el actual audit de fuente/configura identifica la inconsistencia sin reconstruir esas ejecuciones.

La fracción teórica correspondiente de entradas de peso Q/K establecidas en cero es $(N_L - L_{\text{prune}})/N_L \, (d_{\text{head}}/2 - k_0)/(d_{\text{head}}/2)$. Con cinco de 16 capas y 37/128 pares de

coordenadas seleccionados, es 9.033 %, no 14.5 %. Establecer en cero entradas de una matriz densa no modificada no proporciona por sí mismo un ahorro medible en memoria, latencia o cálculo.

Cuadro 21: Pruning sweeps completados en la versión guardada v2. Las diferencias en la precisión son fracciones, p. ej. −0.14 Es una caída de 14 puntos porcentuales.

modelo v2	L_{prune}	Q/K %	PPL	Δ PPL	exactitud	Δ exact.
pythia-1b base	—	0.000	14.9736	0	0.570	0
pythia-1b	11	9.033	16.5365	+1.5629	0.405	-0.165
pythia-1b	13	5.420	15.5490	+0.5753	0.550	-0.020
pythia-1b	15	1.807	15.1637	+0.1901	0.540	-0.030
pythia-1b	16	0.000	14.9736	+0.0000	0.570	+0.000
pythia-2.8b base	—	0.000	11.4973	0	0.615	0
pythia-2.8b	17	12.891	14.0789	+2.5816	0.425	-0.190
pythia-2.8b	19	11.172	13.2943	+1.7969	0.485	-0.130
pythia-2.8b	21	9.453	12.8873	+1.3899	0.475	-0.140
pythia-2.8b	23	7.734	12.4181	+0.9207	0.530	-0.085

El registro v1 de Pythia-1B conserva las filas en 13, 15 y los cortes no-op en 16, 17 y 19. Sus filas de base repetidas no son repeticiones independientes. Los registros v2 contienen las barridas completas mostradas anteriormente, más sus pérdidas registradas en base. Ambos scripts disponibles v2 solicitan 100 bloques de 1024 tokens y 200 ejemplos de LAMBADA. La última métrica corresponde a todos los subtokens objetivo correctos bajo argmax forzado por el maestro, no a una nueva corrida de benchmark de generación.

Cuadro 22: Regresiones descriptivas de la variación del PPL en función del porcentaje de entradas Q/K igualadas a cero. La correlación de v1 en dos puntos no tiene significancia replicable.

Datos seleccionados	n	pendiente OLS	intercepto	r	pendiente de origen
1B v1, puntos positivos	2	0.106627	-0.002562	1.000000	0.106060
1b v2, puntos positivos	3	0.189964	-0.253495	0.969389	0.153883
1B v2, including no-op	4	0.168560	-0.103117	0.969777	0.153883
2.8b v2, puntos positivos	4	0.313581	-1.561517	0.988316	0.167242

Todos los ajustes en esta tabla utilizan los valores almacenados sin redondear y el porcentaje de entradas Q/K como predictor. Cambiar las filas de base de barrido o no-op incluidas cambia la pendiente. Ni $r > 0.99$ para ambos modelos ni un coeficiente común 0.175 ± 0.005 se reproducen. El intervalo reportado no fue un intervalo de confianza. Las observaciones de costo positivo desfavoran la propuesta original de capa sin costo; no establecen una ley lineal universal ni un conjunto privilegiado de componentes de frecuencia geométrica.

Relación con métodos conscientes de frecuencia. FASA [44] es un enfoque empírico separado. Estas ablationes no demuestran equivalencia con sus bloques de frecuencia, establecen que solo deben retenerse los planos no ciclados o proporcionan una sustitución libre de calibración. Un seguimiento posterior debe inspeccionar primero el subespacio

rotatorio real y el emparejamiento de tensores, luego utilizar máscaras emparejadas, presupuestos de intervención iguales y evaluación a nivel de tarea. La propuesta y todas las afirmaciones numéricas anteriores permanecen en la fuente archivada; no se infiere ninguna novedad ni prescripción de implementación de este revision.

12 Extensión del contexto: Álgebra de Candidatos Inversos y Pruebas Empíricas

La candidata basada en configuración Padpuede invertirse exactamente. Su inversa no establece un óptimo para la escalado del kernel neural-tangent, la recuperación o la perplexidad. Para un candidato objetivo especificado $-1 < g < 1$ y $T > 0$, define

$$\alpha_{\text{cand}} = \frac{\theta_{\text{design}}(g,T)}{\theta_{\text{nominal}}}, \qquad \theta_{\text{design}}(g,T) = \frac{T\sqrt{2}(1+g)}{2(1-g)}. \tag{56}$$

La antigua notación α_{opt} es ambigua: algunas tablas históricas utilizan este candidato ratio, mientras que los scripts B1 recortan una expresión exponencial diferente en 64. Ninguna notación establece un óptimo de tarea. La expresión exponencial $e^{C/g}/\theta_{\text{nominal}}$ y la inversa lineal $T\sqrt{2}/(1-g)$ son funciones diferentes; ninguna puede reemplazar la Ecua (56).

Para $T = 2048$ y $\theta_{\text{nominal}} = 10{,}000$, un objetivo $g = 0.5$ da $\theta_{\text{design}} = 4344.464$ y $\alpha_{\text{cand}} = 0.434446$. Un objetivo $g = 0.85$ da $\theta_{\text{design}} = 17860.574$ y $\alpha_{\text{cand}} = 1.786057$. Estos corrigen la aritmética redondeada previamente impresa como 4346 y 17 858. El candidato nominal es aproximadamente 0.7470; reducirlo a 0.5 amplía la familia ideal de ley de potencia. No establece que el modelo correspondiente entrenado se vuelva más global o más local.

Relación con métodos de extensión del contexto. En la convención NTK-aware descrita por Peng et al. [35] un parámetro de contexto s cambia la base por $s^{d_r/(d_r-2)}$ para dimensión rotativa $d_r > 2$ Por lo tanto, un multiplicador de base α corresponde algebraicamente a $s = \alpha^{(d_r-2)/d_r}$ en esa convención; no es s A sí misma. YaRN combina un programa de frecuencia dependiente con la reescalación de atención y tiene su propia evaluación empírica. Una proporción de bases de RoPE no es automáticamente el parámetro de escala de esos métodos. Ni el ajuste de una escala reportada ni el cambio de un campo de configuración por sí solos demuestran la transformación rotativa correcta en tiempo de ejecución. Las implementaciones específicas del modelo y sus cachés deben ser verificados en cualquier nuevo experimento.

12.1 Diseñar una base de candidatos para un exponente especificado

Resolver la ecuación racional da

$$\boxed{\theta_{\text{design}}(g,T) = \frac{T\sqrt{2}(1+g)}{2(1-g)}}. \tag{57}$$

Su dominio es $-1 < g < 1$, con la subfamilia de exponentes positivos $0 < g < 1$ utilizada en la tabla ilustrativa. El límite cuando $g \to 1^-$ es infinito para este candidato. Esto no es un teorema que los modelos entrenados no puedan tener exponente uno en una base

Cuadro 23: Bases inversas-candidato calculadas a partir de la Ecua (57), redondeadas al entero más cercano. Los objetivos son parámetros ilustrativos de familia, no recomendados como configuraciones para una tarea.

g	θ_{design} a $T = 4096$	θ_{design} a $T = 8192$
0.50	8,689	17,378
0.65	13,654	27,308
0.75	20,274	40,548
0.85	35,721	71,442
0.95	112,956	225,912

finita. Los nombres de tareas previamente asociados a estos objetivos se eliminan porque la fórmula numérica no valida esas asignaciones de tareas.

Para referencia, $\theta = 10{,}000$ en $T = 4096$ da $\gamma_P = 0.550831$, no 0.75. El mismo candidato en $T = 8192$ da aproximadamente 0.2664. Todos los celdas de la tabla se generan a partir de la ecuación antes de redondear; la tabla archivada registra los valores desplazados.

Prueba un entorno de candidato. Especifica la tarea y el contexto objetivo antes de elegir entornos. Establece una línea base, implementa y verifica la transformación rotativa real, y mide tanto los cambios en el perfil como la calidad de la tarea bajo datos y precisión coincidentes. Los umbrales anteriores sugeridos $g \leq 0.85$ y $\alpha \leq 8$ no son restricciones universales de estabilidad. Los experimentos de recuperación conservados, incluyendo su degradación en contexto corto a una escala más grande probada, se aplican a las condiciones reportadas en la Sección 18 y la Figura 17. El flujo completo anterior y la callback permanecen en el archivo separado del predecesor indexado por el Apéndice J.

Relación con la compresión. Un candidato objetivo no es un exponente medido. Calcula una ventana de masa a partir de un perfil medido en el soporte intencional, o etiquétalo explícitamente como una familia hipotética de cálculo. La validación de la tarea sigue siendo necesaria incluso si el exponente ajustado coincide con el candidato objetivo.

12.2 El horizonte de información anterior como coordenada inversa

La función anteriormente llamada horizonte de información es

$$d_{\text{inv}}(\theta, g) = \theta\sqrt{2}\,\frac{1-g}{1+g}, \qquad \theta > 0, \quad -1 < g < 1. \tag{58}$$

Es el inverso del candidato con respecto a la longitud del contexto. En particular, $d_{\text{inv}}(\theta, \gamma_P(\theta, T)) = T$ y $d_{\text{inv}}(\theta_{\text{design}}(g, T), g) = T$. Estas igualdades son identidades algebraicas, no predicciones independientes de la longitud del contexto de entrenamiento o del fracaso en la recuperación. A una longitud fija g, $d_{\text{inv}}(\alpha\theta, g) = \alpha d_{\text{inv}}(\theta, g)$, pero una intervención no necesita preservar el exponente ajustado. Para $g > 1$ esta expresión es negativa y no define una longitud de contexto física.

El ejemplo anterior $\theta = 10{,}000$, $\widehat{\gamma} = 0.748$ da aproximadamente 2039 tokens. Su proximidad al contexto de entrenamiento de 2048 tokens no establece un techo universal. EXP-B1v4 registra 88 % en longitud nominal de 3072 y 10 % en 4096 sin escalado base; la afirmación de que valida $L^c = 2d_{\text{inv}}$ es retraída. La distancia esperada de una aguja

uniforme $L/2$ sola no puede determinar una curva de recuperación o umbral de colapso. No se sigue ninguna afirmación general de imposibilidad para el sintonizado, la decodificación o los cambios en el prompt.

Para $z = T\sqrt{2}/\theta > 0$, el mismo álgebra da

$$\log(d_{\text{inv}}/T) + \log(z/2) + 2\,\text{arctanh}\, g = 0.$$

La constante depende de la definición establecida de z. Es una identidad de coordenadas, no una invariancia empírica de un modelo.

12.3 Dirección futura: prueba de un controlador de retroalimentación

El termostato original aumentó θ cuando un exponente medido excedió un objetivo ilustrativo. Esa dirección entra en conflicto con la derivada positiva del candidato en la Sección 3 Una corrección de signo basada únicamente en el candidato se utilizaría en su lugar mediante el *unvalidated* rule

$$\theta_{t+1} = \begin{cases} \theta_t/(1+\eta), & \widehat{\gamma}_t > g_* + \delta, \\ \theta_t, & \text{otherwise}, \end{cases} \qquad \eta > 0. \tag{59}$$

Esta regla mostrada registra la dirección implícita por el candidato; no es una implementación recomendada. Corregir ese signo no establece la respuesta de una red entrenada, estabilidad, reducción de picos de pérdida o preservación de la recuperación. Las constantes propuestas $g_* = 0.85$, $\delta = 0.05$ y $\eta = 0.1$ fueron elecciones ilustrativas sin un rango operativo universal validado.

Un estudio controlado útil mediría la respuesta a pequeños cambios en la base, estimaría la incertidumbre y el retraso, y luego compararía un controlador prespecificado con bases fijas y líneas de base coincidentes. Se reportaría la inestabilidad, la pérdida, los puntajes de tarea y la sobrecarga. No se afirma ningún experimento de entrenamiento nuevo o implementación de callback en esta revisión. El pseudocódigo original se conserva en el archivo previo separado indexado por el Apéndice J.

13 Aboliciones Coordinadas y los Límites de la Interpretación de Bandas

Las intervenciones archivadas siguen siendo evidencia sobre coordenadas particulares y procedimientos de medición. Su anterior interpretación como una separación semántica/posicional no está respaldada por la auditoría de implementación. Tres operaciones distintas deben mantenerse separadas: las ablaciones de la fila B3 en Pythia-1B, el script de amplificación en Pythia-70M, y los barridos de poda de coordenadas en la Sección 11.7. No utilizan las mismas máscaras.

13.1 Planos rotatorios y nombres históricos de grupos

Definición 13.1 (Agrupamiento de planos rotatorios)**.** Sea $d_{\text{rot}} = 2m$ la dimensión rotatoria de una cabeza. En la base NeoX inspeccionada, el plano k contiene coordenadas $(k, k+m)$, $0 \le k < m$; las coordenadas $d_{\text{rot}}, \ldots, d_{\text{head}} - 1$ bypassan la rotación. Una clasificación completa en ciclo en T utiliza las frecuencias reales y la condición $T\omega_k \ge 2\pi$. Es

un conteo geométrico, no una medida de información, utilidad o función semántica. Las palabras `zero_dead` y `zero_alive` a continuación identifican condiciones archivadas solo.

Para la configuración congelada Pythia-1B, $d_{\text{head}} = 256$, $d_{\text{rot}} = 64$, $m = 32$ y hay ocho cabezas de atención. El script B3 utiliza esas dimensiones de cabeza, pero cero pares adyacentes en lugar de los planos divididos a la mitad. Para cada cabeza, localiza correctamente los bloques fusionados de pesos Q y K y selecciona coordenadas 0:46 para `zero_dead`, o 46:64 para `zero_alive`, usando intervalos semicerrados. La primera máscara elimina ambas filas de peso de 14 planos rotatorios y una fila de cada uno de los otros 18; la segunda elimina una fila de esos mismos 18 planos. Por lo tanto, las máscaras son intervenciones coordenadas desiguales, no conjuntos complementarios de planos de frecuencia intactos. Las biases de proyección y los pesos V no se cambian; las filas de peso nulas no garantizan por sí solas activaciones nulas post-proyección. Estas afirmaciones se verificaron ejecutando la máscara archivada y la restauración Las funciones actúan sobre pesos sintéticos. No reconstruyen la ejecución histórica.

13.2 EXP-B3: todas las ventanas y ambas evaluaciones guardadas

El script disponible B3v2 solicita 200 muestras por condición/ventana y 32 tokens de objetivo forzados por muestra en el conjunto de prueba WikiText-2. Dentro de una ventana fija, reutiliza la misma semilla aleatoria y la misma regla de muestreo de objetivo entre condiciones. Entre ventanas, el intervalo de muestreo cambia, por lo que los objetivos muestreados no están emparejados. El prefijo crece desde W to $W + 31$ durante el bloque objetivo; W Es la longitud del contexto inicial. El JSON almacena la PPL agregada, no las pérdidas por objetivo individual o intervalos de confianza.

Cuadro 24: Guardado completo B3v2 PPL sweep en Pythia-1B. Menor es mejor; los nombres de condiciones son etiquetas históricas de máscara de coordenadas.

Comenzando W	baseline	cero_muerto	cero_vivo
32	27.743402	377.737390	37.088815
64	21.773069	386.347909	29.763744
128	16.954068	382.877404	35.995311
256	16.020234	444.796639	108.266915
512	14.211691	512.998765	208.897918
1024	12.960886	551.712647	316.583773

Define el beneficio contextual de manera consistente como

$$B = \text{PPL}(32) - \text{PPL}(1024).$$

La base line da $B = 14.782516$, mientras que las dos máscaras archivadas dan -173.975257 y -279.494958, respectivamente. Los aumentos corto ventana PPL son 349.993988 y 9.345413; su razón es 37.45. Esa razón no es una relación señal-ruido ni un efecto normalizado para tamaño de intervención. A $W = 1024$, los aumentos son 538.751761 y 303.622887. De hecho, `zero_dead` tiene un PPL más alto que `zero_alive` en cada ventana guardada. Los diferentes cambios en el beneficio del contexto no demuestran destrucción selectiva de la semántica versus posición. La posición sigue afectando la predicción a $W = 32$, y el PPL de largo contexto también depende del contenido.

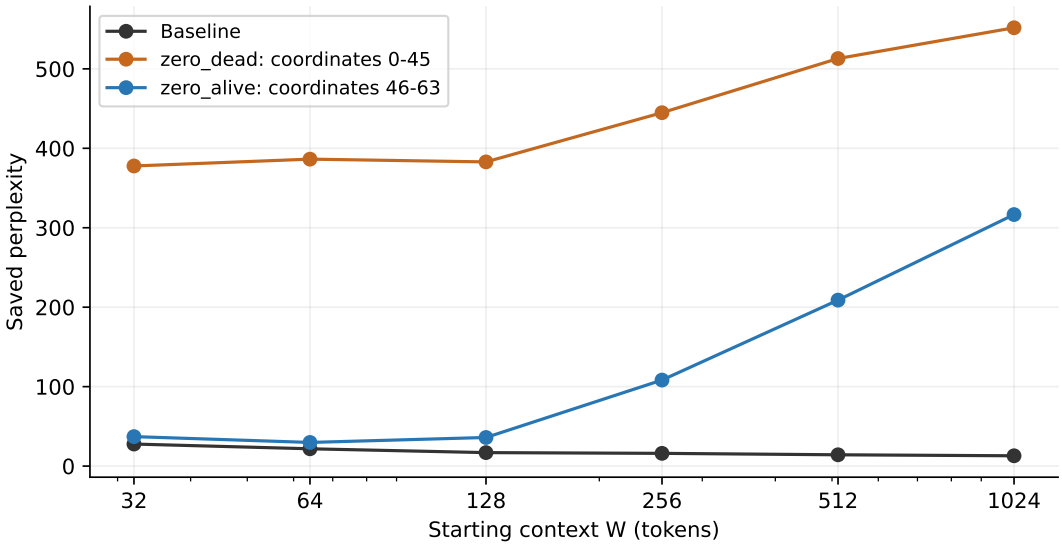


Figura 13: Guardado completo de la ventana B3v2. Las etiquetas de las curvas identifican máscaras coordenadas, no canales semánticos o posicionales. Los puntos provienen de registros agregados; no se pueden reconstruir barras de incertidumbre a partir de esos agregados.

El registro B3 anterior probó las distancias de passkey 50, 200 y 500: la precisión fue nula en todas las tres condiciones. Sus accuracies en LAMBADA fueron 0.005, 0 y 0.005 para baseline, `zero_dead` y `zero_alive`, con $n = 200$; los valores almacenados PPL fueron aproximadamente 454866, 283545 y 2038171. Estos resultados a nivel basal no pueden discriminar las funciones propuestas. Se conservan por separado de B3v2, en lugar de ser descartados o presentados como corroboración. Ambos archivos JSON originales acompañan esta revisión.

13.3 La simulación del puntaje y el registro de amplificación de la GPU

La simulación histórica del puntaje utilizó $s(d) = \sum_k \sqrt{E_k}\cos(d\omega_k)$ y escaló un grupo seleccionado de índices. No es un modelo de probabilidad de atención: las fases aprendidas Q/K, los coeficientes dependientes de la entrada y el denominador softmax causal están ausentes. Su resumen original se mantiene a continuación como informe de simulación, junto con sus cortes históricos de índices. Esos cortes no son los conteos verificados de Pythia en el plano rotatorio. En consecuencia, una pendiente simulada cercana a cero después de eliminar un grupo no puede validar un mecanismo neuronal.

Cuadro 25: Resumen del modelo de puntuación histórica $T = 2000$ La columna old-fit es una medición reportada por separado; las salidas de la simulación no son resultados de softmax.

Modelo	θ	cutoff	ajuste antiguo	$g(0)$	$g(2)$	$g(3)$	candidato
Pythia-70M	10K	23	.748	.113	.484	.685	.752
Pythia-160M	10K	23	.511	.118	.474	.648	.752
Pythia-1B	10K	91	.931	.120	.423	.549	.752
SmolLM2-360M	100K	18	.969	.059	.210	.257	.972
Qwen2.5-0.5B	1M	15	1.028	.043	.218	.282	.997

13.4 Lo que realmente cambia el código de amplificación

El panel de GPU guardado contiene cuatro configuraciones de amplificación en Pythia-70M. Reajustar los arrays de atención media almacenados reproduce sus pendientes. R^2 El script disponible utiliza 50 prompts y longitud de secuencia. $d + 200$ muestreando la última consulta y clave 199. Su nulo uniforme-causal es por tanto $1/(d + 200)$ con pendiente 0.439001 en su cuadrícula de ocho puntos, no la E4 $d+50$ La etiqueta histórica "via E4" ocultó esta diferencia en el protocolo.

Cuadro 26: Registro de amplificación regrabado. La máscara disponible cambia Q, K y V de manera desigual; no se establece un efecto causal solo rotatorio.

α	ajuste medido	R^2	simulación de puntaje
1	0.769227	0.986052	0.33
2	0.727302	0.981538	0.48
3	0.667649	0.971580	0.69
5	0.702202	0.906031	1.03

Existe un error de intervención separado. El script trata las filas fusionadas de NeoX como tres bloques contiguos a nivel del modelo de Q, K y V, pero el diseño real es cabeza, Q/K/V, coordenada. En un tensor sintético con las dimensiones configuradas, la función escala 276 filas de Q, 230 filas de K y 230 filas de V por capa. Las cabezas de Q 0–5 están afectadas, las cabezas de K/V 0–4 están afectadas, y las cabezas 6–7 permanecen inalteradas. En cada componente afectado selecciona las primeras 46 coordenadas; Pythia-70M tiene solo 16 coordenadas rotativas por cabeza. Esto no es una operación exclusiva de Q/K ni una operación exclusiva de rotación.

Los cambios observados en relación con $\alpha = 1$ son aproximadamente -0.041926, -0.101578 y -0.067026. Su signo es opuesto a los aumentos de la simulación 0.15, 0.36 y 0.70. El hecho de que la base sea la mayor entre los cuatro ajustes no establece un óptimo: no se proporciona ningún objetivo de calidad de tarea, búsqueda local de dos lados o incertidumbre, y las entradas guardadas de PPL de WikiText son NaN. La afirmación de que los pesos entrenados ocupan un óptimo geométrico se retira. Alinear una fórmula de configuración no puede reparar la máscara.

Proyección de salida y escasez. El primero W_O experimento reportó menos de 1.3 % cambio de throughput después de ceroing denso y un aumento aproximadamente siete veces en PPL en tokens aleatorios. Estos reportes siguen siendo históricos, no un resultado validado de poda de frecuencia. RoPE rota Q/K, no V; las coordenadas de entrada de W_O no tienen asignación directa de “banda muerta RoPE”. Las entradas nulas en una matriz densa no cambiada no implican una reducción medida de FLOP, latencia o memoria. La sugerida 3.5× proyección de ahorro y el mecanismo atribuido a Q-ROAR no están establecidos por este registro. Un backend escaso, evaluación de tarea coincidente y un mapa de coordenadas explícito serían experimentos separados.

¿Qué requiere un seguimiento discriminativo? Lee la dimensión rotatoria y el paramiento real, verifica el diseño fusionado QKV, mantiene fijas las coordenadas V y no rotatorias, y compara máscaras de igual tamaño de planos completos con controles aleatorizados. Registra si los sesgos son intervenidos. Utiliza las mismas posiciones de objetivo en ventanas de contexto y almacena las pérdidas por ejemplo; mide tareas posicionales y

de contenido por separado. Hasta que esos tests existan, los efectos guardados respaldan la sensibilidad coordinada, no un reservorio semántico, disociación doble o regla de poda segura.

14 Perfiles de atención, razones de peso y hipótesis de circuitos

14.1 Una distribución de referencia condicional no es un detector de circuito

Para cada consulta i, elige un conjunto no vacío J_i de claves pasadas, excluyendo la propia si d^{-g} se utiliza. Normaliza los pesos observados y de referencia en el mismo conjunto:

$$p_{ij} = \frac{a_{ij}}{\sum_{k\in J_i} a_{ik}}, \qquad q_{ij} = \frac{(i-j)^{-g}}{\sum_{k\in J_i}(i-k)^{-g}}.$$

Se requiere masa observada positiva retentada. Reporte la masa excluida de sumidero/automa, selección de consulta y protocolo de ajuste. La estadística por cabeza es

$$\kappa_{h,\ell} = \mathrm{KL}(p^{h,\ell}\|q), \tag{60}$$

con cualquier promedio a través de consultas declarado explícitamente. MaxEnt caracteriza q bajo su restricción especificada; no demuestra que q sea atención con contenido semántico removido. Una divergencia KL no nula puede resultar de sumideros, estructura local, máscara o especificación incorrecta de la referencia. Una pequeña divergencia KL no identifica la función de los vectores de valor. Ninguna de estas estadísticas identifica un circuito de inducción sin evidencia separada de tarea e intervención.

El anterior recetario de cribado utilizó 50 prompts, un umbral 0.05, y afirmó que menos de 15 % de las cabezas fueron marcadas en nueve modelos. Estos son ajustes históricos de cribado y reportes, no sensibilidad ni especificidad validadas, ni una reducción seis veces mayor en el costo de descubrir circuitos. El umbral anterior 0.1 “SWA sink” también no tiene una interpretación causal calibrada. Una ventana no puede truncar una secuencia más corta que ella misma; la configuración Mistral de 4096 no puede explicar un efecto de borde de ventana en las secuencias cortas examinadas aquí. Se necesitaría un benchmark etiquetado de prueba y controles coincidentes de mascarado/no mascarado para hacer una afirmación de detección.

14.2 Medidas en el espacio de pesos: coordenadas y promedio

El grupo de scripts espectrales guardados agrupa coordenadas de salida adyacentes a lo largo de la dimensión completa de la cabeza. La versión TransformerLens utiliza columnas de tensores almacenados como (head, input, output); la versión raw-Hugging-Face utiliza filas de la matriz de pesos lineales. Para un grupo de coordenadas I_k su estadístico de producto de Frobenius es

$$E_{\ell hk} = \|W_{Q,\ell h}[I_k,:]\|_F \|W_{K,\ell h}[I_k,:]\|_F.$$

Es un producto de normas, no una relación señal-ruido basada en varianza. Las configuraciones Pythia inspeccionadas tienen RoPE parcial y la implementación NeoX inspeccionada utiliza parejas de mitad dividida. Los grupos adyacentes de cabeza completa archivados

no pueden ser reetiquetados como planos rotatorios verdaderos invirtiendo su razón. Para las exportaciones de TransformerLens, la conversión/permutación debe verificarse además; su versión histórica no está congelada en el registro de experimentos.

Sean $a_{\ell h}$ y $b_{\ell h}$ las medias de $E_{\ell hk}$ por debajo y por encima del umbral archivado. El resumen guardado es

$$M_{\text{hist}} = \frac{1}{LH} \sum_{\ell,h} \frac{a_{\ell h}}{b_{\ell h}}, \qquad R_{\text{hist}} = 1/M_{\text{hist}}.$$

En general, ni uno iguala la razón media en dirección opuesta o una razón de medias generales. Para Pythia-1B, $M_{\text{hist}} = 0.900342$, $R_{\text{hist}} = 1.110689$, mientras que la media de b/a es 1.113334 y la razón de medias generales de grupo a/b es 0.894128. Para Pythia-70M, los correspondientes M_{hist} y la razón de medias generales son 0.872883 y 0.800258. La convención de agregación importa.

Cuadro 27: Ocho registros de razones de coordenadas guardados. Los umbrales cuentan grupos de cabeza completa adyacentes, no planos rotatorios verificados. La última columna es la razón de las medias totales de los grupos.

Modelo	d_{head}	cutoff/grupos	M_{hist}	$1/M_{\text{hist}}$	razón general
pythia-1.4b	128	46/64	0.884877	1.130101	0.865139
pythia-1b	256	91/128	0.900342	1.110689	0.894128
pythia-2.8b	80	29/40	0.873084	1.145365	0.848381
pythia-410m	64	23/32	0.874479	1.143538	0.803702
pythia-70m	64	23/32	0.872883	1.145629	0.800258
SmolLM2-1.7B	64	22/32	0.787101	1.270485	0.778233
SmolLM2-135M	64	19/32	0.846094	1.181902	0.853030
SmolLM2-360M	64	19/32	0.851445	1.174475	0.850211

La tabla se recalculará a partir de las razones por cabeza almacenadas y perfiles promedio; no es una nueva medición de peso. Las anteriores versiones Pythia-1B 1.38 y $M = 0.725$ no describen este registro completo de cabeza. La figura anterior de 32 barras también era inconsistente con su leyenda de 16 cabezas: el registro completo almacenado tiene 16 capas, ocho cabezas y 128 grupos de coordenadas adyacentes. La figura 14 ahora muestra ese registro con etiquetas de coordenadas y sin asignación semántica.

Otras métricas guardadas deben permanecer distintas. La anterior resumen basado en norma cuadrada utilizó una fórmula diferente, $E_k = (\|W_Q[:, 2k]\|^2 + \|W_Q[:, 2k+1]\|^2 + \|W_K[:, 2k]\|^2 + \|W_K[:, 2k+1]\|^2)/(4d_{\text{model}})$. Sus filas reportadas se conservan a continuación con su corte original y etiquetas de grupo interpretadas únicamente como índices coordenados.

Esas energías redondeadas no determinan todas las razones impresas hasta tres decimales. No pueden sustituirse al panel de producto de Frobenius. De manera similar, el resumen histórico de SVD reportado $k_{\text{obs}} = 21.5 \pm 8.7$ y un contraste de 1.03. Una predicción entera afirmada $2n = 35$ no puede ser dos veces un conteo de plano entero; una aproximación continua tendría que identificarse como tal y no suministra ningún teorema que vincule ciclos con la cobertura de energía de valor singular. El error anterior 38 % y el mecanismo geométrico inferido al comparar diferentes cantidades "SNR" son retirados.

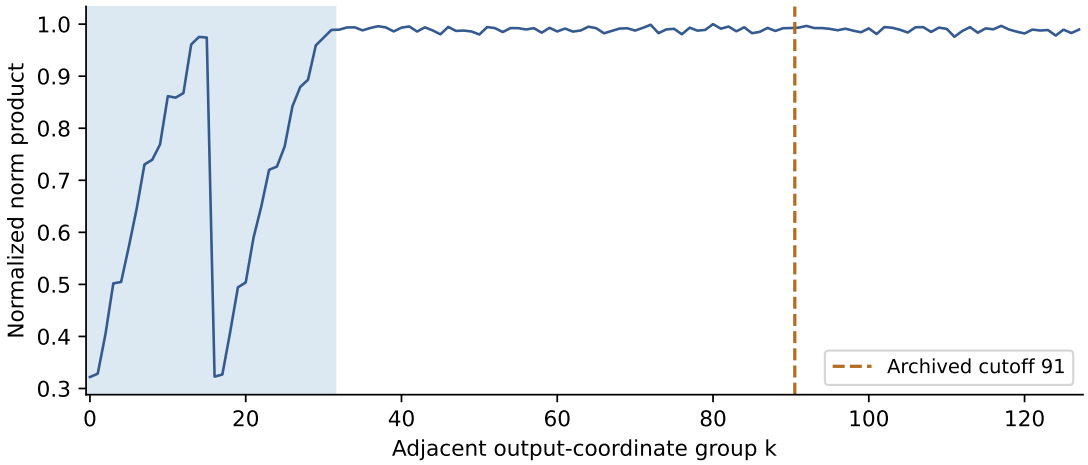


Figura 14: Perfil del producto de norma de coordenadas adyacentes de Pythia-1B guardado, promediado sobre 16 capas y ocho cabezas, normalizado por su máximo. Los primeros 32 grupos sombreados cubren las 64 coordenadas rotatorias en el orden nativo; los grupos adyacentes no son los planos rotatorios divididos a la mitad. La línea discontinua marca el umbral archivado 91. La procedencia de la base de exportación sigue siendo necesaria.

Cuadro 28: Resumen histórico de la norma cuadrada separada, conservada sin interpretación específica de frecuencia.

Modelo	cutoff antiguo	Primer promedio	segundo promedio	ratio antiguo
Pythia-70M	23	.0186	.0278	1.495
Pythia-160M	23	.0067	.0096	1.437
SmolLM2-360M	18	.0321	.0341	1.061
Qwen2.5-0.5B	15	.0008	.0008	.980

Las dos registros agregados de GQA. El archivo separado de Qwen reporta razones globales 0.903216 (0.5B) y 0.880607 (3B), mientras que sus medias de razón por capa son 0.883547 y 0.877954. Sus recíprocos reproducen los valores globales históricamente aproximados 1.107 y 1.136; estos tienen una ruta de agregación diferente a la de la tabla de ocho modelos. Dos observaciones confundidas no establecen ni la ausencia de un efecto GQA ni un mecanismo compartido universal. Los registros de entrada exactos y ambas convenciones de agregación acompañan esta revisión.

Correlaciones y significado de M. En los cinco registros conservados de Pythia, las correlaciones de Pearson con d_{head} son $r(M_{\text{hist}}, d_{\text{head}}) = +0.987556$ y $r(R_{\text{hist}}, d_{\text{head}}) = -0.986580$. La segunda es una fuerte correlación muestral negativa, sin ser una correlación parcial. Esos cinco tamaños no controlan la arquitectura, el entrenamiento ni el cambio en el agrupamiento de coordenadas. La expresión multivariada histórica $1.13 + 0.0009 d_{\text{head}} - 0.008 \log_{10} \theta$ predice 1.3284 para Pythia-1B, frente al valor conservado 1.110689; no se mantiene como predictor ajustado. Los cinco valores Pythia de M_{hist} van de 0.872883 a 0.900342 y el panel de ocho modelos, de 0.787101 a 0.900342. Al añadir las dos razones globales de Qwen se obtiene un rango heterogéneo de diez modelos que llega a 0.903216, no $[0.72, 0.90]$.

M denota una relación de norma coordenada definida explícitamente solo. $M = 1$ es paridad bajo esa definición, no un límite de razonamiento/memoria, capacidad semántica o transición de empaquetamiento fractal. La puntuación normalizada de arquitectura $\widehat{M}$ no tiene una base predictiva validada y se retira como regla de comparación. La figura 15 conserva los siete puntos históricos, con valores corregidos y ejes descriptivos; sus exponentes y razones provienen de protocolos diferentes y no constituyen una taxonomía de arquitectura.

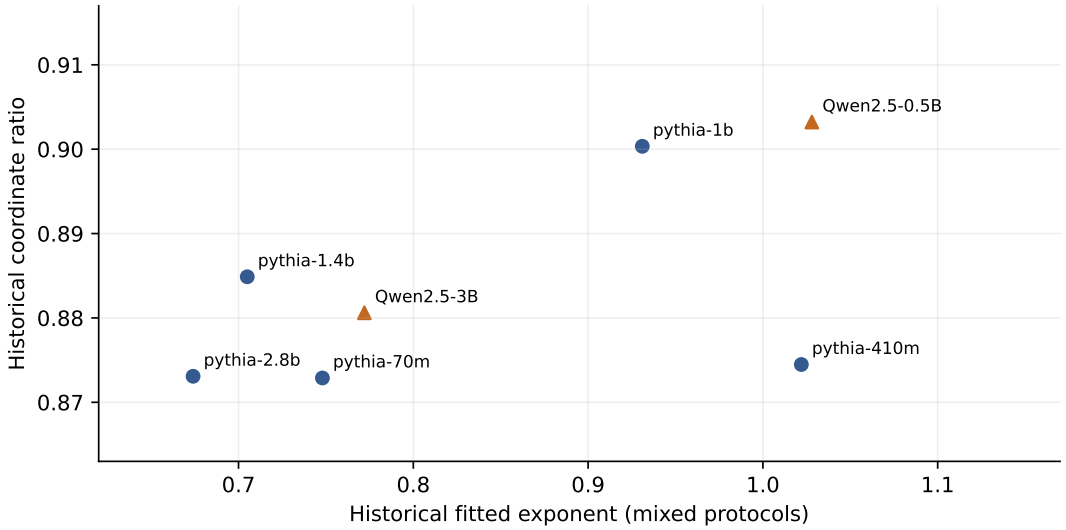


Figura 15: Siete pares de exponente/coordinada históricos. Pythia utiliza la media de las razones por cabeza; Qwen utiliza la razón global guardada por separado. Estos son puntos descriptivos con procedencia mixta, no puntajes semánticos. No se asignan regiones de fase ni interpretación causal.

Consistencia del software. El local `tools/taf/diagnose/m_score.py` no es una implementación de la estadística archivada. Promedia las normas Q/K en lugar de mul-

tiplicarlas, utiliza una fórmula diferente para el corte complementario y divide el tensor fusionado NeoX en tres bloques a nivel del modelo. Una verificación sintética de identidad de fila recupera solo 256 de las 2048 filas Q esperadas y ninguna de las filas K esperadas para la forma 1B. También omite la restricción parcial-rotary. Su salida no debe citarse como reproducción de estas tablas. Esta auditoría del manuscrito no modifica ni certifica el software desplegado; las instantáneas de código fuente y las verificaciones sintéticas se incluyen en el paquete de reproducibilidad.

Panel de exponentes históricos extendido. El mapa de 26 puntos conserva su panel histórico, con la corrección del contraste Pythia-2.8B al malla 30–100 coincidente. Sus afirmaciones sobre protocolos cruzados y arquitectura requieren reconciliación con la evidencia a nivel de fila; no se infiere umbral de inducción, mecanismo SWA o prescripción de controlador.

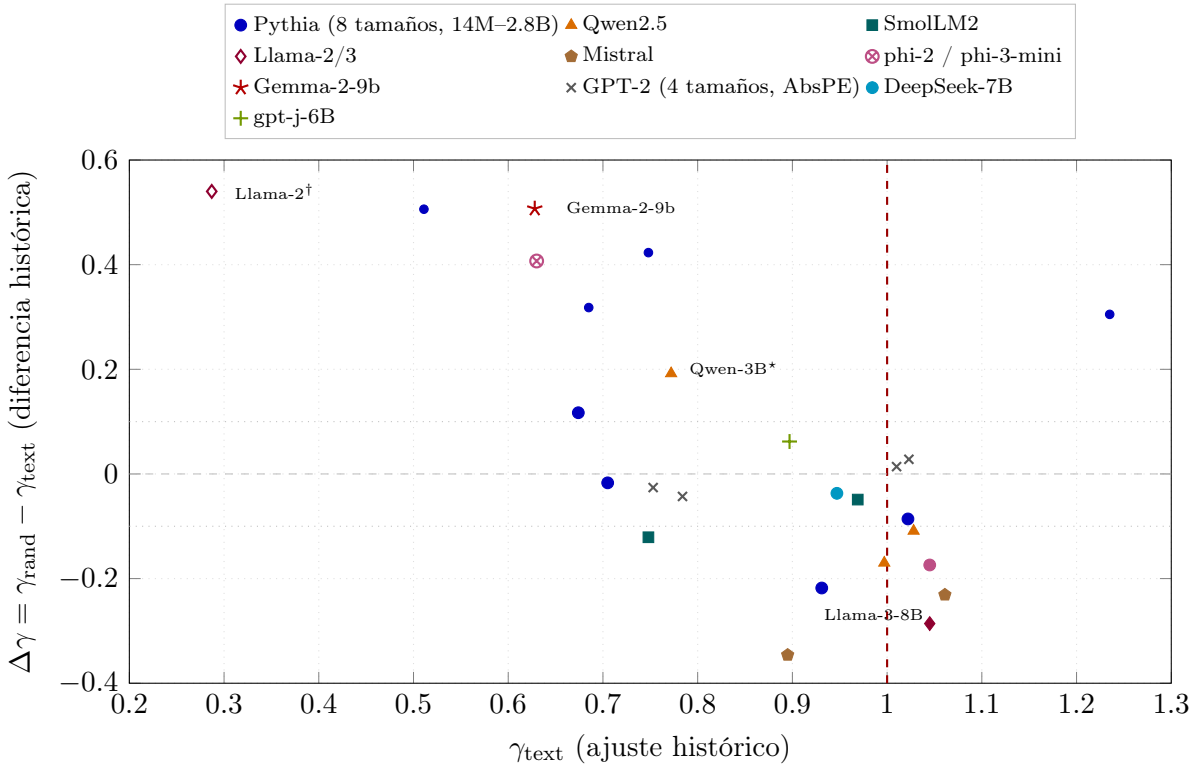


Figura 16: Panel de diferencia de exponentes de 26 puntos históricos, con la corrección del contraste Pythia-2.8B derivada de la desigualdad de soporte −.123 to +.117 en la cuadrícula común 30–100 (Sección 8.4). Los colores y marcadores agrupan las series; los valores combinan protocolos históricos y no son clasificaciones de circuitos o arquitectura. En particular, la diferencia mostrada de Qwen-3B es +0.192 la anterior leyenda −0.225 Se describe una comparación diferente y no debe sustituirse. Las líneas discontinuas son valores de referencia, no límites funcionales validados.

14.3 El campo de γ: heterogeneidad y su nulo instrumental

El extractor de campo disponible mide la última consulta en longitud de secuencia $d + 60$, siempre en comparación con la clave 59. Promedia prompts de tokens generados aleatoriamente en cada distancia, luego ajusta cada capa/cabeza por separado. Esto no

es ni el protocolo October QL fijo-consulta ni el procedimiento $d + 50$ E4. Su perfil causal uniforme es

$$A_0(d) = \frac{1}{d + 60}.$$

Ejecutar el extractor y el ajustador en matrices sintéticas causales-uniformes da $\widehat{\gamma}_0 = 0.755252$, $R^2 = 0.985521$ en $\{30, 50, 100, 200, 500, 1000, 2000\}$. Con la malla de seis puntos que termina en 1000 la pendiente nula es 0.710309. Los cinco registros guardados usan diferentes conteos de prompts y, para 2.8B, una malla diferente; esos ajustes son parte de la observación, no repeticiones intercambiables.

Cuadro 29: Se guardaron registros de campo con ajustes reales y estadísticas recalcularse de celdas finitas. Los ajustes globales y las medias de pendientes ajustadas difieren.

Pythia	$L \times H$	prompts	$d_{\text{máx}}$	global	mean	std	NaN %
1.4b	24×16	100	2000	0.690	0.907	0.463	4.688
1b	16×8	60	2000	0.753	0.943	0.331	0.781
2.8b	32×32	50	1000	0.969	1.309	0.631	0.195
410m	24×16	100	2000	0.979	1.328	0.602	37.500
70m	6×8	80	2000	1.122	1.579	0.817	37.500

Los cinco utilizan semillas 42 y 123. Los conteos en la tabla son prompts por semilla y distancia. El JSON almacena matrices ajustadas, no arreglos de atención por distancia o resultados a nivel de prompt. Las medias y desviaciones estándar anteriores se recalculan sobre celdas ajustadas finitas. El script acepta una cabeza si al menos tres medias de distancia son finitas y superan 10^{-12}; no aplica ningún umbral de aceptación R^2. Un NaN marca el fracaso de ese criterio numérico/datos, no una cabeza semántica identificada ni el fracaso de una prueba de bondad de ajuste de ley de potencia.

Pythia-70M tiene 18 celdas NaN: dos en L3 y ocho cada una en L4 y L5, utilizando índices basados en cero. L4–L5 en sí mismas representan $16/48 = 33.3\,\%$, no la cota completa de 37.5 %. Pythia-410M también tiene 37.5 % NaNs, todos los 16 cabezales de cada capa L15–L23. Este panel no identifica umbral de inducción a 400 millones de parámetros ni establece L_{crit} a partir de ajustes faltantes.

El ajuste global está por debajo de la media de celda finita en todos los cinco registros. La derivada de mezcla dependiente de la distancia en la Sección 10 explica por qué las pendientes promedio difieren del ajuste de una mezcla, pero no demuestra que estos perfiles medidos sean componentes exactos de ley de potencia o que la pendiente global reportada sea un exponente asintótico. El global `nanmean` también permite que el conjunto de cabezas participantes varíe según la distancia. Las correcciones nulas específicas de cabeza requerirían cada soporte aceptado y su ponderación; la nula común no debe restarse indiscriminadamente de las celdas ajustadas con menos puntos.

La regla de outliers en el script es la media ±2 desviaciones estándar, no umbrales fijos de 0.1 y 2. Para el registro de 1B, los cabezales guardados altos son (L0,H4) 2.204, (L1,H3) 2.772 y (L3,H0) 1.810, y el cabeza bajo es (L1,H0) 0.077. El registro de 2.8B contiene un máximo de 7.389 en L30,H0; L30 es penúltimo, no la última capa de atención. Marcar un cabeza no lo excluye del ajuste global reportado ni identifica su tarea. La asociación inversa entre la dimensión del cabeza y la dispersión del campo es descriptiva; no sigue de un argumento de suma de Riemann ni aísla la especialización inducida por el entrenamiento.

Cuadro 30: Registros completos de puntos de control guardados de Pythia-1B; la pendiente global uniforme nula coincidente es 0.755252.

Paso	ajuste global	media de celda finita	desviación estándar	NaN %
1	0.757234	0.757522	0.016958	0.000
1,000	0.851255	0.877925	0.213461	0.000
10,000	0.875856	0.885266	0.352252	0.000
143,000	0.717686	0.891705	0.338022	0.781

Los cuatro puntos de control de entrenamiento de Pythia-1B. Estos registros utilizan 100 consultas por semilla y la misma cuadrícula de siete puntos. La pendiente global del paso 1, 0.757234, está solo 0.001982 por encima del nulo uniforme correspondiente. Su proximidad a 0.75 no demuestra que haya surgido un atractor geométrico aprendido. Las cuatro pendientes globales son menores que uno: el punto de control 0.875856 no pertenece al régimen $g > 1$ del manuscrito. La dispersión entre celdas finitas sube de 0.016958 a 0.352252 y luego baja a 0.338022. Este registro muestra heterogeneidad bajo el estimador indicado; sin evidencia independiente de tareas, no mide capacidad semántica, sobreentrenamiento ni diversidad funcional. El punto de control final 0.717686 y el panel separado 0.753024 tienen distintos números de consultas y no deben reducirse a un único valor exacto.

Perfiles de clase de token: protocolo separado y reproducción. El experimento léxico Pythia-1B guardado utiliza 50 prompts de longitud 1100 y ajusta distancias. $30, 50, 100, 200, 500, 1000$. Extrae las consultas elegibles a partir de la posición 60 en adelante, a diferencia del extractor del campo last-query anterior. Su selección es estrictamente $R^2 > .5$, not $R^2 \geq .5$. Las diferentes poblaciones promedio explican por qué la media

Cuadro 31: Resúmenes léxicos retransmitidos sobre 128 celdas de capa/cabeza. Gamma mediana utiliza celdas seleccionadas; media R^2 utiliza todos los finitos R^2 cells.

Clase	PUNCT	COMMON	MID	RARE	WHITESP	NUMERIC
Median g	.612941	.754994	.651151	.853760	.830354	.921402
Células seleccionadas	50	39	47	32	92	113
Mean R^2	.412156	.351339	.384916	.310053	.609364	.738798

R^2 puede estar por debajo del umbral de selección. Las etiquetas de clase utilizan heurísticas basadas en tokens decodificados y cortes de identificadores de token (COMMON por debajo de 1000, RARE por encima de 30000), no bins de frecuencia medida del corpus. El número almacenado como consultas totales suma las contribuciones a través de distancias y cabezas; no es un conteo de consultas distintas e independientes. Los conjuntos de cabezas seleccionados difieren entre clases, por lo que sus medianas no son contrastes emparejados y no establecen multifractalidad.

Aunque la longitud del contexto es fija, el conjunto de consultas elegibles cambia con la distancia: una consulta q contribuye solo si $q \geq d$. Para la atención causal uniforme, la media específica de clase es el promedio de $1/(q+1)$ sobre las consultas de clase elegibles para esa distancia, y no necesita ser plana. Las matrices guardadas omiten las posiciones de consulta y los conteos por distancia de clase, impidiendo una reconstrucción exacta del nulo instrumental léxico. Estas limitaciones también se aplican a la interpretación de los

resúmenes exploratorios inter-familia. La dispersión mediana $.921402 - .612941 = .308461$ es descriptiva, no una dimensión geométrica medida ni una fluctuación crítica.

15 Comparaciones de Base/Instruct bajo una consulta de longitud variable

15.1 Estimando, protocolo y procedencia

Usamos $\delta\gamma = \widehat{\gamma}_{\text{instruct}} - \widehat{\gamma}_{\text{base}}$ como contraste entre puntos de control. Es distinto del contraste entre corpus aleatorio y textual $\Delta\gamma$ de la Sección 16 y del residuo frente a γ_P. Los ocho pares D3 se conservan como observaciones comunicadas; no aíslan el aprendizaje por refuerzo frente a SFT, DPO u otros cambios. Un punto de control de instrucciones no prueba por sí solo que se haya usado RLHF.

El disponible `cloud_d3_rlhf_kaggle.py` especifica 50 prompts de tokens aleatorios por distancia, semilla 42, fp16 y atención entusiasta. Genera longitudes $n = d + 60$, lee la consulta final en la posición de clave 59 (basado en cero), agrupa valores de atención de cabeza/layer finita, y ajusta OLS en espacio logarítmico en $d \in \{30, 50, 100, 200, 500, 1000\}$. Las distancias 10 y 20 son medidas pero no ajustadas. La atención causal uniforme por lo tanto produce

$$A_0(d) = \frac{1}{d+60}, \qquad \widehat{\gamma}_0 = 0.7103091963, \qquad R_0^2 = 0.984052729.$$

Este es el nulo D3 emparejado, no el $d+50$ nulo E4. La función de ajuste real y un cálculo independiente de 50 dígitos coinciden. No se ejecutó ninguna nueva inferencia de punto de control. Restar este nulo común de ambas pendientes deja su diferencia inalterada; no elimina los sumideros dependientes del modelo, los efectos de token o los valores faltantes de atención.

La tabla de ocho pares del manuscrito es el resumen numérico disponible D3. Su versión completa `d3_rlhf_results.json` no fue encontrada en los directorios de resultados locales y experimentos buscados. La columna instruct y tres entradas base conservan por tanto *procedencia del manuscrito redondeada*. Cinco valores base tienen un contraparte numérico recuperable en los arrays de tokens aleatorios de redondeo-2: SmolLM2-1.7B (0.866259), Qwen2.5-0.5B (0.906234), Qwen2.5-1.5B (0.980402), Qwen2.5-3B (0.746480), y Llama-3.2-3B (0.678947). Todos reproducen las entradas base D3 después de redondear a tres decimales. Los dos scripts disponibles comparten la regla de tokens aleatorios, $d+60$, semilla 42, 50 prompts, ajuste de cuadrícula y configuración fp16 para estos modelos. Esto establece valores contrapartes reproducibles bajo protocolos declarados coincidentes, pero coincidir en valores redondeados no prueba que D3 haya reutilizado esas corridas exactas. No reconstruye Los arrays de instrucciones faltantes o la incertidumbre emparejada. Un informe JSON de dos pares diferente del 21 de abril presenta valores distintos y sigue siendo un experimento separado. El panel planeado por el script también contiene una pareja Phi-3 (la misma checkpoint dos veces); esa entrada no es una pareja base/instrucción independiente y no se incluye en las ocho filas.

15.2 Qué las diferencias reportadas respaldan

En el punto de corte descriptivo explícitamente elegido 0.02 los conteos correctos son **dos aumentos, tres disminuciones y tres pequeños cambios**«Small» significa

Cuadro 32: Resumen del manuscrito D3, redondeado a tres decimales. Cinco entradas básicas tienen coincidencias en los reproducciones agregadas de la ronda 2 descritos en el texto; los datos emparejados completos siguen sin estar disponibles. Las categorías utilizan un umbral descriptivo, no una prueba de significancia.

Par	base	instruct	$\delta\gamma$	categoría descriptiva
SmolLM2-135M	0.913	0.922	+0.009	cambio pequeño
SmolLM2-360M	0.890	0.887	−0.003	cambio pequeño
SmolLM2-1.7B	0.866	0.916	+0.050	aumento
Qwen2.5-0.5B	0.906	0.848	−0.058	disminución
Qwen2.5-1.5B	0.980	0.973	−0.007	cambio pequeño
Qwen2.5-3B	0.746	0.698	−0.048	disminución
Llama-3.2-1B	0.829	0.757	−0.072	disminución
Llama-3.2-3B	0.679	0.730	+0.051	aumento

$|\delta\gamma| \leq 0.02$ en esta tabla, no estadísticamente indistinguible ni mecánicamente inalterado. El mayor cambio absoluto reportado es 0.072 No se muestra un par que cruce uno. Estos son propiedades de ocho estimaciones redondeadas, no límites sobre el entrenamiento posterior o evidencia para una barrera de fase. Los signos opuestos en los dos pares Llama-3.2 no identifican cuál componente de la receta causó cada desplazamiento. Un desplazamiento escalar despreciable no muestra que el aprendizaje actuara solo sobre las capas MLP o los logits de salida.

El script reinicia el generador aleatorio en cada punto de control. Las secuencias de tokens idénticas dentro de un par también requieren límites iguales en el vocabulario del tokenizador; el resumen archivado no establece todas las revisiones del tokenizador. A lo largo de las distancias, el generador avanza y las entradas difieren. Los perfiles de atención por consulta y los resultados de pérdida/tarea no se recuperaron para D3, por lo que esta auditoría no puede construir intervalos de confianza emparejados ni evaluar si el umbral elegido excede la variabilidad de muestreo. No hay registro de preregistro ni conjunto de prueba independiente que respalde la correlación anterior y las afirmaciones de alerta temprana.

Consecuencias para el monitoreo. El exponente puede registrarse junto con las métricas de pérdida y tarea bajo un instrumento invariante, pero el monitoreo predictivo requiere validación prospectiva. El anterior $C_V > 20$ la alarma de olvido catastrófico se retira: en $N = 2000$ el máximo de familia finita es aproximadamente 6.92 Ni los ocho pares ni su ordenamiento de tamaño establecen una 4×10^8 umbral de cabeza de inducción, calidad de alineación, seguridad o un efecto universal post-entrenamiento.

16 Contrastos de Corpus y la Proporción de Arquitectura No Validada

16.1 Una estadística descriptiva con una convención de signo fija

Definición 16.1 (Corpus contrastado)**.** Para un checkpoint especificado, precisión, implementación de atención, cuadrícula de distancia y regla longitud/consulta, define

$$\Delta\gamma = \widehat{\gamma}_{\text{random}} - \widehat{\gamma}_{\text{text}}. \tag{61}$$

Un contraste positivo significa un perfil aleatorio de tokens más pronunciado y un perfil de texto más plano en ese soporte; un contraste negativo significa lo contrario. Ningún signo identifica una arquitectura o un circuito de inducción. Un residuo contra γ_P Es una cantidad diferente y no debe sustituirse aquí. Es necesario coincidir con el instrumento; no hace que los corpus, los tokenizadores o las historiales de entrenamiento sean intercambiables entre modelos.

16.2 Comparaciones históricas y correcciones de arquitectura

Cuadro 33: Resúmenes separados de corpus históricos con metadatos de arquitectura corregidos. Las entradas de Falcon reportan puntajes de ajuste potencia/mixto, no exponentes. Estas filas no se agrupan con el panel de nube.

Modelo	G	ventana	aleatoria	texto	$\Delta\gamma$
Gemma-2-9B	2	4096	1.135	.628	+.507
Llama-3-8B	4	–	.931	1.046	-.115
Mistral-7B-v0.1 (fp16)	4	4096	.830	1.061	-.231
Pythia-70M	1	–	.748	.748	≈ 0
Falcon-7B (RoPE, MQA)	71	–	R^2: .88 / .999		unavailable

La comparación anterior enumera cinco modelos, no siete; Falcon no tiene un par de exponentes en esa tabla. Estos resúmenes forman un panel distinto del panel de rondas en la nube siguiente. Los dos valores 0.748 de Pythia-70M no demuestran igualdad entre corpus; para esa comparación prevalece el análisis de cuadrícula correspondiente de la Sección 8.4. Del mismo modo, los valores -0.287, -0.115 y -0.034 de Llama-3-8B proceden de registros distintos y no son estimaciones repetidas bajo un protocolo común establecido. Los valores $+0.192$, -0.225 y -0.053 de Qwen2.5-3B tampoco deben fusionarse en una firma arquitectónica única. Cada valor debe acompañarse de su registro y cuadrícula.

La configuración de referencia de Gemma-2-9B de Google tiene 16 cabezas de consulta, 8 cabezas KV. $d_{\text{head}} = 256$ y un patrón local/global alternante con ventana local de 4096; así $G = n_q/n_{\text{kv}} = 2$, not 8.[1] Falcon-7B utiliza RoPE y atención multi-query, en lugar de ALiBi, como se registra en la auditoría de configuración. El marco de ventana configurado en Mistral-7B-v0.1 también es de 4096. Para secuencias de hasta 2050 tokens en el panel E4 anterior, o 1060 tokens en el protocolo en la nube, una máscara local de 4096 tokens no excluye ninguna clave causal adicional. Por lo tanto, un corte activo de ventana deslizante no puede explicar estas diferencias medidas. El entrenamiento con atención local podría seguir afectando los pesos aprendidos; esa hipótesis separada no fue probada por estas comparaciones.

[1] Configuración de referencia https://github.com/google/gemma_pytorch/blob/main/gemma/config.py, `get_config_for_9b` Inspección realizada el 8 de noviembre de 2026.

El antiguo árbol de decisión basado en signos y la afirmación de que los archivos de configuración pueden reemplazarse por dos ajustes escalares son retirados. La arquitectura debe leerse desde la configuración exacta y la implementación. La identificación de circuitos requiere evidencia conductual e intervención, como los experimentos con modelos pequeños de Olsson et al. [34]; su trabajo no proporciona umbral universal de parámetro 4×10^8.

16.3 Rondas en la nube: estimador compartido, profundidad diferente de evidencia

Los scripts de ronda 2 y ronda 3 utilizan 50 prompts, semilla 42, $d + 60$ la regla y las mismas seis distancias ajustadas que D3. Comparan IDs aleatorios con fragmentos de texto de entrenamiento WikiText-2 contiguos, promediados a través de valores finitos de head/layer. Los fragmentos de texto se muestrean nuevamente por longitud; los prompts aleatorios y de texto no son contrapartes semánticas emparejadas. La precisión varía según el modelo, y en la ronda 4 se utiliza 4-bit NF4 para Mistral. La nula uniforme coincidente se repite nuevamente 0.7103091963.

Hay un riesgo adicional de procedencia: el camino de fallo en la carga de texto en la etapa 2/3 reutiliza el resultado aleatorio como "text". Esa rama fuerza $\Delta\gamma = 0$ sin medir el texto. No afirmamos que ocurriera en las filas exitosas que siguen, cuyas estimaciones del corpus difieren; las salidas futuras deben registrar explícitamente el éxito en la carga del corpus y tratar la caída como ausente, en lugar de neutral.

El archivo proporciona medias de distancia completas para seis modelos exitosos de ronda 2 y Mistral en ronda 4. Todos los catorce ajustes de corpus y sus valores R^2 se reproducen con precisión numérica. La ronda 3 proporciona resúmenes redondeados para Pythia-2.8B, Llama-3-8B y una repetición de Llama-3.2-3B. La última no se cuenta como modelo nuevo. Para Llama-3-8B se muestra el valor $0.821 - 0.856 = -0.035$, mientras que el contraste redondeado guardado es -0.034; conservamos el último tal como se reportó, sin inventar dígitos inaccesibles. Las cargas fallidas y modelos incompletos no son mediciones.

16.4 Recalcular la propuesta de ratio

La heurística histórica fue

$$R_c = \frac{\log(d_{\text{head}}/n_{\text{kv}})}{\log(N/N_0)}, \qquad N_0 = 4 \times 10^8. \tag{62}$$

Aqui N_0 es una escala de referencia elegida, no um umbral de inducción establecido. El numerador no es el factor de compartición GQA n_q/n_{kv} y no es una capacidad de canal medida. La razón matemática es singular en $N = N_0$; el ayudante archivado además devuelve NaN para $N \leq N_0$ o $n_{\text{kv}} \geq d_{\text{head}}$. Esas restricciones de software no son límites físicos. El ejemplo anterior de MHA incorrectamente omitió n_{kv}: con $d_{\text{head}} = 64$, ocho cabezas KV y $N = 10^9$, la razón es $\log 8/\log 2.5 \simeq 2.269$, no $\log 64/\log 2.5 \simeq 4.539$.

Tabla 34 utiliza los conteos nominales de parámetros y metadatos de la cabeza de los scripts; no certifica los totales exactos de parámetros de cada checkpoint. La fila fallida de Gemma tiene un $n_{\text{kv}} = 16$ archivado incorrecto (la configuración de referencia tiene ocho), y no se promueve a una observación exitosa. Los dos ajustes de corpus para Mistral son 0.834184 y 0.981382, dando -0.147198, pero el puntaje de token aleatorio R^2 es solo

Cuadro 34: Nueve modelos únicos. Las rondas 2 y 4 tienen medios de distancia reutilizables; las dos entradas de la ronda 3 son resúmenes redondeados. El redondeo explica la discrepancia en la resta de Llama-3-8B mostrada.

Modelo	redondeado	R_c	aleatorio	texto	$\Delta\gamma$	R^2_{rnd}
pythia-2.8b	3	0.473	0.810	0.769	+0.041	0.961
SmolLM2-1.7B	2	0.479	0.866	0.846	+0.020	0.997
Llama-3-8B	3	0.924	0.821	0.856	−0.034	0.966
Mistral-7B (NF4)	4	0.957	0.834	0.981	−0.147	0.678
Yi-6B	2	1.275	0.966	0.842	+0.124	0.980
Llama-3.2-3B	2	1.331	0.679	0.594	+0.085	0.942
Qwen2.5-3B	2	2.034	0.746	0.799	−0.053	0.988
Qwen2.5-1.5B	2	3.085	0.980	1.063	−0.083	0.959
Qwen2.5-0.5B	2	17.078	0.906	0.861	+0.045	0.991

0.677965. Su perfil aleatorio es no monótono, incluyendo un valor grande a distancia 100. Por lo tanto, es especialmente débil evidencia de una firma de ley de potencia. Su diferencia con el contraste fp16 más antiguo −0.231 no es un efecto de cuantización controlado.

16.5 ¿Qué falla y qué permanece sin probar

Utilizando el signo de los nueve contrastes reportados, los rangos mínimos-máximos son

$$R_c(\Delta\gamma > 0) \in [0.472621, 17.077567], \qquad R_c(\Delta\gamma < 0) \in [0.924358, 3.085057].$$

Su intersección de intervalos es $[0.924358, 3.085057]$. Esta es la intersección de dos envolturas de rango muestral, no una zona densamente medida, intervalo de confianza, frontera de fase o evidencia de que cada punto en ella respalda ambos signos. Los intervalos anteriores $[0.4, 1.0]$, $[0.47, 1.33]$ y $[0.5, 1.3]$ no representan esa calculación. Los umbrales de clasificación también cambian históricamente entre cero, 0.02 y 0.1; no deben mezclarse. Aquí los signos son descriptivos y no se asignan etiquetas de "pre-IH" o "post-IH" de verdad.

La afirmación "todos los modelos con $R_c \geq 2$ tienen contraste negativo" es falsa: Qwen2.5-0.5B tiene $R_c = 17.077567$ y $\Delta\gamma = +0.044788$. Su proximidad a la escala de referencia singular elegida explica la sensibilidad de la proporción, no la permisión para eliminar el contraejemplo después de verlo. Llama-3-8B y Mistral tienen contrastes negativos por debajo de uno, mientras que Pythia-2.8B y SmolLM2-1.7B tienen contrastes positivos allí. Estos modelos seleccionados heterogéneos no establecen una regla monótona o un efecto causal de compartir KV.

A valores fijos de d_{head} y n_{kv} con numerador positivo y $N > N_0$,

$$\frac{\partial R_c}{\partial N} = -\frac{\log(d_{\text{head}}/n_{\text{kv}})}{N[\log(N/N_0)]^2} < 0.$$

Así que el anterior consejo de elevar N para elevar R_c fue invertido algebraicamente. Reducir n_{kv} sí eleva esta proporción, pero el archivo no muestra que hacerlo mejore la inducción, la calidad, la seguridad o la eficiencia del entrenamiento. Las prescripciones

arquitectónicas basadas en proporciones y su extensión α_S son retiradas; el Apéndice H conserva sus definiciones y ejemplos como una construcción histórica no validada.

Un seguimiento adecuado debería revisar actualizaciones de checkpoint/config/tokenizer/library, mantener la precisión y los contextos fijos, registrar perfiles por secuencia y el estado de carga del corpus, y comparar familias de decaimiento de candidatos en secuencias de validación. Los intervalos de confianza deben resampling de secuencias, no tratar cabezas o distancias como replicados independientes. Una afirmación causal sobre KV-head también requiere una intervención arquitectónica controlada y resultados directos de circuitos/tareas. Ni un contraste de signo ni una buena ajuste log-log en muestra no suministran esas pruebas.

17 Manual de Diagnóstico y Enciclopedia del Modelo

Revisión del 7 de noviembre de 2026

El modelo enciclopedia y ejemplos de comandos permanecen para la continuidad. Sus campos de configuración, medición y recomendación deben diferenciarse. No se debe rellenar un exponente medido faltante con una fórmula de configuración ni tratar un veredicto histórico "seguro"/"requiere NTK" como una recomendación actual. La interfaz de software y los metadatos del modelo pueden haber cambiado; esta revisión no afirma que todos los comandos impresos o las salidas remotas actuales hayan sido reprobados nuevamente.

17.1 Usando la CLI de diagnóstico archivada

La instantánea del 5 de noviembre contiene `cli/diagnose_model.py` Su ejemplo de comando, en relación con ese software de checkout, es

```
python cli/diagnose_model.py --model EleutherAI/pythia-70m --fast --cpu
```

El snapshot inspeccionado define cuatro distancias rápidas $10, 50, 200, 1000$ y 30 prompts intentados por distancia, no 20. En modo completo se intentan 80 prompts por distancia en la cuadrícula de nueve puntos

$$10, 20, 30, 50, 100, 200, 500, 1000, 2000.$$

Los ajustes exitosos dependen de las salidas disponibles de atención y de muestras válidas. Su protocolo registrado utiliza longitudes $d + 50$ y la consulta final, por lo que se aplican las precauciones de coincidencia nula. El tiempo de ejecución, la memoria de CPU y el soporte del modelo dependen del checkpoint real, las bibliotecas y el hardware; las antiguas promesas universales de tres/quince minutos son retiradas. La cuantización es un backend opcional con sus propios requisitos, no una garantía para cada modelo por encima de un umbral de tamaño. Los comandos impresos son documentación de origen, no comandos ejecutados en esta auditoría.

17.2 Guía de decisión para medición

Registra la configuración exacta del punto de control, tokenizador, precisión, implementación de atención, corpus, regla de longitud/consulta, semillas, soporte de distancia

y agregación. Calcula una nula coincidente mediante el mismo estimador. Ajusta familias alternativas de decaimiento y examina residuales en secuencias de validación; una baja ley de potencia R^2 no identifica ALiBi o SWA. Reporta contrastes del corpus sin mapear sus signos a arquitecturas.

Cuadro 35: Interpretación revisada de las observaciones diagnósticas. Ninguna de estas observaciones por sí sola justifica la poda o la selección del modelo.

Observación	Seguimiento
Mala ajuste de ley de potencia	Inspeccione el soporte, máscaras, sumideros y familias competidoras; no infiera arquitectura de R^2.
Ajuste cerca o por encima de uno	Calcule cantidades de soporte finito; valide cualquier política de caché con resultados de tareas.
Contraste de corpus no nulo	Preserva ambos estimadores y su incertidumbre; verifica los metadatos coincidentes y el éxito en la carga del corpus.
Diversidad en cabeza o capa	Inspeccionar perfiles individuales; establecer la función del circuito con pruebas directas.
Gran masa cuantil	Trátalo como un cálculo distribucional, no como una garantía de contexto útil retentido.

El umbral anterior de flujo se sustituye por esta guía basada en la medición; su texto exacto y la fuente del diagrama permanecen en el archivo de revisión.

17.3 Diseño Arquitectónico: Prescripciones Escalares Retiradas

La anterior cinco-paso árbol mapeado tarea nombres a exponentes objetivo, invertido γ_P para una RoPE base, requerido $R_c \geq 2$ y se utilizó α_S para decidir si entrenar. Estas prescripciones se retiran. El panel de nube contiene un modelo de contraste positivo arriba $R_c = 17$ El umbral de inducción afirmado no está establecido; y aumentar el tamaño del modelo reduce la proporción escrita bajo sus suposiciones usuales en el dominio positivo (Sección 16.4 Un inverso algebraico de un ajuste candidato no es un controlador de arquitectura validado.

Especifica claramente los requisitos de contexto, memoria, latencia y calidad, luego compara las implementaciones reales en tareas representativas de conjunto de prueba. Un pilotaje debe medir la incertidumbre y comparar una baseline antes de seleccionar el compartimiento de valores clave, la escalabilidad RoPE o una política de caché. Estas mediciones no pueden reemplazarse por una puntuación escalar "ignición".

El propuesta de hardware asociada también carece de apoyo por parte de las ablaciones de banda archivadas: sus máscaras no seleccionan clases complementarias de planos rotatorios y una implementación cambia V (Secciones , 14.2 y 13.4). Las estimaciones históricas de 25 % cálculos de atención y 7–15 % ahorros totales de energía son propuestas, no medidas de aceleración del núcleo o garantías de preservación de la calidad. Se necesitaría correcto índice de plano, un núcleo implementado esparso y mediciones end-to-end de calidad/rendimiento.

17.4 Búsqueda de modelos con dominios de cálculo explícitos

Tabla 36 utiliza los 21 identificadores actuales del panel en lugar de mezclar valores gamma obsoletos, etiquetas de precisión no verificadas y recomendaciones de hardware.

Cada fila establece una familia de potencia. $1, \ldots, 2000$ Esta extrapolación no es una nueva medida de atención. Para hacer el lookup exactamente reproducible, las entradas son los exponentes mostrados con seis decimales de la Tabla. 10 La columna discreta es el menor entero que captura al menos .9 de la masa de esa familia; la columna continua utiliza la medida de Lebesgue en $[1, 2000]$ Ni es una prescripción de caché que preserve la calidad.

Cuadro 36: Ventanas de masa condicional $N = 2000$, $f = .9$ Véase la tabla de perfiles para soportes de ajuste y límites de adquisición.

ID	Modelo	entrada g	discreto	continuo
P01	phi-3-mini (text)	0.629563	1527	1532.07
P02	Falcon-7B	0.924685	1070	1106.76
P03	CodeLlama-13B	0.382302	1689	1689.14
P04	Gemma-2-9B (text)	0.627646	1529	1533.78
P05	Mistral-Nemo-Instruct	0.428448	1667	1667.50
P06	Pythia-2.8B	0.674162	1482	1489.31
P07	Pythia-1.4B	0.705073	1447	1456.02
P08	SmolLM2-135M	0.747996	1392	1404.20
P09	Qwen2.5-3B	0.772033	1358	1372.05
P10	Pythia-70M	0.747602	1392	1404.70
P11	SmolLM2-360M	0.969173	963	1008.40
P12	Pythia-1B	0.931108	1055	1093.12
P13	Qwen2.5-7B	0.996695	892	943.26
P14	Qwen2.5-0.5B	1.028374	807	864.95
P15	Mistral-7B-v0.1	1.060750	717	782.24
P16	Llama-3-8B	1.045476	760	821.52
P17	Gemma-2-9B (aleatorio)	1.134796	515	591.04
P18	phi-3-mini (aleatorio)	1.036602	784	844.14
P19	Llama-2-7B (Kaggle)	1.210761	334	410.37
P20	Llama-2-13B (Kaggle)	1.054530	735	798.28
P21	BLOOM-7b1	1.217824	319	395.16

El catálogo anterior de 14 filas, incluyendo sus recomendaciones no respaldadas y configuraciones de modelos inconsistentes, se conserva literalmente en el registro de reemplazo y en el PDF anterior. Los valores de Pythia-14M y OLMo de otros registros no se sustituyen silenciosamente en este panel identificado de 21 filas. Los esquemas de posición solos no eligen una familia empírica de decaimiento: compara potencias, exponenciales y alternativas mixtas en el soporte real.

18 Limitaciones y Preguntas Abiertas

Revisión del 7 de noviembre de 2026

Se conservan las limitaciones, tablas y el atlas de fallos originales, con límites adicionales indicados localmente. Se corrige la restricción exclusiva de la Fase A para el cuantil continuo finito: es finito para todo γ real y vale N^f cuando el exponente es uno. Se retiran la derivación geométrica de un exponente aprendido, la supuesta validación independiente del horizonte inverso y las garantías de tarea deducidas solo de la masa. Las 17 observaciones de octubre ausentes y la procedencia incompleta de los pesos siguen sin resolverse. Las adiciones reutilizan la evidencia conservada, sin reconstruir lo que falta por inferencia.

Listamos honestamente y en un solo lugar las principales limitaciones del marco.

L1: la derivación geométrica propuesta no está establecida. La familia de leyes de potencia sigue condicionalmente de un puntaje de distancia logarítmica o de una restricción MaxEnt prescrita; los conteos de ciclos rotatorios solos no implican ni uno ni otro. El candidato Padé no tiene una derivación universal aceptada. Los porcentajes de error de panel seleccionado no reparan sus condiciones de frontera inválidas o el paso de promedio no lineal (Apéndice B).

L2: el esquema de posición no es un diagnóstico empírico de familia. GPT-2 tiene posiciones absolutas; BLOOM utiliza ALiBi; Falcon-7B emplea RoPE/MQA. La cuantil de familia finita de potencia es matemática para cualquier exponente, condicional a que familia. Ningún codificador posicional por sí mismo valida la familia o una garantía de caché a nivel de tarea o extensión de contexto.

L3: los efectos coordenados no establecen funciones específicas de frecuencia. Las secciones revisadas 13 y 14.2 separan las máscaras B3, la amplificación GPU, el podado coordenado y las estadísticas de relación norma. Las etiquetas históricas B3 no son intervenciones de plano completo; sus dos máscaras afectan a un número diferente de coordenadas. El script de amplificación también cambia V y deja los heads sin alterar. Las resúmenes espectrales de cabeza completa no aíslan el subespacio rotatorio parcial. Sus efectos y correlaciones por lo tanto no pueden validar un reservorio semántico, un óptimo geométrico o un podado seguro. Las etiquetas originales 1.38 y 0.87 no deben fusionarse con las audité 1B $M_{\text{hist}} = 0.900342$ y su recíproca 1.110689. No se afirma ningún nuevo lanzamiento de modelo por parte de las comprobaciones del tensor sintético o las retransmisiones agregadas.

L4: la evidencia de base/instrucción es limitada. La resumen del manuscrito D3 de ocho pares tiene dos aumentos, tres disminuciones y tres pequeños cambios en el umbral 0.02 (Sección 15). Su salida bruta no fue encontrada; el script disponible proporciona la nula uniforme coincidente 0.710309. Un JSON anterior de dos pares es evidencia separada. Estas observaciones no aíslan el RLHF ni validan las advertencias basadas en atención sobre olvido o cambios de alineación.

L5: alcance matemático versus uso empírico. La cuantil de masa continua finita se aplica a cada exponente real, con N^f en uno. La $g < 1$ aproximación grande-N puede ser inexacta cerca de uno. Un cuantil de masa no garantiza la calidad de la tarea, y el

instrumento E4 de longitud variable produce una pendiente no nula incluso para atención uniforme.

L6: KV guardados, escalado de base y experimentos de recuperación. Los paneles completos de dos modelos y siete modelos B2 permanecen separados en la Sección 11. Prueban ventanas asintóticas con diferentes bloques de objetivo y no validan un intervalo universal favorable de exponente. Los experimentos B1 que siguen cambian un multiplicador nominal de RoPE antes de cargar los mismos pesos; no afinan esos pesos. Su α denota el multiplicador de base, no el factor de extensión de contexto de cada método llamado escalado NTK.

EXP-B1v3: PPL de secuencia larga y su denominador. El registro conservado de Pythia-1B incluye cuatro configuraciones básicas y cuatro bandas objetivo de 64 tokens, con índices desde cero: $[960, 1024)$, $[1984, 2048)$, $[3008, 3072)$ y $[4032, 4096)$. El script solicita hasta diez segmentos de prueba de WikiText-2 de 4096 tokens y proporciona todo el prefijo hasta el final de cada banda. Así, 4096 es la longitud de secuencia, no el número de secuencias. El JSON conserva la PPL agregada, sin las pérdidas por secuencia ni los metadatos inmutables de ejecución. Pythia-1B se entrenó con contexto 2048: las bandas A y B terminan dentro de él, mientras que C y D prueban extrapolación. Sus pérdidas sin escalar, muy grandes, no constituyen una referencia de compresión dentro del contexto.

Cuadro 37: EXP-B1v3: PPL absoluto y ratios históricos dentro del contexto. Se conservan los cuatro entornos y dieciséis agregados.

α	PPL A	PPL B	PPL C	PPL D	B/A	C/A	D/A
1	11.949	12.659	620.950	1424.147	1.059	51.966	119.183
4	13.878	14.549	16.084	12.366	1.048	1.159	0.891
16	40.117	59.789	72.969	77.539	1.490	1.819	1.933
64	79.969	143.510	172.836	219.855	1.795	2.161	2.749

Las razones históricas dividen cada configuración's B/C/D PPL por su propio A PPL. Comparan diferentes bandas de objetivo y suprimen los cambios en el denominador. En $\alpha = 4$, A aumenta de 11.949197 a 13.877564 (+16.14 %), y B de 12.659133 a 14.549244 (+14.93 %), en comparación con $\alpha = 1$ en la misma banda. Sus C y D PPL son mucho menores que los valores no escalados. Esto representa un tradeoff entre contexto corto y largo registrado, no una extensión sin costo. La histórica 1.80× en $\alpha = 64$ es B/A dentro de ese ajuste; B es en su lugar 11.34× el valor no escalado de B. No se puede recuperar ninguna incertidumbre emparejada de estos agregados.

El valor 64 proviene de recortar la expresión legada $e^{\log(10000)/0.5}/10000 = 10000$ en 64. No es ni el valor del candidato inverso 0.434446 en $T = 2048$ ni un óptimo demostrado. Los datos no contienen ninguna prueba de ajuste fino que pudiera validar esta prescripción después de la adaptación de pesos. YaRN utiliza un horario diferente y una reescalación de atención [35]; sus resultados no certifican este multiplicador base.

Los protocolos anteriores B1 permanecen distintos. El script v2 PPL recorta 124, 148 o 150 tokens de contexto más 64 tokens de objetivo y no proporciona posiciones de desplazamiento: sus entradas son solo de 188 a 214 tokens de longitud. Los grandes desplazamientos del texto fuente no prueban contextos posicionales largos. Todos los 20

agregados v2 PPL se conservan en el paquete de evidencia. El registro original NIAH contiene 25 celdas de 0/40 en cinco longitudes y cinco configuraciones básicas. Su criterio de argmax para el siguiente token, comparando solo el primer token de cada clave tokenizada por separado, difiere de la tarea posterior de clasificación de candidatos; los ceros se conservan, no se descartan como mediciones fallidas ni se combinan con v4/v5.

EXP-B1v4/v5: clasificación determinista de candidatos NIAH Los scripts obtienen una clave numérica correcta contra nueve entradas de distractor en repetidos rellenos de tema laboratorio. Las 50 fracciones de inserción forman una cuadrícula fija desde 0.1 hasta 0.9; no son posiciones de aguja aleatorias. Las claves y distractores se seleccionan determinísticamente de un pool filtrado por tokenizador. Un empate con un distractor se cuenta como fracaso. Esto corresponde a un rango de candidatos restringido, no a recuperación de generación libre o razonamiento. Los registros del modelo conservan etiquetas exponenciales históricas .748, .931, .705 y .674 para Pythia-70M, 1B, 1.4B y 2.8B, respectivamente. El script de recuperación v5 no vuelve a medir estos exponentes después de cambiar la base.

Una reconstrucción usando el tokenizador Pythia-70M congelado local produce 75 entradas en el pool pero solo 61 claves distintas; los primeros 50 ensayos usan 42 claves distintas, y algunos ensayos tienen ocho distractores distintos entre las nueve entradas. El constructor de pool no deduplica y solo utiliza su fallback de tres dígitos si se encontraron menos de 20 entradas. Todos los longitudes de contexto reconstruidas son $L + 1$ porque BOS se agrega después del presupuesto de relleno; la puntuación agrega también el token clave. Estos son propiedades del código guardado con la instantánea del tokenizador especificada, no prueba del exacto tiempo de ejecución histórico. El JSON agregado omite los identificadores de token, los buffers de frecuencia de tiempo de ejecución, las puntuaciones por ensayo y las revisiones inmutables de biblioteca/pesos necesarias para esa afirmación más fuerte.

La línea trazada 10 % es la referencia nominal para diez entradas candidatas intercambiables. Las claves determinísticas, duplicadas y de frecuencia diferente no establecen esa referencia como la precisión nula instrumental del modelo. No se guarda aquí ninguna prueba de no-needle coincidente o control de etiquetas mezcladas. Las claves reutilizadas y el relleno también impiden tratar las 50 pruebas como muestras independientes del corpus. Reportamos conteos sin fabricar significancia emparejada a partir de agregados.

Cuadro 38: EXP-B1v4: Precisión registrada por Pythia-70M en porcentaje; 50 ensayos por celda. Las etiquetas de columna son presupuestos nominales de contexto.

α	512	1024	2048	3072	4096
1	88	92	92	88	10
4	88	90	90	88	86
8	88	84	88	88	88
16	74	84	82	82	84

Cuadro 39: EXP-B1v5: panel completo de cuatro modelos, precisión en porcentaje, 50 ensayos por celda.

Model	α	512	1024	2048	3072	4096
pythia-70m	1	88	92	92	88	10
	4	88	90	90	88	86
	8	88	84	88	88	88
	16	74	84	82	82	84
pythia-1b	1	100	100	100	16	16
	4	100	100	100	100	100
	8	100	100	100	100	100
	16	100	100	92	90	88
pythia-1.4b	1	100	100	100	16	12
	4	100	100	100	100	100
	8	100	100	100	100	100
	16	100	100	98	98	98
pythia-2.8b	1	100	100	100	16	24
	4	100	100	100	100	100
	8	100	100	100	100	100
	16	100	100	100	100	100

Todos los 80 contadores de celdas v5 reproducen sus precisiones almacenadas; los 20 celdas v4 son iguales al panel Pythia-70M v5 y no se cuentan como replicación independiente. A nivel nominal $L = 4096$, las precisiones no escaladas son $10, 16, 12, 24\,\%$; a $\alpha = 4$ son $86, 100, 100, 100\,\%$. A $\alpha = 8$ son $88, 100, 100, 100\,\%$. Los tres modelos más grandes alcanzan 50/50 bajo ambos 4 y 8; este techo observado no garantiza la precisión poblacional.

No existe una dominancia estricta de Pareto por $\alpha = 4$: en Pythia-70M pierde dos puntos porcentuales frente a $\alpha = 1$ en ambos 1024 y 2048, y en 4096 queda dos puntos atrás de $\alpha = 8$. La fila de atención de Pythia-2.8B representada por $\alpha = 16$ permanece en $100\,\%$ en cada longitud probada, contradiciendo un fracaso universal de sobre-extensión en 16. Las grandes caídas no escaladas no identifican por sí mismas al aliasing rotatorio como su causa. Apoyan la comparación restringida de tareas registrada, no un óptimo derivado del exponente, una cota universal segura o el umbral $L^c = 2d_{\text{inv}}$.

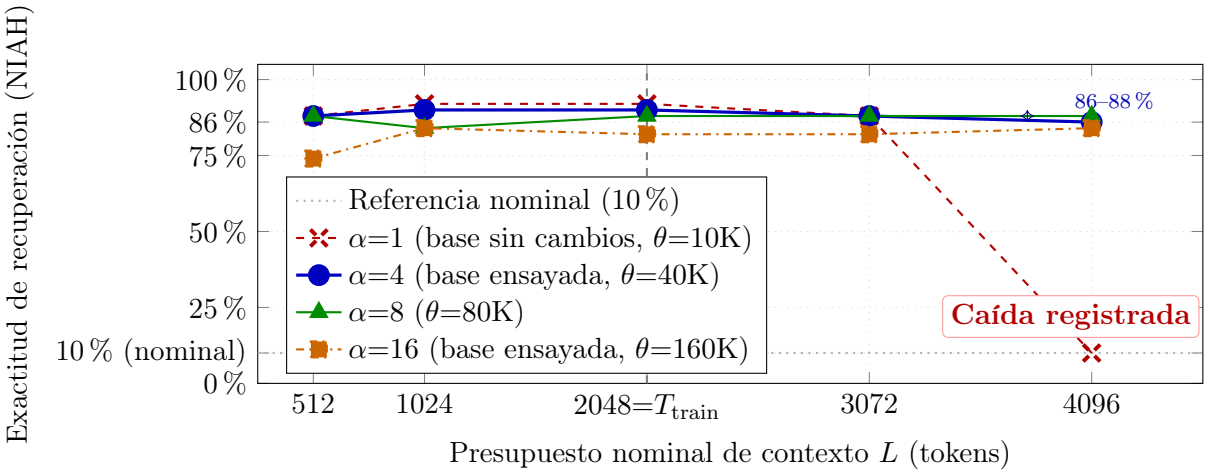


Figura 17: Exactitudes registradas de clasificación entre candidatos para Pythia-70M (EXP-B1v4; 50 ensayos deterministas por celda). El panel v5 de modelos pequeños es idéntico. Las coordenadas conservan todos los valores comunicados; α multiplica la base. La referencia nominal de diez candidatos no es una medición sin aguja. A $L = 4096$, $\alpha = 4$ obtiene 43/50 y $\alpha = 8$ obtiene 44/50. A 1024 y 2048, $\alpha = 4$ obtiene 45/50 frente a 46/50 sin cambio de base. Estas observaciones no prueban un óptimo universal ni un mecanismo de aliasing.

L7: la capacidad térmica de la familia finita no es respuesta a temperatura de muestreo. Nakaishi et al. [33] estudia correlaciones y comportamiento crítico en el texto generado a medida que varía la temperatura de muestreo de los tokens. Esa obra no establece la afirmación sobre capacidad térmica negativa anteriormente atribuida a ella aquí; la atribución se retira. Su variable de muestreo y observables difieren de la familia de distancia de soporte fijo utilizada en este manuscrito.

Nuestra cantidad de familia finita es

$$C_V(\gamma, N) = \gamma^2 \partial_\gamma^2 \log Z_N(\gamma) = \gamma^2 \operatorname{Var}_{p_\gamma}(\log d), \qquad Z_N(\gamma) = \sum_{d=1}^{N} d^{-\gamma}.$$

Es no negativo y finito para todo exponente real fijo en N finito. Solo para $\gamma > 1$ fijo tiende el normalizador de soporte infinito a $\zeta(\gamma)$; en esa familia límite C_V diverge como $\gamma \to 1^+$. Para $N = 2000$, el máximo finito de exponente positivo está en vez cerca de $\gamma = 1.256885$, con $C_V = 6.919812$. El factor γ^2 es esencial: la varianza sola es información de Fisher, no capacidad térmica. Ninguna de estas identidades suministra la respuesta de muestreo de tokens o una transición física en un modelo entrenado.

L8: la arquitectura de mixture-of-experts requiere definiciones específicas de la arquitectura. Un experto no necesita poseer un operador de atención separado. Por ejemplo, los expertos de Mixtral son bloques feedforward[19]. Por lo tanto, no se puede asumir un exponente de atención por experto a partir del término MoE. Las comparaciones de perfiles condicionadas por la ruta requerirían poblaciones de consultas explícitas y mediciones controladas; el candidato inverso no valida un horizonte por experto.

L9: una única potencia ajustada pierde la heterogeneidad. Las mediciones del campo en la Sección 14.3 conservan la variación cabeza/estrato y los ajustes faltantes. Una familia global es una elección de modelado. La capacidad térmica alta de la familia no

necesita ocurrir exactamente en uno, ni un exponente ajustado ni una puerta proxy establecen la sensibilidad de la tarea de la cabeza. Las intervenciones directas y los resultados pareados siguen siendo necesarios.

L10: las afirmaciones originales de falsificación no sobreviven a esta auditoría. El predictor geométrico universal propuesto ya falla en la comparación inicial mixta de paneles y no reproduce la comparación de octubre de QL. Esas observaciones no satisfacen el criterio anterior de error del cinco por ciento. La nula E4 también muestra que una ley de potencia ajustada no requiere geometría rotatoria. Comparar diferentes familias de modelos no puede aislar el efecto de reemplazar una codificación posicional; una prueba causal de ese tipo requiere controles de entrenamiento y medición coincidentes.

Una relación negativa entre g y la entropía calculada a partir de la misma familia finita positiva-g es algebraica: $\mathrm{d}H/\mathrm{d}g = -g\,\mathrm{Var}(\log d)$. No es una prueba empírica independiente de un mecanismo físico de entropía. Los objetos empíricos adecuados son las distribuciones medidas, los efectos de intervención y las predicciones en ausencia de datos, junto con su incertidumbre y controles.

Preguntas abiertas. (0) **Olvidar y capacidad calorífica familiar.** A valores fijos finitos N, C_V no diverge en uno. En la familia de exponentes positivos en $N = 2000$ el máximo numérico es aproximadamente 6.919812 near $g = 1.256885$ El umbral anterior $C_V > 20$ no es umbral operativo en ese contexto. Un cambio de pendiente de 0.05 También es una regla de advertencia no validada, especialmente cuando el estimador mismo es dependiente del contexto. No se establece ninguna corroboración empírica independiente de esa regla propuesta de advertencia. Probar una regla de advertencia requiere puntos de control coincidentes y resultados predefinidos en tareas de validación cruzada, no coincidencia de coeficientes o terminología.

(1) El panel inicial tiene una estimación aleatoria Pythia-31M 1.539824; no respalda el anterior límite superior de 1.35. Un máximo empírico entre modelos dependiera en cualquier caso del protocolo de adquisición y selección de modelos.

(2) El pilotaje de suma modular EXP-D1 tiene 150 puntos de control guardados. Su rango ajustado de gamma va desde .319592 hasta .456514, siempre por debajo de uno. La precisión de validación supera por primera vez .90 en el paso 3400 y .95 en el paso 3500; el umbral posterior define el paso reportado de grokking, no una transición únicamente ubicada. La prueba utiliza secuencias aleatorias de cuatro tokens y solo distancias 1, 2 y 3, promediando diferentes conjuntos de consultas en cada distancia. Tres puntos no establecen una cola de atención asintótica. El campo guardado llamado C_V es en realidad $1/|1 - g|$ (con un límite), no la capacidad de familia finita. Por lo tanto, el registro no respalda una transición de uno, pero tampoco identifica una reestructuración interna de circuitos o resuelve un mecanismo de grokking de contexto largo. El modelo de entrenamiento ha aprendido posiciones absolutas, no RoPE.

(3) Predecir el olvido a partir de un puntaje de coordenada inversa sigue sin ser probado. (4) Una relación entre el rango de vocabulario y los exponentes de distancia de atención requiere observables y controles coincidentes; no se establece una dualidad exacta.

L11: la validación downstream sigue siendo restringida. Los experimentos con la caché guardada y escalado de base utilizan PPL de WikiText-2 y clasificación con clave numérica restringida. Ninguno de ellos valida la razón multi-paso, un intervalo universal

del exponente de la caché o un óptimo para un exponente objetivo. Las medidas de PPL evalúan la pérdida predictiva en los tokens especificados, no la coherencia como resultado semántico independiente. Los efectos en matemáticas, código y preguntas de calidad en documentos largos siguen abiertos; la masa de atención descartada sola no predice la pérdida de precisión en esos casos.

L12: validación no controlada de MoE. Existe un registro aleatorio de token E4 de OLMoE-1B-7B-0924 guardado, por lo tanto la afirmación anterior de que no se recolectó ningún perfil de MoE es incorrecta. No es un estudio de intervención densa/MoE emparejado ni un análisis condicionado a expertos, y no valida una extensión de MoE de la propuesta de horizonte.

L13: El sesgo de ALiBi no es el perfil completo de atención. Un sesgo de distancia negativa lineal contribuye un factor exponencial antes de la normalización. Los puntajes de contenido, las máscaras y la precisión numérica aún son relevantes. El cero histórico de BLOOM-1b7 a distancia 5000 no puede por sí solo distinguir entre subflotación, máscara y masa positiva muy pequeña. La familia con corte exponencial es un modelo matemático condicional; su adecuación para ALiBi debe medirse, no inferirse únicamente del esquema de posición.

L14: no establecido atrayente SGD-a-candidato. El resultado de máxima entropía restringido fija una familia después de prescribir su momento de distancia logarítmica; no determina cuál exponente aprenderá SGD. La trayectoria de coordenada racional en la Sección 20 no es una trayectoria observada del optimizador. Ni un atrayente de entrenamiento ni la convergencia al candidato basado en configuración se sigue del piloto guardado. El puente dinámico requerido sigue sin demostrarse.

L15: auditoría de base controlada, instrumento y objetivo de entrenamiento. El piloto guardado contiene tres corridas de un transformador de estilo GPT-NeoX de 4 capas, con tamaño oculto 128, cuatro cabezas, tamaño de cabeza 32, dimensiones rotatorias completas y 13,659,136 parámetros. El código solicita la misma semilla 42, procedimiento de inicialización, ordenamiento de recorte aleatorio y configuraciones de AdamW, mientras cambia la base entre $10^4, 10^5, 10^6$. Establece tanto los campos de configuración de base legada como los más recientes. No hay una verificación de frecuencia de ejecución guardada ni una revisión inmutable de la biblioteca disponible para certificar su interpretación histórica.

El script limita el flujo de entrenamiento de WikiText-2 a 2,000,000 tokens y solicita 4000 pasos de ocho cortes de 512 tokens: 16,384,000 presentaciones programadas de tokens de entrada, con reutilización. Esto no es un millón de tokens de entrenamiento o 2.4 millones de tokens distintos del corpus. Las pérdidas no finitas pueden omitir pasos y AMP puede omitir actualizaciones, por lo que el horario no es un conteo verificado de actualizaciones del optimizador. Las nueve pérdidas registradas por ejecución terminan en 6.466924, 6.475903 y 6.480629. Estas pérdidas de entrenamiento no son perplexidades de validación cruzada ni certificados de convergencia.

Defecto de alineación de objetivos en el script guardado. Los cortes ya establecen $y_t = x_{t+1}$ y llaman al modelo causal de lenguaje con `labels=y`. La pérdida causal estándar desplaza internamente las etiquetas: empareja el logit en t con $y_{t+1} = x_{t+2}$ en vez de x_{t+1}. La inspección de la fuente instalada de Transformers 5.13.0 y una comprobación

de pérdida con logits sintéticos verifican esta composición; no hace falta ejecutar el modelo. La versión del entorno histórico no está fijada, así que se demuestra un defecto de la receta disponible bajo esa convención de pérdida, sin reconstruir la ejecución. El mismo desplazamiento interno figura en la implementación pública anterior.[2] El piloto no puede ser certificado como preentrenamiento estándar para la predicción del siguiente token sin resolver este desajuste.

Protocolo de medición. El extractor utiliza 30 cultivos para cada distancia. $d \in \{10, 20, 30, 50, 100, 200, 300, 500\}$ con longitud de secuencia $d+5$, todas las capas y cabezas, y la última consulta a distancia d Reutiliza el flujo de entrenamiento; las longitudes varían entre 15 y 505, en lugar de una longitud fija de evaluación de 512. El candidato se evalúa sin embargo en $T = 512$ Esto es distinto del protocolo E4 anterior. Los tres archivos por base coinciden con el JSON combinado, y todas las pendientes y puntuaciones de ajuste que se muestran a continuación se reproducen a partir de ocho medias guardadas.

Cuadro 40: Piloto de base controlada: exactitud agregada reproducida redondeada para visualización. El error relativo utiliza al candidato como denominador, no la incertidumbre.

θ	$\gamma_P(512)$	$\widehat{\gamma}$	R^2	Error relativo %)
10000	0.930122	0.977318	0.996753	5.074
100000	0.992785	0.959179	0.996762	3.385
1000000	0.999276	0.959578	0.996552	3.973

Para la atención causal uniforme, la misma entrada seleccionada es igual a $1/(d+5)$. Ajustar ese nulo en las ocho distancias da $\widehat{\gamma}_0 = 0.9103752731$ y $R^2 = 0.9978936065$. En consecuencia, una ajuste de potencia por encima de .99 y una pendiente cercana a uno no establecen una organización geométrica aprendida. Los excesos por encima de este nulo son aproximadamente .067, .049 y .049; restarlos no aísla un exponente causal. Las pendientes observadas caen de .9773 a .9592 y luego suben ligeramente a .9596, mientras que el candidato sube de .9301 a .9993. En particular, el cambio observado de 10^4 a 10^5 tiene el signo opuesto a $\partial_\theta \gamma_P > 0$ a valores fijos de T. Esto es un desacuerdo descriptivo en la dirección del candidato, no una afirmación de significancia proveniente de este pilotaje de una sola semilla, confundido por el instrumento. Los errores relativos de 3.39–5.07 % son aritmética descriptiva no es una prueba de aceptación calibrada ni una confirmación causal.

Los registros almacenados contienen una semilla, ninguna dispersión por cultivo, ningún perfil base no entrenado y ninguna validación de tarea reservada. Una explicación de "atractor subentrenado" está por tanto sin probar. Un estudio futuro útil comenzaría corrigiendo el alineamiento del objetivo, verificando las frecuencias de tiempo de ejecución y la nula del estimador, y luego usaría semillas coincidentes y mediciones reservadas. Simplemente extender la receta antigua a más pasos no resolvería estos problemas. Las pérdidas completas guardadas, las medias, los scripts y el texto desplazado permanecen en el paquete de evidencia; ningún nuevo entrenamiento o inferencia de modelo se afirma en esta auditoría.

L16: explicaciones alternativas y efectos de medición permanecen abiertos. El candidato en sí no es una explicación suficiente establecida de los perfiles observados. La

[2]Véase el bloque de pérdida de Transformers v4.44.2 fuente GPT-NeoX, Consultado 12 de noviembre de 2026. Este es un ejemplo de implementación de referencia, no una versión declarada para el experimento guardado.

coocurrencia en el corpus, la optimización, la estructura aprendida de consulta/clave, las máscaras y la normalización variable del estimador son posibles contribuyentes. El uniforme $d + 5$ nulo en L15 demuestra que una pendiente sustancial de ley de potencia puede surgir sin preferencia aprendida de distancia. Sus contribuciones relativas requieren controles emparejados y no pueden identificarse seleccionando al candidato solo a partir de la configuración. Las asociaciones cruzadas entre modelos anteriores y los resultados negativos conservados permanecen como evidencia descriptiva.

18.1 Atlas de Fallos: limitaciones observadas e identificaciones no probadas

Este atlas separa los fallos aritméticos/ámbito, las observaciones guardadas y las analogías no probadas. No asigna una causa raíz donde el experimento no identificó una. El compendio registra el alcance actual, no una lista de recetas universalmente efectivas.

Cuadro 41: Atlas de fallos corregido. Los informes empíricos y las limitaciones matemáticas tienen diferentes estatus evidencial.

Topic	Observación e implicación
Cerca de $g = 1$	La aproximación sub-uno grande-N tiende a cero a medida que N permanece fijo; El cuantil continuo finito tiende en cambio a N^f. El normalizador discreto es finito. Utiliza el cálculo adecuado con soporte finito.
$g > 1$	Ambos cuantiles finitos permanecen definidos. Una aproximación asintótica en fase A No es apropiado; el rendimiento de la caché medido sigue siendo un resultado separado.
Base/instruct	Ocho pares redondeados D3 tienen dos aumentos, tres disminuciones y tres pequeños cambios. La ausencia de salida bruta impide un replay completo o una afirmación sobre RLHF causal.
GQA ratio	Los signos observados no siguen la puerta monotona propuesta. Modelos mixtos y una referencia singular elegida no identifican un mecanismo causal.
ALiBi	Un cero numérico reportado en 5000 no identifica únicamente la ley de potencia Colapso. Inspecciona mascaras, contenido y precisión antes de interpretarlo.
Mittag–Leffler	La razón mediana reportada .23 utiliza un denominador diferente de la 1.30 Comparación de poder-familia. Cada razón de normalización exacta se enfocaría en uno; ninguno establece un proceso de tiempo de espera.
Tanh trayectoria	La histórica cuatro puntos de control $R^2 = .15$ y no monótona Las pendientes no respaldan la trayectoria monótona probada, o identifican un eje de tiempo oculto.
Pruning coordenado	Los costos descriptivos no validan una frontera de banda rotatoria; Las máscaras inspeccionadas incluyen coordenadas no rotatorias. La ausencia de un paso no prueba la ausencia de todos los hubs causales.

Topic	Observación e implicación
Analogía de polímero	$3g/(1+2g)$ es una sustitución no probada, no una medida Exponente de propagación. No se establece una interpretación dinámica del polímero.
Controladores adaptativos	Las compuertas de toda la secuencia fallan en la causalidad del prefijo. Su guardado Los puntajes forzados por el profesor no certifican la perplexidad autoregresiva causal.
Möbius cluster	Todas las 23 configuraciones quedan fuera del radio de prueba Eje horizontal solo; el cero observado está estructuralmente forzado.

19 Coordenadas Efectivas y Propuestas de Control Dinámico

Intervenciones existentes y conexión con la Parte III

La campaña del 5 de noviembre contiene 8.064 intervenciones sobre 32 prompts test compartidos por Pythia-70M, Pythia-410M y GPT-2, con 1.152 controles de intensidad cero. Al reducir un 10 % la salida de atención, el predictor por gradiente reduce el MAE frente a efecto cero en 67.7 %, 95.9 % y 94.3 %, respectivamente. Son reducciones del error predictivo, no mejoras de acierto o capacidad. La ablación completa es más difícil: Pythia-70M no mostró una ventaja clara frente a efecto cero con el análisis de incertidumbre emparejado.

Un seguimiento seleccionado por discrepancias incluye 18 casos, cuatro pasos y dos variantes de precisión: 144 registros de diferencias finitas. NeoX eager en Transformers 5.13.0 calcula softmax en FP32 aun con pesos FP64; la referencia que mantiene softmax FP64 recupera convergencia en los casos investigados. Es un diagnóstico numérico exploratorio, no una réplica independiente de todos los casos. Se conservan backend original y registros de campaña.

La Parte III distingue una identidad afín de lectura con normalización final fijada de las derivadas sobre la secuencia completa. La escritura directa de un componente no determina su efecto causal posterior. Se usa esta revisión del dataset TAF:

`f240c0638bbfe8253a7e8098a4bc651b486ebabb`

19.1 Tres superficies de control propuestas

La heterogeneidad medida de la cabeza motiva experimentos a más de un nivel de agregación. No proporciona a un modelo aprendido un exponente global única ni establece un espacio euclidiano de dimensión $1/\gamma$ Un solo contador de umbral de peso ni define una métrica ni excluye modelos lorentzianos o hiperbólicos. La componente métrica por cabeza introducida previamente g_{vv} no se definió y no se utiliza aquí.

Tres propuestas permanecen conceptualmente distintas de las siguientes hooks históricas:

1. Una regla por token podría comparar una fila de atención medida con una familia objetivo usando divergencia KL. Calcular esa estadística cuesta $O(N)$ por cabeza/consulta una vez que la fila está disponible; extraer filas y cambiar las frecuencias rotatorias puede incurrir en costos adicionales y requerir consistencia en el caché.

2. Una regla por cabeza podría calibrar una perturbación diferente contra cada perfil real de la cabeza, luego probar el efecto resultante en la tarea. Una pendiente objetivo no por sí sola determina una intervención efectiva.
3. Una regla de retroalimentación por tarea podría adaptar un parámetro usando solo estadísticas pasadas. Su estabilidad, latencia, regla de parada y beneficio en conjunto necesitan medirse. La propuesta en tiempo de entrenamiento de la Sección 12.3 no establece estas propiedades en inferencia.

No se valida aquí ningún controlador impulsado por KL o orientado a pendiente. La campaña de predicción de intervención de noviembre mencionada anteriormente es una evidencia separada con sus propios protocolos y controles.

19.2 Puntos de anclaje históricos: implementación y validez causal

Los experimentos guardados utilizaron Pythia-1B, solicitaron fp16/eager attention y 50 fragmentos de prueba de WikiText-2 de longitud 1024, muestreados con semilla 42. El script reinicia esa semilla para cada política y reporta la exponencial de la pérdida media por fragmento. El JSON contiene medias y desviaciones estándar agregadas, no pérdidas por fragmento emparejadas o revisiones de entorno de ejecución inmutables. El siguiente análisis de auditoría ejecuta definiciones de hooks extraídas sobre tensores sintéticos; no vuelve a ejecutar el modelo histórico ni certifica su entorno de biblioteca exacto.

El cooler de salida de atención no es una regla causal por cabeza. El hook se conecta a la salida del módulo de atención. En la implementación NeoX inspeccionada que se encuentra después de la proyección de salida; redimensionar sus coordenadas ocultas en bloques iguales no recupera las salidas originales de la cabeza. Para cada tal bloque, el código promedia la varianza de coordenada sobre *secuencia completa* normaliza a través de bloques y establece

$$c_b = \max\{0, 1 - v_b/\overline{v}\}, \qquad a_b = (1 + \epsilon c_b)^{-1}.$$

Se enfría por debajo de la dispersión promedio, no la parte superior del 25 % de cabezas. Ni v_b ni c_b se mide curvatura de Fisher o entropía de atención.

Porque la puerta utiliza posiciones posteriores, cambiar un sufijo modifica la salida gateada en un prefijo anterior idéntico. Ejecutar la definición original con dos bloques de ancho dos, $\epsilon = .5$, y primer vector $(1, -1, 1, -1)$ da como primer salida $(.680203, -.680203, 1, -1)$ cuando el segundo vector es $(1, -1, 8, -8)$, pero $(1, -1, .680203, -.680203)$ cuando es $(8, -8, 1, -1)$. La diferencia máxima en el prefijo es .319797; la clave de control de fuerza nula devuelve exactamente su entrada. Los puntajes forzados por el maestro con esta puerta estiman, por tanto, una perplexidad causal autoregresiva ordinaria. Esta falla estructural es independiente de si una puntuación guardada mejoró o empeoró. No cuantifica el tamaño de la fuga en esa ejecución.

El gancho rotatorio también difiere de su intervención publicitada. Multiplica tanto las coordenadas coseno como seno desde el índice $\lfloor .7D \rfloor$ onward. At $D = 64$ cambia las coordenadas 44–63, solo parte de una mitad duplicada de frecuencia, en lugar de planos rotatorios completos. Con $\epsilon = .5$ Una coordenada cambiada $\cos^2 + \sin^2 = 2.25$ En lugar de uno. Esto representa una perturbación anisótropa de amplitud, no una rotación ortogonal, un cambio de base o una intervención establecida de "banda muerta".

Cuadro 42: Puntajes de múltiples controladores guardados. Las filas que contienen el enfriador adaptativo tienen el defecto de causalidad prefijo mencionado anteriormente; todas las filas permanecen como agregados históricos.

Política	rotatoria ϵ	más fría ϵ	puntaje registrado
Baseline	0	0	15.714571
Rotary solo	.5	0	42.763976
Cooler solo	0	.5	15.779193
Combined	.5	.5	39.498063
Rotary strong	1	0	520.755205
Cooler strong	0	1	15.912687
Combinado fuerte	1	1	385.132282

Todos los siete agregados de múltiples controladores. En comparación con el exceso exclusivo de rotación sobre la base line, el .5 combinación recupera 3.265913 unidades de puntuación, o 12.073880 % En relación con la *sum* de los dos excesos de single-hook it recupera 3.330534 units, or 12.283435 % Estos denominadores no deben intercambiarse. La aditividad de las perplexidades es a su vez una referencia elegida; no se implica por la aditividad de las pérdidas. El primero 12.5 % La predicción de Fisher utilizó una incorrecta top-25 % Descripción y un ajuste de ponderación no medida. Se retiran, así como las afirmaciones de acuerdo cuantitativo, seguridad o viabilidad del controlador.

Reglas de thermostat-v2 y la comparación temprano/tarde. El mismo problema del bloque post-proyección se aplica a thermostat-v2. Su entropía y sus proxies gamma se agrupan a lo largo de toda la secuencia; ejecutar ambas reglas originales en tensores con prefijos iguales/sufijos diferentes también cambia la salida temprana. Las compuertas de identidad y solo capa pasan esa prueba sintética de prefijo. La regla solo de capa utiliza

Cuadro 43: Todas las reglas de thermostat-v2, en relación con la base de comparación guardada 15.714571. La compuerta reportada es la primera capa en el último bloque evaluado.

Rule	Diferencia de puntaje	puerta reportada	Prueba de prefijo
Identidad	.000000	1.00	pasada
Índice de entropía de cuartil superior	2.372792	.875	fallido
Top-quartile gamma proxy	1.658443	.875	failed
Capas tardías	5.110877	1.00	pasaron
Capas iniciales	68.918053	.50	superaron

$\ell/(L-1) > 1/2$ or $< 1/2$ con base cero ℓ Para 16 capas, estos seleccionan ocho capas cada una. Su ratio de puntuación excedente es 13.484584 Este registro refleja la asimetría de intervención, no una prueba de que las capas iniciales realizan trabajo posicional irremplazable o se encuentren en un punto de nacimiento del RG. La información de Fisher de soporte finito no diverge en uno. Las puertas de proxy no miden el exponente de distancia ajustado.

Los anteriores corridas similares a v1 y expertas compartían una puerta cercana a sigmoid(5) = .993307, en lugar de una identidad exacta; su puntaje registrado fue 15.723432, mientras que una puerta cercana a cero dio 4013.370283. Estos son corridas

separadas, no controles adicionales de v2. El archivo single-cooler contiene todos los cinco ajustes de fuerza:

$\epsilon = 0$	.1	.3	.5	1
15.714571	15.720480	15.743350	15.779193	15.912687

Repetir la regresión OLS logarítmica sobre los cuatro puntajes de exceso positivos da el exponente 1.522543; el aumento relativo de .1 es .037599 %. Ejecutar el GammaController extraído de este script sobre el mismo par de prefijos sintéticos reproduce la infracción .319797. Son puntajes registrados recuperables, pero no establecen un margen causal de seguridad ni un paso de Fisher de régimen lineal a cuadrático. No se conserva la incertidumbre de ejecuciones independientes repetidas.

La referencia corregida en la Sección 19.3 implementa una regla causal en el límite de proyección deseado. Sus comprobaciones sintéticas reparan defectos en la implementación, mientras que los antiguos puntajes agregados permanecen como observaciones históricas no demostradas. La eficacia de los modelos entrenados requiere evaluación separada y emparejada con procedencia inmutable.

19.3 Referencia de implementación corregida y su alcance verificado

El código de referencia nuevo se proporciona en `corrected_reference.py`. Este corrige los defectos de implementación mencionados anteriormente sin sobrescribir el código histórico o reasignar sus puntajes medidos. La nueva regla es una política de dispersión por consulta; no estima la entropía, la curvatura de Fisher o el exponente de la ley de potencia.

Sea $h_{b,t,a,r}$ la salida actual de las cabezas antes de la proyección de salida, con lote b, consulta t, cabeza a y coordenada r. Escalen todas las coordenadas y cabezas en cada (b, t) por su valor absoluto común máximo (utilizando un piso positivo de punto flotante en cero), obteniendo u. Esta escala común previene el desbordamiento y no mezcla las consultas. La referencia define

$$\begin{aligned} v_{b,t,a} &= \operatorname{Var}_r(u_{b,t,a,r}), \\ c_{b,t,a} &= \left[1 - \frac{v_{b,t,a}}{\operatorname{máx}(\operatorname{mean}_a v_{b,t,a}, \tau)}\right]_0^1, \\ h'_{b,t,a,r} &= \frac{h_{b,t,a,r}}{1 + \epsilon c_{b,t,a}}. \end{aligned}$$

donde la varianza utiliza el denominador de la población, $\tau > 0$ es un piso numérico y $\epsilon \geq 0$. En $\epsilon = 0$ la entrada se devuelve exactamente. La regla utiliza solo las salidas de la cabeza de la consulta actual. Su factor radica en $[1/(1+\epsilon), 1]$; esta afirmación de amplitud no es una garantía de pérdida.

Proposición 19.1 (Causalidad del referente corregido)**.** *Asuma que cada salida de la cabeza anterior en la consulta t depende solo de las posiciones de entrada en o antes de t. La puerta mostrada seguida de una proyección de salida por consulta preserva esa causalidad prefix. Porque no utiliza ninguna agregación de secuencia ni estado mutable, dividir el mismo tensor de salida de cabeza en fragmentos de consulta no cambia el resultado.*

Demostración. La varianza, la media de la cabeza, el recorte y la escala en (b, t) dependen solo de $h_{b,t,:,:}$. Componer estas funciones y una proyección lineal por consulta con el mapa

ascendente asumido no puede introducir una dependencia de un token posterior. Los límites de los fragmentos no aparecen en sus definiciones. □

Para el diseño NeoX inspeccionado, el adaptador es un *pre-hook en* `attention.dense`, que opera sobre su entrada de cabeza concatenada. Verifica el número declarado de cabezas y su dimensión. Una arquitectura general requiere su propia convención verificada de concatenación y proyección; adjuntar este adaptador después de una proyección de mezcla reintroduciría el error semántico. El módulo no se adjunta automáticamente a otras arquitecturas.

El referente rotatorio construye ángulos a partir de posiciones explícitas y un vector de frecuencias únicas en plano. Duplica ángulos según sea split-half o pairing intercalado y calcula sus seno y coseno. Así, cada par tiene $c_i^2 + s_i^2 = 1$ y una rotación que preserva la norma 2×2; las coordenadas fuera del subespacio rotatorio declarado se copian sin cambios. Estas identidades son válidas en aritmética exacta; las comprobaciones de punto flotante usan tolerancias dependientes del tipo de datos, y una salida rotada no representable genera un error. Una escala positiva por plano cambia la frecuencia angular, no la amplitud. No se afirma que ninguna escala detecte bandas muertas, optimice la longitud del contexto o mejore el rendimiento del modelo.

Medición, ablación y alineación del objetivo. El mismo módulo incluye tres correcciones adicionales. Su rutina de perfil de distancia utiliza un conjunto explícito de consultas elegible en *every* distancia ajustada en un solo tensor de atención. En la atención causal uniforme, su nulo es la constante media $|Q|^{-1} \sum_{q\in Q} 1/(q+1)$ Esto se verifica mediante consultas de longitudes distintas. Mantiene los ejes de lote/cabeza, rechaza la atención futura y rechaza la elegibilidad específica de distancia en lugar de modificar silenciosamente el conjunto. La rutina de ablación de fused-NeoX anula completamente los pares rotatorios declarados de Q/K, dejando los V y coordenadas no rotadas sin cambios. No selecciona esos pares mediante una clasificación semántica o de banda muerta.

El procedimiento de pérdida recibe IDs brutas de entrada y alinea logits en t con el objetivo en $t+1$ exactamente una vez. Una máscara de par de válidos excluye el relleno en cualquiera de las posiciones. Devuelve una suma de pérdidas por token y un conteo de objetivo válido; la perplexidad del corpus es $\exp(\sum \mathrm{NLL}/\sum n)$, no una media ponderada de los medios de los fragmentos cuando los conteos difieren. Las pérdidas no finitas y los conjuntos de objetivo vacíos generan un error en lugar de ser omitidos silenciosamente. Los logits se centran por fila antes de evaluar la pérdida, evitando la cancelación causada por un desplazamiento común grande; la suma final de la pérdida también se verifica para desbordamiento. El llamador debe suministrar IDs brutas: las formas coincidentes no pueden detectar una secuencia desplazada externamente. Estos procedimientos reparados no reconstruyen conjuntos históricos de consultas faltantes, pérdidas por token o corridas de entrenamiento.

El paquete registra 145 implementaciones y verificaciones aritméticas, incluyendo sustitución de sufijos, truncamiento de prefijos, independencia de lote, equivalencia de tokens/piezas, un control negativo de mezcla-proyección, un cálculo sintético de atención causal, ambos pares rotatorios, rotaciones inversas y colas parciales de RoPE. Otros controles verifican objetivos de desplazamiento único, ponderación desigual de piezas, nulos fijos de consulta y ablationes completas de planos Q/K sin cambios en V. Sumas y cuantiles de alta precisión independientes verifican las rutinas numéricas, incluyendo $g = 1$ y su vecindad. Otros controles prueban grandes desplazamientos logit comunes, desbordamiento de suma de pérdida y rotaciones no representables. Estos son tests de tensores y

fórmulas de CPU; no se cargó ningún modelo preentrenado y no se afirma una nueva PPL de modelo entrenado.

20 Coordenadas Racionales y los Límites de la Analogía RG

El candidato racional permite la diferenciación exacta y la iteración. Estas operaciones se refieren a una fórmula escalar. No se establece ninguna mapa de granulación gruesa sobre los estados de red, Hamiltoniano, exponentes críticos medidos o clase de universalidad mediante ellas. Las identificaciones anteriores con el flujo Wilson-Fisher, dinámica Ising y una dimensión anómala física son retiradas. Un modelo separado de percolación para transformadores entrenados en un lenguaje formal[28] no proporciona esas identificaciones faltantes para este manuscrito.

20.1 Definiciones candidatas de coordenadas y exponentes legados

Para base positiva y contexto, escribir

$$g = \gamma_P = \frac{2\theta - T\sqrt{2}}{2\theta + T\sqrt{2}} = \frac{1-u}{1+u}, \qquad u = \frac{T}{\theta\sqrt{2}} > 0, \qquad -1 < g < 1. \tag{63}$$

Positive g también requiere $T < \theta\sqrt{2}$. Este rango pertenece al candidato, no a todas las pendientes medidas de atención: varios perfiles guardados exceden uno. Llamar a g una dimensión anómala requeriría una observable y ley de escala definidas por separado. No es una consecuencia de RoPE.

El anterior listado de exponentes críticos se conserva como un conjunto de definiciones:

$$\beta_{\text{leg}} := g - 1, \qquad \nu_{\text{leg}} := (1-g)^{-1}, \qquad \eta_{\text{leg}} := g - 1, \quad g \neq 1. \tag{64}$$

Estos símbolos no son medidas independientes del parámetro de orden, la longitud de correlación o los exponentes anómalos. El subíndice distingue a β_{leg} del campo vectorial de ecuaciones diferenciales que se presenta a continuación. Para ilustrar, las entradas redondeadas $g = .930, .674, .997$ dan respectivamente $(\beta_{\text{leg}}, \nu_{\text{leg}}, \eta_{\text{leg}}) =$, $(-.070, 14.286, -.070)$, $(-.326, 3.067, -.326)$ y $(-.003, 333.333, -.003)$. La divergencia de este recíproco definido en uno no implica la divergencia de la información Fisher de la familia finita.

20.2 Traectoria escalar exacta y estabilidad

En condiciones fijas θ, se define $\ell = \log u$. Derivación directa da

$$\frac{dg}{d\ell} = \beta_{\text{flow}}(g) = \frac{g^2 - 1}{2}. \tag{65}$$

Con $x = (1-g)/2$, esta es la ecuación logística

$$\frac{dx}{d\ell} = x(1-x), \qquad 0 < x < 1. \tag{66}$$

Es un cambio de coordenadas, no una derivación de dinámicas de entrenamiento o capa. Una antiderivativa que satisfaga $\beta_{\text{flow}} = -V'$ es

$$V(g) = \frac{g}{2} - \frac{g^3}{6} + C.$$

La antigua expresión $g - g^3/3$ tiene un factor dos incorrecto para este mismo ℓ. Este cubo no es ni un potencial doble pozo cuártico ni un potencial $\mathbb{Z}_2$ simétrico. La paridad del campo vectorial no hace que el flujo sea invariante bajo $g \mapsto -g$ con orientación temporal inalterada.

Para $g(\ell_0) = g_0 \in (-1, 1)$ la solución es

$$g(\ell) = \tanh\left(\operatorname{arctanh} g_0 - \frac{\ell - \ell_0}{2}\right). \tag{67}$$

La derivada del campo vectorial es $\beta'_{\text{flow}}(g) = g$. Consecuentemente, $+1$ es repelente y -1 atractivo para incrementos de ℓ. Ningún extremo se alcanza desde el intervalo abierto en tiempo finito ℓ; inicializar exactamente en cualquiera de los extremos da una solución constante. Los anteriores puntos de "nacimiento" en tiempo finito y las afirmaciones contradictorias sobre estabilidad son por tanto incorrectas. Aquí $\ell = \log(T/(\sqrt{2}\,\theta))$ con θ fijo; no es el tiempo de optimización transcurrido ni el número de capas.

Lo que realmente verifica el cálculo del gradiente de fase. Para un plano rotatorio k en un subespacio de dimensión d_{rot}, $\varphi_k(T, \theta) = T\theta^{-2k/d_{\text{rot}}}$ es suave para $T, \theta > 0$. Al definir $A = d\varphi_k$ se obtiene $\partial_T A_\theta - \partial_\theta A_T = 0$ porque las derivadas mixtas conmutan. La forma uno es globalmente exacta en este dominio, por lo que su integral alrededor de un bucle cerrado se anula. No se definió ningún estado cuántico o conexión de Berry. Este cálculo no descarta la memoria o la dependencia de camino en los pesos de las redes entrenadas, y no identifica un haz de Berry de modelos.

Comparaciones de trayectorias guardadas. La exploración histórica del eje de tamaño reporta un mejoramiento $R^2 \simeq .12$ para ocho tamaños de Pythia bajo sus propuestas ejes de log-tamaño/profundidad. La secuencia independiente de cuatro puntos de control tiene pendientes globales $.757234, .851255, .875856, .717686$ at steps $1, 1000, 10000, 143000$ (Table 30 Es no monótona, mientras que la trayectoria anterior es monótona. $(-1, 1)$ Lo anteriormente reportado $R^2 = .15$ pertenece a un modelo ajustado de tanh con escala libre, no una confirmación directa de la Ec. (67) Estos resúmenes ajustados son históricos, no son nuevas repeticiones de entrenamiento. Su mala ajuste no identifica un tiempo RG dependiente de los datos oculto. Esa identificación requeriría una operación especificada y pruebas independientes.

20.3 Mm̈apa de Mobiüs: dominio correcto y punto fijo algebraico

Sea $M(z) = (2 - z)/(2 + z)$ para complejo $z \neq -2$. Satisface

$$|M(z)| < 1 \quad \Longleftrightarrow \quad \Re z > 0,$$

pues $|2 + z|^2 - |2 - z|^2 = 8\Re z$. Así, lleva el *semiplano derecho* al disco unitario. No mapea el semiplano superior al disco: $z = -1 + i$ tiene $M(z) = 1 - 2i$ y $|M(z)|^2 = 5$. La matriz $\left(\begin{smallmatrix} -1 & 2 \\ 1 & 2 \end{smallmatrix}\right)$ El determinante es -4; no existe una múltiplo escalar real con determinante $+1$. No es un elemento de $\mathrm{PSL}(2, \mathbb{R})$ actuando en el semiplano superior. De hecho $\Im M(z) = -4\Im z/|z + 2|^2$. La anterior nota correctiva también fue demasiado lejos al negar toda interpretación de semiplano a disco; la afirmación del semiplano derecho es válida.

Resolver $M(z) = z$ da lugar a las dos raíces $(-3 \pm \sqrt{17})/2$ y por tanto

$$z^* = \frac{\sqrt{17} - 3}{2} \simeq .561553. \tag{68}$$

En esta raíz positiva,

$$M'(z^*) = -\frac{4}{(2+z^*)^2} \simeq -.609612.$$

El multiplicador negativo da convergencia alternada local. La convergencia global desde $z_0 > 0$ se sigue al considerar $M^2(z) = (3z+2)/(z+6)$: mapea $[0,\infty)$ en $[1/3, 3)$ y tiene derivada $16/(z+6)^2 \leq 4/9$ allí. Su único punto fijo positivo es z^*, por lo tanto, ambas subsecuencias convergen. Esto es iteración algebraica; cuando $z_0 > 2$, incluso su primera iteración es negativa y no puede ser una relación contexto/base positiva. No es una secuencia de intervenciones físicas NTK.

Para el definido recíproco $\chi(g) = 1/|1-g|$,

$$\chi(z^*) = \frac{5+\sqrt{17}}{4} \simeq 2.280776. \tag{69}$$

Es la *mayor* de las dos raíces positivas de $2y^2 - 5y + 1 = 0$, no su única raíz positiva. Es un número algebraico, no una constante de atención establecida empíricamente.

Por qué la prueba de agrupamiento de 23 filas no es informativa. La tabla conservada calcula $z = T\sqrt{2}/\theta$ y la distancia euclídea a $(\log z^*, z^*)$ en el plano $(\log z, \widehat{\gamma})$. Al reproducir las 23 filas, 0/23 quedan dentro del radio .5, pero incluso la menor distancia horizontal es .685814870. Por tanto, todas las filas quedan fuera del círculo con independencia de su $\widehat{\gamma}$ medido. La antigua referencia de área rectangular 6.5 % no está condicionada a estas configuraciones elegidas; no convierte ese cero impuesto por la geometría en evidencia contra un atractor del entrenamiento. No se midió ninguna trayectoria del modelo bajo transformaciones repetidas. Se archivan las 23 filas completas, sin tratarlas como una réplica independiente ni atribuir únicamente al entrenamiento la dispersión dentro de una familia.

20.4 Propuesta del umbral NIAH: no validada por el panel guardado

La propuesta histórica combinó la coordenada inversa con el doble de la distancia media de la aguja:

$$L^c_{\text{NIAH}} = 2d_{\text{inv}}(\theta,\gamma) = \frac{2\theta(1-\gamma)\sqrt{2}}{1+\gamma}. \tag{70}$$

Sustituyendo al candidato se obtiene $L^c = 2T$ por construcción. Para el histórico .748 exponente, $2d_{\text{inv}}$ es aproximadamente 4078, pero v4 muestrea cinco longitudes y no localiza umbral agudo allí. Sus agujas se encuentran en una rejilla fija de profundidad; no sobrevive ninguna curva por prueba en el JSON. La dimensión geométrica de la atención y un soporte percolante no fueron medidos. La caída observada certifica por tanto ni la precisión del umbral 0.4 % ni un techo universal de recuperación. Esta propuesta es independiente del experimento de percolación en lenguaje formal citado anteriormente.

20.5 La puntuación de la razón de norma no tiene una interpretación establecida de empaquetamiento.

El auditorio M_{hist} es un resumen de norma-ratio coordenado con un orden de promedio especificado (Sección 14.2 La igualdad con uno denota paridad normativa bajo esa convención, no una transición razonamiento/memoria. No se ha definido ningún límite, volumen

interior o capacidad semántica. Un experimento controlado de capacidad requeriría un índice plano válido, presupuestos de parámetros emparejados y resultados de validación cruzada. NR-12 se refiere a la descomposición suave y no proporciona tal prueba.

20.6 Laplaciano fraccionario: las suposiciones del operador faltante

En una dimensión el laplaciano fraccionario de orden $\alpha \in (0, 2)$ usa un kernel de salto proporcional a $|x - y|^{-1-\alpha}$ junto con la resta de valores de función y un dominio prescrito[22] Una formal match de tail powers sería

$$s_{\text{formal}} = \frac{\gamma - 1}{2}, \qquad \alpha_{\text{formal}} = 2s_{\text{formal}} = \gamma - 1. \tag{71}$$

El rango admisible de este emparejamiento es $1 < \gamma < 3$. Para $\gamma < 1$ el índice es negativo y no puede describir una difusión α-estable; en uno el índice cero está fuera del rango indicado. El caso de Cauchy es $\alpha = 1$ (formalmente $\gamma = 2$), no $\gamma = 1$. El punto final browniano $\alpha = 2$ es un operador local limitante con núcleo de calor gaussiano, no un punto final ordinario $|x - y|^{-3}$ de densidad de salto.

Una matriz de atención fila-estocástica P preserva constantes, $P\mathbf{1} = \mathbf{1}$; un Laplaciano las anula. El generador de Markov es $P - I$; con la convención de operador positivo $(-\Delta)^{\alpha/2}$, un generador de difusión tendría el signo opuesto, $-(-\Delta)^{\alpha/2}$. Identificar tal límite de escala aún requiere suposiciones de homogeneidad espacial, cola, normalización, frontera y convergencia. La atención causal finita dependiente del contenido no proporciona estas automáticamente. La atención neural fraccionaria específica de Qu et al. [36] construye su propio mecanismo de difusión; no establece la conversa para el softmax ordinario.

También falla el intento de cierre mediante una función de Green. Al sustituir $s = (g - 1)/2$ y $d_{\text{eff}} = 1/g$ en el exponente formal $2s - d_{\text{eff}}$, se obtiene $g - 1 - 1/g$. La igualdad con $-g$ tiene raíces $g = 1$ y $g = -1/2$. Restringir a $g > 0$ deja la primera, pero allí $s = 0$, fuera de la condición habitual del núcleo de Riesz $0 < 2s < d$. Por tanto, ni siquiera la raíz algebraica positiva identifica una función de Green física. Un exponente de conteo de umbrales no es una dimensión espacial medida.

20.7 Analogía conforme: una etiqueta de exponente no es una CFT

La reescritura

$$d^{-g} = d^{-2h_{\text{formal}}}, \qquad h_{\text{formal}} = g/2 \tag{72}$$

Define una etiqueta. No construye un operador primario, un estado, simetría conforme, una expansión de producto de operadores o bloques de Virasoro. En dos dimensiones, el peso holomorfo y la dimensión de escala completa también requieren una convención; un solo exponente de distancia real no determina ambos pesos. RoPE proporciona frecuencias particulares, no invariancia exacta de escala de la atención aprendida. Perfiles exponenciales y otros no son excluidos por esta etiqueta.

La ley logarítmica del intervalo de Calabrese–Cardy se refiere a la entropía de von Neumann de un estado cuántico reducido bajo condiciones especificadas de CFT.[5]. La presente S_N Es la entropía de Shannon de una distribución de distancia. Coincidir con su coeficiente logarítmico principal. $c/3$ no mide $c = 3$. La fija $/g < 1$, $g = 1$ y fijo$/g > 1$ Regímenes de Shannon también difieren (Sección 7.4 Se retiran la propuesta de carga

central universal, la identificación de exponentes críticos y las predicciones de puntos de orden superior.

20.8 Identidades condicionales y sus dominios

Las siguientes son consecuencias algebraicas de las definiciones, no pruebas independientes sobre diez o más modelos:

1. $(1/g)(g/2) = 1/2$ for $g \neq 0$ La interpretación del umbral utiliza además $g > 0$ Ninguno de estos factores es una dimensión geométrica o conformal medida.
2. $(g-1)/(1-g) = -1$ for $g \neq 1$ No se trata de una ley de hiperscaling de Widom demostrada entre observables definidos de forma independiente.
3. For $-1 < g < 1$, el definido $d_{\text{inv}}\chi = \theta\sqrt{2}/(1+g)$ tends to $\theta/\sqrt{2}$ as $g \to 1^-$ Los factores singulares no están definidos en uno.
4. Componer la base inversa con γ_P Devuelve el objetivo exactamente. Re medir un modelo modificado no necesita devolver ese objetivo. Por lo tanto, la afirmada propiedad empírica de "idempotencia NTK" se retira.

Para referencia cruzada, el punto fijo algebraico es

$$z^\star = \gamma_P(z^\star) = \frac{\sqrt{17}-3}{2}. \tag{73}$$

Esto da $T/\theta = z^\star/\sqrt{2} \simeq .397078$, no .396 exactamente. Multiplicar θ por $a > 0$ a valor fijo de T envía z a z/a, no a $M(z)$. A menos que $a = 1$, cambia al candidato incluso en $z^\star$. La iteración de M y la escala repetida de la base son operaciones diferentes.

A valores fijos θ la afirmación diferencial equivalente es

$$\frac{dg}{d\log z} = \frac{g^2-1}{2}. \tag{74}$$

Su solución interior general tiene $z(1+g)/(1-g) = C$; el candidato particular selecciona

$$z\frac{1+\gamma_P(z)}{1-\gamma_P(z)} = 2, \qquad z > 0. \tag{75}$$

Esta es una identidad coordenada. Sustituir pendientes medidas no la convierte en una invariante bajo una intervención real en una red, y su residuo singular cerca de $g = 1$ no es una métrica calibrada de error físico.

Representación zeta finita. Para entero $N \geq 1$,

$$Z_N(g) = \zeta(g) - \zeta(g, N+1), \qquad g \neq 1,$$

con continuidad analítica donde sea necesario. Los polos aparentes se cancelan en uno, dando lugar a H_N, no a una divergencia de contexto finito. Para un $g > 1$ fijo, el límite de soporte infinito es $\zeta(g)$. Los valores $\pi^2/6$ y $\pi^4/90$ pertenecen a ese límite en $g = 2, 4$, respectivamente, no a una suma finita. Además, los positivos θ, T restringen a γ_P a $(-1, 1)$; el candidato inverso no puede ingeniar esos exponentes enteros.

Otras analogías propuestas. La comparación histórica de 39 curvas de Mittag–Leffler comunicó una razón mediana de .23 (desviación estándar 5.5) frente a $1/\Gamma(1-g)$ y otra de 1.30 (desviación estándar .56) frente a un normalizador de potencia truncada. Cada razón tendería a uno bajo su propia normalización exacta; 1.30 no es el valor previsto de la primera razón. Las sumas residuales citadas para Pythia-2.8B, .0068 frente a .0053, favorecen el ajuste de potencia pura en esa comparación. Estos resúmenes heterogéneos de perfiles no refutan la familia matemática de Mittag–Leffler ni identifican un proceso de tiempos de espera.

Escribir $p(d) = |\psi(d)|^2$ con $\psi \propto d^{-g/2}$ restablece la normalizabilidad. La suma con soporte infinito cambia en uno; no es una banda de movilidad de Anderson sin un Hamiltoniano, desorden y estados propios. De igual manera, insertar $d_F = 1/g$ en la expresión propuesta de Flory da $3g/(1+2g)$ algebraicamente, pero no se ha medido aquí ninguna ley de expansión de polímeros o dependiente del tiempo. Los valores históricos .861, .878, .976 son ejemplos de sustitución, no predicciones dinámicas validadas.

La correspondencia propuesta con el horizonte BTZ también es incorrecta. En la convención usual de BTZ rotatorio no cargado,

$$r_-^2 = \frac{M\ell^2}{2}\left(1 - \sqrt{1 - \frac{J^2}{M^2\ell^2}}\right)$$

para $M > 0$, $\ell > 0$ y $|J| \le M\ell$[1]. No es la fórmula inversa racional en contexto escrita. En particular $J = 0$ da $r_- = 0$, mientras que sustituyendo $g = 0$ da $d_{\text{inv}} = \theta\sqrt{2} > 0$. Ningún cambio especificado de variables o solución cargada corrige la identificación afirmada $M = \theta$, $g = Q/M$.

Finalmente, para $g < 1$ sea $C(g) = -\log(1-g) - g/(1-g)$. A lo largo del flujo escalar,

$$\frac{dC}{d\ell} = \frac{g(1+g)}{2(1-g)}.$$

Es negativo en $(-1, 0)$ y positivo en $(0, 1)$. Esto es un cálculo elemental de monotonía, no una confirmación o violación de un teorema de CFT c, cuyas suposiciones físicas no han sido proporcionadas. En el centro elegido de la puerta alpha-S, establecer su puntaje en uno fuerza igualmente a $ND = e^2 N_0 D_0$ por normalización; no descubre una constante de inducción.

20.9 Hallucinación como exceso del horizonte: una propuesta no calibrada

La hipótesis original relacionaba un error factual con una coordenada inversa candidata. La propuesta sigue disponible para su prueba, pero ni esa coordenada ni el factor $\chi = 1/(1-\gamma)$ es un predictor validado de forma independiente de las alucinaciones. En particular, el último no es la información de Fisher de familia finita de la Sección 7.3.

Conjetura 20.1 (Hipótesis empírica, no un modelo de probabilidad establecido). En un panel de modelo y tarea prespecificado, una puntuación de coordenada inversa podría estar asociada con la frecuencia de errores factuales después de controlar la longitud del contexto, el modelo y el protocolo. Esta asociación y su calibración permanecen sin probar aquí.

Para $\theta > 0$, $L > 0$ y $0 < \gamma < 1$, conserve la puntuación candidata

$$s_{\text{hall}}(L; \theta, \gamma) = \text{máx}\{0, 1 - (d_{\text{inv}}/L)^{1-\gamma}\} \frac{\sqrt{\chi}}{1 + \sqrt{\chi}}, \qquad \chi = \frac{1}{1 - \gamma}. \tag{76}$$

Su nombre anterior $P_{\text{hallucinate}}$ sugería una probabilidad calibrada, la cual no ha sido establecida. En $L = 2d_{\text{inv}}$ el factor geométrico es $1 - 2^{\gamma-1}$: es igual a 0.292893 en $\gamma = 0.5$ y a 0.066967 en $\gamma = 0.9$, corrigiendo el valor anterior 0.65. A medida que $L/d_{\text{inv}} \to \infty$ en un valor fijo de γ, el puntaje tiende a $\sqrt{\chi}/(1 + \sqrt{\chi})$. En $L \leq d_{\text{inv}}$ es cero por definición, no evidencia de fiabilidad factual o inmunidad a la hipnosis. Las ilustraciones numéricas anteriores de Pythia y Yi se conservan solo en el archivo del predecesor separado; no son tasas de error medidas.

El índice relacionado con la ilusión geométrica es también solo un puntaje candidato:

$$\text{GHI}(\theta, \gamma, L) = \text{máx}\{0, 1 - d_{\text{inv}}/L\} \frac{1}{|1 - \gamma|}, \qquad \theta, L > 0, \quad 0 < \gamma < 1. \tag{77}$$

No es una probabilidad, y su extensión mediante $\gamma = 1$ o a una longitud inversa negativa requeriría una nueva definición. Una salida nula no implica un riesgo nulo. Una prueba requiere resultados etiquetados, calibración en datos de validación cruzada y comparación con bases más simples en los mismos ejemplos; TruthfulQA, FEVER o tareas de comparación entre documentos controlados fueron propuestas pero no fueron evaluadas para esta fórmula. Ninguna de estas definiciones establece dos modos de fallo causales independientes.

21 TAF Agente: Registro Histórico de Software y Límites Evidenciales

La aplicación expone fórmulas, campos de configuración y perfiles guardados. Esta sección describe la interfaz archivada del manuscrito, no una certificación de la implementación actual. Un evaluador de fórmulas puede reproducir un cálculo sin validar sus premisas científicas. Las decisiones seguras/inn seguras no respaldadas en interfaces anteriores son superadas por los dominios matemáticos y los límites empíricos en esta edición. Ninguna aplicación fue desplegada o actualizada en esta auditoría.

Entradas declaradas revisión del modelo, tokenizador, contexto, población de consultas precisión, máscara y agregación.
Comprobaciones de dominio distinguir la dimensión de cabeza y la dimensión rotatoria, el contexto de la inversa y el exponente medido; rechazar entradas no compatibles.
Cálculos devolver momentos y cuantiles de la familia y coordenadas candidatas con sus hipótesis.
Evaluación comparar la intervención implementada con una referencia emparejada sobre resultados reservados; informar la incertidumbre y los recursos reales.

Figura 18: Flujo de trabajo corregido de cálculo y evaluación. Un cálculo de configuración no proporciona un veredicto de tarea. La imagen de la interfaz obsoleta se conserva únicamente en el archivo del predecesor.

21.1 Puntos finales registrados y disposición de la fuente

El manuscrito registra las siguientes ubicaciones. La disponibilidad, autenticación, licenciamiento y comportamiento del servicio actual no se probaron en esta revisión local.

Application	https://karlesmarin.github.io/tafagent
Guía de campo (ES/EN)	https://karlesmarin.github.io/transformers-guide/
Repository	https://github.com/karlesmarin/tafagent
Mirror	https://huggingface.co/spaces/karlexmarin/taf-agent
Registry	https://github.com/karlesmarin/tafagent-registry

El diseño archivado incluye `python/taf_browser.py` para cálculos, `js/main.js` para orquestación, `js/i18n.js` para cadenas de lenguaje, una capa HTML/CSS y un script de medición CLI. La instantánea del 5 de noviembre congela archivos relevantes para comparación; no es una reconstrucción completa y recientemente probada de cada liberación histórica. La arquitectura descrita combina Python en el navegador mediante Pyodide y síntesis opcional de WebGPU con recuperación remota de modelos y configuraciones. La ejecución numérica local no implica que identificadores o solicitudes nunca dejen la pestaña: búsquedas en Hub y descargas remotas requieren solicitudes de red. La anterior afirmación general de privacidad, el acceso automático a modelos restringidos, la cobertura universal de arquitectura y la promesa perpetua de costo cero son retiradas. Su verificación requeriría el despliegue exacto, rutas de solicitudes, autenticación y licencias.

Históricamente, la interfaz ofrecía una tarjeta TAF, formularios de comparación, un componente de enrutamiento de modelo de lenguaje pequeño, un constructor de comandos CLI y un visualizador de fase. El CLI mide perfiles cuando es compatible; una fórmula de configuración de navegador no los mide. Las filas almacenadas en el visualizador de fase son un panel con límites de procedencia, no observaciones generadas al abrir la aplicación. Las cadenas en inglés, español, francés y chino se describieron en el registro de interfaz; este auditorio no certifica la completitud de la traducción ni las claves actuales.

21.2 Las cinco recetas históricas y su alcance actual

X-1: comparación de costos. Una comparación de costos requiere precios explícitos, throughput, utilización y un objetivo de calidad probado. La antigua puerta alpha-S no puede decidir si el entrenamiento o el autoalojamiento cumplirán con ese objetivo (Apéndice H). No se hace ninguna recomendación actual sobre precios ni despliegue aquí.

X-2: viabilidad a largo contexto. La antigua composición de γ_P, d_{inv} y $2d_{\text{inv}}$ devuelve una longitud algebraica. No predice la precisión de recuperación: $d_{\text{inv}}(\theta, \gamma_P) = T$ por construcción. Los juicios de recuperación exclusivamente basados en configuración y los multiplicadores óptimos implícitos se retiran. La sección 12 conserva las comparaciones específicas de tarea y los costos a corto contexto guardados.

X-3: presupuesto de entrenamiento. La cantidad de parámetros y el presupuesto de tokens son entradas de recursos. La supuesta cota de 400M y la puntuación alpha-S no establecen circuitos de inducción ni un criterio de go/no-go para el entrenamiento. Un futuro estimador necesita el throughput medido y un objetivo de calidad especificado de forma independiente.

X-5: estimaciones de hardware. El catálogo archivado de GPUs y las salidas de precio/tasa son estimaciones históricas. Para un cache homogéneo, su almacenamiento básico es $2BLNH_{\mathrm{KV}}d_{\mathrm{head}}b$ bytes, con lote B, capas L, posiciones retentivas N, cabezas KV H_{KV} y b bytes por escalar; los layouts reales pueden agregar relleno, escalas, metadatos y sobrecarga de asignación. Una ventana hipotética puede cambiar N, pero D_f no predice la tasa de throughput ni una relación de compresión segura. Ancho de banda, implementación del kernel y calidad medida son entradas separadas. No se certifica ningún ranking de USD/token o tokens/segundo.

X-19: masa de caché y resultados de tareas. Las sumas acumulativas continuas de cuantil y exactas discretas cuantifican la masa condicional sobre la familia ajustada. Ninguna proporciona una penalización universal PPL. La estrategia de clasificación basada únicamente en configuración es retirada; compara las políticas reales con una base no comprimida en la carga de trabajo intencional y registra los conteos de cabezas KV, precisión, memoria y tiempo.

21.3 Catálogo de modos posterior y reglas diagnósticas retiradas

La descripción histórica de las versiones v0.4–v0.7 afirmaba 22 modos, pero la lista mostrada tiene 18 filas nombradas más Ask, Compare y Phase diagram: 21 modos nombrados bajo esa convención. Recetas y modos son conceptos de interfaz superpuestos. No se infiere ni certifica un modo vigésimo segundo no explicado mediante esta auditoría.

X-21: hipótesis de impresión dependiente del tamaño. La fórmula propuesta fue $\gamma_{\mathrm{random}} = \gamma_P + \nu \log_{10}(P/(14\,\mathrm{M}))$, con $\nu = -1/(2\pi)$. Un período de rotación no deriva que la pendiente de regresión. Los resultados reportados $p = .022$ de un panel seleccionado no validan el coeficiente fijo ni identifican un entrenamiento anómalo. La regresión de cinco ejes también utilizó predictores derivados de la respuesta (Sección 5.5). Se requieren perfiles directos con protocolos coincidentes; un error relativo a esta fórmula no puede diagnosticar un fallo de conversión o ajuste fino.

X-22: estadística descriptiva de cálculo/contexto. La antigua $K = \gamma \log(N^2 D)$ utilizaba $D = 20N$ en lugar de los tokens de entrenamiento observados de cada modelo. Su valor depende de las unidades de conteo; una versión explícitamente adimensional requeriría conteos de referencia N_0, D_0. La histórica 51.2 ± 16.8 es una media y dispersión de panel, no una invariante ni umbral calibrado de anomalía. El ejemplo $K = 74.6$ tiene $(74.6 - 51.2)/16.8 = 1.392857$ desviaciones estándar, no un estatus establecido de outlier estadístico. Se encuentra fuera del intervalo citado y redondeado de una dispersión [34, 68]; esas dos afirmaciones son compatibles pero no implican un defecto en el modelo.

X-23: contrastes en el corpus. La antigua receta utilizaba texto menos aleatorio, opuesta a la actual $\Delta\gamma = \widehat{\gamma}_{\mathrm{random}} - \widehat{\gamma}_{\mathrm{text}}$. Ningún signo detecta universalmente los encabezados de inducción. El subgrupo histórico significa que $-.19$ y $+.03$ no pueden transferirse entre convenciones ni usarse como sustituto de las mediciones de circuito y las pruebas de ICL en conjunto de prueba.

Cuadro 44: Nombres históricos de modo, función intencional y limitaciones. Este es un catálogo, no una característica o certificación de rendimiento actual.

Mode	Función intencional y cualificación necesaria
Perfil	Recopilar metadatos, fórmulas y perfiles registrados; diferenciar campos medidos y predichos.
Desenmascara	Inspeccionar ajustes de posición/escalado; los límites de configuración no miden la calidad del contexto práctico.
NIAH para la razón	estimador de recuperación/reasonamiento histórico; las tasas de pasos solo arquitectónicos no están validadas.
LongScore	Presenta comparaciones de contexto largo almacenado; retiene versiones de referencia y protocolos en lugar de certificar clasificaciones actuales.
Quant	Estimador propuesto del efecto de cuantización; los campos arquitectónicos solos no establecen una ley de potencia o cambio de PPL.
Inspeccionar configuración	Mostrar campos proporcionados; los campos faltantes o no soportados requieren manejo explícito.
Contaminación	Resuma hipótesis de exposición; las fechas de liberación y los priors no detectan contaminación real durante el entrenamiento.
Drift	Compara metadatos de evaluación; una brecha en puntaje por sí sola no identifica un error o cuantifica el ruido numérico.
Arena CI	Ajustar un modelo de comparación pareada y la incertidumbre cuando los votos brutos y las suposiciones de muestreo lo permitan.
Saturación	Comparar rangos de puntajes almacenados; la saturación aparente y otros benchmarks alternativos requieren evidencia fechada.
Soluciones	Index documentan problemas de evaluación y herramientas; un remedio vinculado no es universalmente aplicable.
Plantilla de chat	Identifica convenciones de formato y construye comandos; una plantilla omitida no tiene un efecto universal de precisión factor dos.
Diagnose CLI	Genera un comando de medición; debe verificarse el soporte del modelo actual y las dependencias de ejecución.
"traduccion": "JSON CoT"	Orden de los campos de salida; el orden del esquema en sí no demuestra ni refuta la calidad del razonamiento.
PEFT Lint	Inspecciona scripts en busca de patrones conocidos; las verificaciones estáticas no pueden establecer una ejecución de entrenamiento exitosa.
Cache Diff	Estima los efectos específicos del caché de proveedor utilizando reglas explícitas actuales y hits medidos.
Spec-Decode	Compara tokenizers; el desajuste es una preocupación de compatibilidad, no una prueba de que cada método sea más lento que la base.

Mode	Función intencional y cualificación necesaria
Token Tax	Contar tokens con versiones de tokenizers nombradas; las diferencias en el conteo son independientes de los precios del servicio y la precisión de la tarea.

Capas de verificación. La interfaz antigua describió doce comprobaciones algebraicas en tiempo de ejecución y 37 sellos de teoremas. Un residuo numérico, una simplificación simbólica y un teorema del asistente de prueba tienen contratos diferentes. La ecuación $2\eta^2 + \eta\gamma_\chi + 1 = 0$ necesita su propia definición o restricción γ_χ No es una identidad para cantidades arbitrarias y medidas de forma independiente. Del mismo modo $\beta\chi = -1$ Utiliza las definiciones específicas del legado en el subdominio uno. Un recuento manifiesto no certifica todas las afirmaciones de este manuscrito. El anterior auditorio de 20 lemas de Isabelle mantiene el alcance resumido en el Anexo. A Esta edición además integra los ocho grupos verificados en el Anexo K Su sesión completa de Isabelle/HOL auditó 300 hechos nombrados del proyecto. El conteo incluye hechos intermedios y no es un conteo de afirmaciones empíricas independientes.

21.4 Mapa de fórmulas: cálculo versus interpretación empírica

El mapa enumera las entradas matemáticas reales, incluyendo un exponente ajustado cuando sea necesario. Es un mapa de dependencias de cálculos, no una cadena causal establecida desde la configuración al desempeño de la tarea.

Cuadro 45: Mapa de fórmulas revisado. Las definiciones no proporcionan efectos de intervención.

Object	Inputs	Producción y alcance
γ_P	$\theta, T > 0$	Candidato en $(-1, 1)$; no es una pendiente medida.
d_{inv}	θ, g	Coordenada inversa para $-1 < g < 1$; no corte de información.
$2d_{\text{inv}}$	θ, g	Propuesta de umbral histórica; no garantía de tasa de recuperación.
D_f	g, N, f	Cuantil masivo; especificar soporte continuo o discreto.
Decaimiento suave w	g, d_{inv}, d	Peso multiplicativo, o $\log w$ sesgo en su dominio; PPL requiere evaluación.
Base ratio	$g_{\text{target}}, T, \theta$	Multiplicador inverso-candidato; no optimalidad universal NTK.
Base inversa	g_{target}, T	Base positiva para $-1 < g_{\text{target}} < 1$; adecuación de la tarea no validada.
α_S	R_c, N, D, escalas de referencia	Heurística histórica; interpretación go/no-go retirada.
n_{active}	$d, \theta, d_{\text{rot}}$	Conteo exacto de planos no ciclados con paring especificado; no demostrado el funcionamiento semántico.

Object	Inputs	Producción y alcance
L_{crit}	Protocolo definido de medición	Hipótesis de profundidad histórica; no un límite certificado solo por arquitectura.
C_V	g, N	Respuesta de familia finita, finita en uno; no alarma de olvido calibrada.
χ	$g \neq 1$	Recíproco definido $1/\lvert 1-g\rvert$; distinto de la información de Fisher finita.
S_N	g, N	Distancia de Shannon; la asintótica requiere una especificación fija g regime.
$\Delta\gamma$	Ajustes aleatorios/textuales coincidentes	Contraste descriptivo, no un clasificador de circuitos o arquitectura.
e^{S_N}	g, N	Número efectivo de distancias soportadas, no un techo de razonamiento.
I_{total}	Convención de coordenadas establecida	Información histórica presupuesto-proyecto, no derivado capacidad del canal.
Prueba de pendiente de entropía	Varios soportes y ajustes	Propuesta de diagnóstico de tamaño finito; no es un clasificador garantizado de fase.
R_{patch}	Intervención y línea base	Estadística de recuperación específica del protocolo; las afirmaciones causales requieren parcheo controlado.

21.5 Rol de aplicación y trabajo restante

Una interfaz útil debe exponer la procedencia, los dominios matemáticos, los campos faltantes y la incertidumbre. Las correcciones de la aplicación y la guía del 18 de noviembre se documentan por separado. Esta auditoría del manuscrito no verifica el servicio en vivo ni publica cambios en la aplicación. El acuerdo entre una fórmula mostrada y un ejemplo resuelto es una comprobación aritmética; la validación empírica requiere observaciones independientes.

Lectura y software complementarios. La documentación actual enlaza las fórmulas de la familia finita, el candidato Padé y los perfiles medidos con recorridos distintos. El capítulo de medición de la guía explica cómo conservar el protocolo de consultas y ajuste; el capítulo de TAF Agent explica las rutas del navegador y de la CLI. El contrato de fórmulas especifica el soporte finito, los dominios y la procedencia de los valores mostrados. Las correcciones anteriores de la aplicación y la guía preceden a esta edición del manuscrito del 19 de noviembre. Estos enlaces permiten leer y calcular, pero no constituyen evidencia independiente sobre un modelo entrenado. El Blueprint complementario relaciona enunciados, pruebas, figuras y cuatro laboratorios ilustrativos. El DOI de esta versión es 10.5281/zenodo.24379538; el registro anterior de Zenodo corresponde a la edición histórica.

22 Conclusión

Auditoría de coordenadas y campo. Las observaciones preservadas B3 y de amplificación describen las operaciones implementadas de coordenadas; sus anteriores interpretaciones causales específicas de frecuencia son retiradas. El campo nulo reproduce una pendiente cerca del primer punto de verificación, y las estadísticas espectrales requieren convenciones explícitas de base y agregación. Estas son correcciones necesarias a la interpretación empírica, no resultados nuevos de modelos entrenados. Las intervenciones en plano coincidente y los controles en contexto fijo permanecen necesarios antes de cualquier prescripción mecanística o de despliegue.

El control instrumental precede la interpretación física. El control uniforme-causal E4 demuestra que la medición puede producir una ley de potencia aparente alta R^2 sin atención aprendida. Esto debilita la comparación inicial-panel como evidencia para una ley geométrica. El control fijo de consulta QL no comparte esa confusión de pendiente en su rango de ajuste auditado. Las futuras comparaciones de modelos deben preservar los conjuntos de consultas, soportes y convenciones nulos; los controles sintéticos solos no resuelven cuánto comportamiento real del modelo permanece.

Los perfiles de distancia-atención proporcionan descripciones empíricas útiles cuando se especifica su corpus, soporte y protocolo de medición. La familia de potencias finitas proporciona identidades exactas para la normalización, la entropía, la información de Fisher y una capacidad térmica formal. La entropía máxima selecciona esa familia bajo un momento log-distancia prescrito y una medida de referencia. La geometría rotatoria sola no establece que la atención entrenada resuelva este problema de optimización.

La expresión basada en configuración Padśigue siendo un candidato, no una predicción universal del exponente medido. Su derivación retirada no puede ser restaurada mediante la inversión algebraica de la misma expresión. La comparación inicial mixta con 13 bases registradas tiene error absoluto medio del exponente 0.167766 y error medio relativo absoluto 19.2144 %. La estadística del caso seleccionado se retira porque su pertenencia y cálculo no pudieron ser reconstruidos. En el reemplazo de octubre aparte, 25 de 42 identificadores históricos han recuperado mediciones, y el candidato almacenado tiene correlación de rango -0.445 frente a 19 ajustes disponibles QL. Los cambios en el protocolo impiden interpretar esas discrepancias como una estimación controlada de un efecto arquitectónico o instrumental.

La compresión y los experimentos de contexto siguen siendo observaciones empíricas bajo sus condiciones registradas. Una familia finita de cuantiles no garantiza la precisión de tareas autoregresivas, y una base inversa de candidatos no determina un ajuste óptimo de extensión de contexto. El límite del error de salida de fila fija es útil dentro de sus suposiciones de puntuación fija y valor fijo. La campaña de intervención de noviembre proporciona una prueba separada de predicción de efectos en 32 promt compartidos para tres modelos; las ganancias de predicción no establecen una mejora en la capacidad del modelo.

Los siguientes pasos empíricos son la estimación con protocolo emparejado y incertidumbre de secuencia, la comparación prospectiva contra familias alternativas de perfiles, y la evaluación controlada de las intervenciones propuestas. Las interpretaciones circuitales, geométricas, físicas y de ilusión necesitan evidencia más allá del ajuste del exponente. El apéndice K proporciona ocho grupos verificados que refinan la parte matemática: calibración nula, cotas de familia finita y mezcla, orden cuantil, estabilidad de familia finita,

truncamiento, softmax y una aproximación condicional de coseno a potencia. Estos resultados abarcan sus hipótesis declaradas; no certifican ni la tubería empírica ni la novedad científica.

Esta edición integra la revisión completa de consistencia del documento del 12 de noviembre. El flujo racional y el mapa han corregido los dominios; las identificaciones físicas sin las estructuras requeridas son retiradas. Los controladores históricos adaptativos fallan en la causalidad prefijada, el panel de punto fijo es informativo por construcción, y la comparación V1/V2 utiliza la temperatura de atención en la ranura de contexto. Los argumentos desplazados exactos y los predecesores completos están archivados; el Apéndice J registra su reemplazo; los argumentos obsoletos se conservan fuera del manuscrito actual. Las lagunas en la procedencia histórica y las hipótesis empíricas no probadas permanecen abiertas. La auditoría no es un certificado de que todas las afirmaciones científicas sean libres de errores.

Origen del código y los datos. Los cálculos archivados y sus limitaciones se describen en la Sección 8. El proyecto complementario es `github.com/karlesmarin/tafagent`. Esta edición no introduce ningún nuevo modelo de inferencia ni despliegue de software. Incorpora los desarrollos formales verificados previamente del 19 de noviembre sin modificar sus fuentes de prueba. El paquete de origen incluye la instantánea verificada exacta, sus recibos de compilación y dependencias, un mapa de teoremas a código fuente, y las comprobaciones de integración y compilación del manuscrito.

A Verificaciones Simbólicas y Alcance de la Prueba Formal

Los scripts históricos de SageMath registran exploraciones simbólicas, no un certificado general para cada fórmula mostrada o interpretación. La sesión local aceptada de Isabelle del anterior auditorio contiene 20 lemas con alcance, con verificaciones rápidas y sucias desactivadas y sin dependencias reportadas de oráculos. Su lema de variación continua asume los primeros dos momentos; no mecaniza la integral. Sus limitaciones permanecen sin cambios debido al nuevo apéndice. El anterior auditorio de antecedentes encontró ocho precedentes directos en la propia repositorio y uno precedente PDI después de reparametrización; no se estableció novedad para ninguno de los veinte lemas.

La extensión del grupo de ocho se presenta ahora en el Apéndice K, junto con su sesión completa de 300 hechos en Isabelle/HOL y un mapa a las fuentes de prueba proporcionadas. Ese desarrollo es independiente del anterior auditorio de 20 lemas.

Las verificaciones matemáticas corregidas son:

1. En el continuo $[1, N]$, $N > 1$, el normalizador es $Z_c = (N^{1-g} - 1)/(1 - g)$ con límite $\log N$ en uno. En distancias discretas es $Z_N = \sum_{d=1}^{N} d^{-g}$ con valor H_N en uno.

2. En la familia finita discreta, $U_N = -\partial_g \log Z_N$ y $S_N = \log Z_N + gU_N$ son exactas. La última corresponde a la entropía de Shannon, no a una fórmula de entrelazamiento o de Cardy.

3. $I = \mathrm{Var}(\log d)$ y $C_V = g^2 I$ son finitas para fijo N. El valor $(\log N)^2/12$ en uno pertenece a la familia continua.

4. El f-cuantil continuo es $[(1-f)+fN^{1-g}]^{1/(1-g)}$ con límite N^f en uno. La expresión en el régimen grande-N $Nf^{1/(1-g)}$ requiere fijo $g < 1$; no es uniforme cerca de uno y no corresponde al cuantil discreto exacto.

5. Para $N > 1$, $dU/dg = -I < 0$ y $dS/dg = -gI$, por tanto $dS/dU = g$. La familia degenerada de un punto no define este cociente.

6. $T\sqrt{2}/(1-\gamma_P) = \theta + T/\sqrt{2}$ es una composición exacta con el *lineal* inverso, no una aproximación de primer orden al inverso de Padé.

El auditorio profundo del 12 de noviembre invoca además módulos llamables desde Curiosity, Excovery, Socrates y Triangulum. Registra hashes de módulos, residuales simbólicos, comprobaciones de sumas finitas de alta precisión, controles positivos y negativos, y detección de detección de candidatos. Varios herramientas comparten SymPy; el acuerdo no es cuatro pruebas independientes. Algunos parsers/checkers devuelven resultados inconclusivos, conservados en la evidencia legible por máquina en lugar de promovidos a certificados. La extracción de candidatos no es una validación semántica exhaustiva.

B Auditoría del intento de derivación geométrica

El anterior "demostración completa" mezcló una identidad exacta de softmax con suposiciones estadísticas no probadas y condiciones de frontera inconsistentes. Su texto completo se conserva en la fuente predecesora y en el registro de reemplazo. Los pasos que siguen indican con precisión qué se conserva y qué falla.

Paso 1: puntajes rotatorios y conteos de ciclos. Con el subespacio rotatorio real $d_{\text{rot}} = 2m$, un puntaje rotatorio contiene tanto coeficientes coseno como seno: $\sum_i [a_i \cos(d\omega_i) + b_i \sin(d\omega_i)]$, más cualquier contribución no rotada y el factor de escala de atención. Completar un ciclo a una distancia particular no promedia un senoide a cero. Para un valor fijo de d, promediar entradas aleatorias no promedia la fase en sí. Tampoco las identificaciones de token independientes establecen representaciones proyectadas independientes de media cero después de transformaciones aprendidas. Coeficientes isótropos constantes, una media residual coseno positiva y cancelación de planos ciclados requieren evidencia separada.

Si, como hipótesis adicional, el puntaje esperado fuera $b + a\widetilde{n}_{\text{active}}(d)$ en un rango log-lineal no recortado, su coeficiente de distancia logarítmica sería

$$\kappa = \frac{am}{\log \theta}. \tag{78}$$

El anterior $d_{\text{head}}/(2\log\theta)$ omitió el coeficiente a y igualó la dimensión total de la cabeza al subespacio rotatorio. Incluso el coeficiente condicional correcto no determina un exponente aprendido.

Paso 2: promedio y el teorema condicional válido. En general, $\mathbb{E}[\text{softmax}(s)] \neq \text{softmax}(\mathbb{E}[s])$. La concentración de un denominador solo no sustituye $\mathbb{E}[e^{s_d}]$ por $e^{\mathbb{E}[s_d]}$. Por ejemplo, con dos logits $(X, 0)$ donde $X = 0$ o 2 ocurren con la misma frecuencia, el peso medio del primer softmax es aproximadamente .690399, mientras que el softmax de los logits medios da .731059. El intercambio requiere un control adicional de error, no solo un promedio de muchos prompts.

La afirmación válida es exacta y condicional: si cada logit en el soporte elegido tiene forma $s_d = c - g \log d$, entonces

$$\operatorname{softmax}(s)_d = \frac{e^c d^{-g}}{\sum_{u=1}^{N} e^c u^{-g}} = \frac{d^{-g}}{Z_N(g)}.$$

Esto demuestra la consecuencia de un perfil de puntaje prescrito. No deriva ese perfil o g a partir de una geometría de conteo cíclico. El teorema de MaxEnt por separado parte igualmente de una restricción prescrita de distancia logarítmica.

Paso 3: el candidato racional no está fijado por las condiciones anteriores. Para $M_a(z) = (a-z)/(a+z)$, $M_a(0) = 1$ para cada no nulo a; esta condición no puede seleccionar $a = 2$. El último valor da $M_2(1) = 1/3$, no cero como se afirmó previamente. Su cero ocurre en $z = 2$, o $T = \theta\sqrt{2}$, no en $T = \theta/\sqrt{2}$. En $T = \theta$ da $(2 - \sqrt{2})/(2 + \sqrt{2}) \simeq .171573$, no aproximadamente uno. Se puede obtener $a = 2$ imponiendo la derivada del candidato $M_a'(0) = -1$, pero esto es una condición local establecida; aún no se deriva de proceso de atención medido. Los porcentajes de error anteriores en un subconjunto "no anómalo" ni reparan estas contradicciones ni identifican la isotropía como causa de todos los residuos. Véase los paneles completos en la Sección 5.4.

Tres mapeos distintos de base. En un valor positivo fijo T, resolver $\gamma_P(\theta, T) = g$

$$\theta_P(g, T) = \frac{T\sqrt{2}(1+g)}{2(1-g)}, \quad -1 < g < 1. \tag{79}$$

El candidato lineal tiene en cambio inverso $\theta_L(g, T) = T\sqrt{2}/(1-g)$ para $g < 1$. Evaluar ese inverso diferente en γ_P da

$$\theta_L(\gamma_P(\theta, T), T) = \theta + \frac{T}{\sqrt{2}} = \theta(1 + z/2), \qquad z = T\sqrt{2}/\theta. \tag{80}$$

La expresión de Taylor anterior utilizaba $\theta(1+z)$ y se desviaba en un factor de dos en su incremento. No se necesita aproximación de Taylor para la identidad anterior. Finalmente, la relación impuesta por separado $g = \log\theta/\log b$ tiene inverso $b = \exp(\log\theta/g)$ para $g \neq 0$; no es ninguno de los otros dos inversos. En $\theta = 10000, T = 1000$, el inverso Padé devuelve exactamente 10000 y la composición lineal da 10707.106781. Su diferencia es 7.071068 % de 10000, no un error de inverso sub-0.5 %. En $T = 5000$, es 35.355339 %, no por debajo de 2 %.

C Tabla de Fórmulas

El catálogo que se presenta a continuación muestra los objetos matemáticos utilizados en el resumen corregido. El catálogo completo anterior se conserva en el archivo sucesor separado indexado por el Apéndice J. Su presencia en dicho archivo no otorga validación ni novedad. El estado actual de los veinte grupos originales de hallazgos se detalla en el Apéndice G.

Cuadro 46: Catálogo de fórmulas con soporte e interpretación expresados explícitamente.

Objeto	Expresión	Ámbito
Familia finita	$p_\gamma(d) = d^{-\gamma}/Z_N$, $Z_N = \sum_{d=1}^{N} d^{-\gamma}$	Entero $N \geq 1$, real γ, medida de conteo.
Logaritmo de momento	$U_N = \sum_d p_\gamma(d) \log d$	Equals $-\partial_\gamma \log Z_N$.
Entropy	$S_N = \log Z_N + \gamma U_N$	Entropía de Shannon exacta para suma finita; $0 \leq S_N \leq \log N$.
Métrica de Fisher	$I(\gamma) = \mathrm{Var}_{p_\gamma}(\log d)$	Finito en cada real γ para N fijo; evalúa directamente en cero.
Capacidad térmica formal	$C_V = \gamma^2 I(\gamma)$	Susceptibilidad familiar, no una sensibilidad de tarea identificada.
Longitud intrínseca	$\left\lvert \int_{g_1}^{g_2} \sqrt{I(g)}\, dg \right\rvert$	Distancia a lo largo de esta familia estadística.
Normalizador continuo	$Z_c = (N^{1-\gamma} - 1)/(1-\gamma)$; $Z_c(1) = \log N$	Medida de Lebesgue en $[1, N]$, $N > 1$; no el normalizador discreto.
Varianza en uno	$C_{V,c}(1) = (\log N)^2/12$	Solo continuo. Para distancias discretas use la varianza de $\log d$.
Coordenada de energía libre	$-\log Z_N$	Coordenada adimensional; la división por γ da la convención térmica para $\gamma \neq 0$.
Ventana de masa discreta	$\min\{k : \sum_{d=1}^{k} d^{-\gamma} \geq f Z_N\}$	Distancia entera cuantil; $0 < f < 1$.
Ventana de masa continua	$[(1-f) + fN^{1-\gamma}]^{1/(1-\gamma)}$; N^f en $\gamma = 1$	Cuantil de la familia continua, no una cota de precisión.
Candidato de Padé	$(2\theta - T\sqrt{2})/(2\theta + T\sqrt{2})$	$\theta, T > 0$; rango $(-1, 1)$; identificación empírica no establecida.
Base inversa	$T\sqrt{2}(1+g)/(2(1-g))$	$-1 < g < 1$, $T > 0$; inversión candidata, no un óptimo de tarea.
Contexto inverso	$d_{\mathrm{inv}} = \theta\sqrt{2}(1-g)/(1+g)$	$-1 < g < 1$, $\theta > 0$; no es un techo de recuperación validado.
Base desplazada	$b(T) = \theta + T/\sqrt{2}$	Definición exacta de coordenada; diferente de la base inversa.
Logaritmo de la razón base	$\log[b(D)/b(N)]$	Cambio de coordenadas firmado, no establecida pérdida de información.
Exponente umbral	$d_{\mathrm{eff}} = 1/\gamma$, $\gamma > 0$	Registra un exponente de conteo umbral de peso, no una dimensión geométrica medida.
Soporte efectivo	$\exp S_N$	Entropía del número efectivo dentro de la familia elegida, no un límite del contexto de tarea.
Cota de fila congelada	$\lVert o - o_K \rVert \leq \epsilon\ \mathrm{diam}\{v_j\}$	Puntajes y valores fijos, fila renormalizada conservada, masa eliminada $0 \leq \epsilon < 1$.

Objeto	Expresión	Ámbito
Decaimiento suave	$w(d) = (1 - d/d_{\text{inv}})^{\gamma}$ dentro de $0 < d < d_{\text{inv}}$	Para valores positivos de γ, d_{inv}: utiliza $\log w$ como sesgo del puntaje; una propuesta empírica, no un teorema de geometría.
Corpus contrastante	$\widehat{\gamma}_{\text{random}} - \widehat{\gamma}_{\text{text}}$	Estadística descriptiva del protocolo emparejado; no un identificador de arquitectura causal.

La entropía asintótica es un enunciado separado. Para fijado $0 < \gamma < 1$,

$$S_{\gamma}(N) = \log N - \log(1 - \gamma) - \frac{\gamma}{1 - \gamma} + O(N^{\gamma - 1} \log N).$$

En uno es $\frac{1}{2} \log N + \log \log N + O(1)$; por encima de uno converge a $\log \zeta(\gamma) - \gamma \zeta'(\gamma)/\zeta(\gamma)$ para $\gamma > 1$ fijo. Estos regímenes no pueden fusionarse en la anterior expresión $(1-\gamma) \log N$.

D Zipf y la atención: auditoría de panel guardado

Para dos miembros de la misma familia de potencias finitas con exponente entero $N > 1$ La igualdad de los momentos logarítmicos identifica exponentes iguales en el mismo soporte y medida de referencia (Sección 6.3 Momentos iguales de distribuciones arbitrarias no implican esta conclusión. Los rangos de vocabulario y las distancias de tokens son observables diferentes. El archivo Zipf guardado contiene ocho intentos de entrada: siete estimaciones exitosas y un fallo de acceso restringido para Gemma. Reproducir los siete pares da Pearson $r = -.505314$ Esto es una estadística descriptiva válida por filas, no un cálculo de constante-vector indefinido.

Cuadro 47: Se recuperaron los siete pares válidos de Zipf y atención. Los valores Gamma provienen de los registros históricos del texto E4; los protocolos no son intercambiables.

Modelo	exponente de rango	exponente de atención textual
Pythia-70M	.997365	.747602
Pythia-1B	.997365	.931108
Pythia-1.4B	.997365	.705073
Llama-2-7B	1.232283	.287057
Mistral-7B	1.164898	1.060750
Qwen2.5-7B	1.015546	.996695
Llama-3-8B	1.007281	1.045476

Hay tres filas de Pythia, no cuatro: 2.8B no está presente en este JSON. Sus exponentes de rango son exactamente iguales en el archivo almacenado, por lo que la correlación de Pearson de ese subgrupo no está definida: no vale cero ni aproximadamente $-.5$. Sus exponentes de atención abarcan .705073–.931108. La tokenización compartida reduce la variación independiente en la estadística de rango, pero no demuestra que la tokenización haya sido la causa única de la correlación negativa combinada. No se proporciona una correlación positiva ajustada. Las categorías de tokenizadores “BPE” y “SentencePiece” no son mutuamente excluyentes.

El campo delta del JSON histórico utiliza texto-minus-aleatorio, opuesto a la convención actual de aleatorio-minus-texto. El registro Llama-2 conserva su procedencia no resuelta de adquisición/quantización y no puede establecer un artefacto de precisión. Una comparación controlada requeriría corpus coincidentes, tokenización, soporte, adquisición y incertidumbre, con una hipótesis que realmente vincule los dos observables. Variar la base sola no prueba la igualdad de sus exponentes.

E Tarjeta de Referencia Rápida

Propósito. Separar mediciones, cálculos de familia y pruebas de intervención. La configuración original solo de despliegue se conserva únicamente en la tarjeta de despliegue del predecesor separado indexada por el Anexo J

1. Define y mide lo observable. Registra revisiones inmutables del modelo y del tokenizer, cuando estén disponibles, instantánea del corpus, precisión, backend de atención, selección de consultas, soporte de distancia, bins y agregación. Ajusta el perfil declarado y reporta la incertidumbre a nivel de secuencia. Un gran R^2 en sí mismo no identifica una arquitectura ni certifica una ley de potencia. Compara familias alternativas sobre las mismas observaciones y pérdida.

2. Indica cuál exponente se está utilizando. Use $\widehat{\gamma}$ para un ajuste medido γ para un parámetro familiar. La configuración candidata $\gamma_P = (2\theta - T\sqrt{2})/(2\theta + T\sqrt{2})$ lies in $(-1, 1)$ para entradas positivas y no puede sustituir a cada exponente observado. Utiliza los paneles de comparación identificados explícitamente en la Sección 5.4, no la afirmación sobre la precisión del caso seleccionado retirada. Las contrastes corporales requieren un protocolo emparejado y no identifican solamente circuitos.

3. Calcule la ventana de masa adecuada. Para distancias de tokens enteros use

$$D_{f,N} = \text{mín}\left\{k : \sum_{d=1}^{k} d^{-\gamma} \geq f Z_N(\gamma)\right\}.$$

Para la densidad continua distinta en $[1, N]$, use

$$D_{f,c} = \begin{cases} [(1-f) + fN^{1-\gamma}]^{1/(1-\gamma)}, & \gamma \neq 1, \\ N^f, & \gamma = 1. \end{cases}$$

Se establecen $N > 1$ y $0 < f < 1$. Estas son cuantiles de masa bajo una familia elegida, no ahorros de memoria medidos ni garantías sobre la precisión.

4. Interpreta el Pad inverso como álgebra.é Para $-1 < g < 1$ y $T > 0$,

$$\theta_{\text{inv}} = \frac{T\sqrt{2}(1+g)}{2(1-g)}, \qquad \alpha_{\text{cand}} = \frac{\theta_{\text{inv}}}{\theta_{\text{nominal}}}.$$

No task optimum follows. Sustituyendo al candidato en su contexto inverso devuelve T identicamente y no puede validar independientemente un horizonte de recuperación. Los inversos exponencial y lineal son funciones diferentes.

5. Evalúa la intervención. Utiliza ejemplos emparejados de prueba para medir la perplexidad o la calidad de la tarea junto con las ahorros reales de recursos. Incluye el modelo sin cambios y las bases de comparación relevantes. Mantén las mediciones faltantes distintas de los resultados nulos. Reporta el alcance de la incertidumbre y los casos de fallo. Una puntuación de anomalía sugiere una hipótesis; la identificación de circuitos causales requiere intervenciones adicionales.

F Resultados Negativos Productivos

Revisión del 7 de noviembre de 2026

El registro de resultados negativos se conserva en su totalidad. Los resultados numéricos conservan sus límites originales de protocolo y procedencia; su preservación no constituye una replicación fresca. Un fracaso en una analogía física no establece una teoría física de reemplazo. Las distinciones corregidas de base y de instrumento se aplican tanto a los resultados positivos como a los negativos.

Los identificadores históricos que se presentan a continuación se conservan, incluyendo el salto original de NR-10 a NR-12; no se proporciona resultado de NR-11. No son nueve replicaciones independientes. Un candidato fallido, una correlación débil y un experimento no informativo tienen significados diferentes. Ninguno establece una ley de potencia o una invariancia en el tiempo de inferencia de la atención aprendida.

NR-1: una mala parametrización de familia beta. Los informes históricos de regresión simbólica de 14 modelos reportan $R^2 = .052$ para la parametrización geométrica elegida. Esto no descarta todas las distribuciones beta ni prueba que las familias de potencia y exponencial agoten las alternativas. La unicidad de MaxEnt se mantiene bajo su restricción establecida; un fracaso en la regresión no es una prueba adicional de unicidad.

NR-2: razón no constante entre radio y longitud de entrenamiento. El coeficiente de variación histórico de 14 modelos .885 describe las razones seleccionadas $D_{.9}/T_{\text{train}}$. No establece ninguna invariante de reemplazo universal, ni deriva la dispersión empírica desde γ_P.

NR-3: La aplicabilidad de H3 no fue cuantificada por la cota propuesta. La cuenta legada mezcla distancias cortas/largas 50/100 con una posterior afirmación sobre distancia 1000, y reporta diferencias no positivas en las probabilidades logarítmicas para sus modelos RoPE probados. Estos ajustes del protocolo no pueden ser reconciliados solo a partir de ese texto. La escala continua de cero conteo $2\pi\theta$ no es un umbral identificado de degradación, y la propuesta de razón $n_{\text{dead}}(d_{\text{long}})/N > 1.5$ no tiene teorema de aplicabilidad validado. Los contrastes observados no establecen la causa raíz reclamada; se necesita un protocolo recuperable, un contraste de base no degenerado y parcheo controlado.

NR-4: Proximidad entropía atención ($|r| \leq 0.23$). Hipotetizamos que la entropía de atención por capa $H_{\text{attn}}(\ell)$ es una proxy inversa para la proporción de recuperación H3 $R(\ell)$: cuando el modelo no puede recuperar ($R = 0$), la atención debería ser uniforme

($H_{\text{attn}} \approx \ln T$); cuando la recuperación es completa ($R = 1$), la atención debería concentrarse ($H_{\text{attn}} \approx 0$). Correlaciones de Pearson en tres variantes de GPT-2 en la cabeza crítica h_0:

Modelo	$r(H_{\text{attn}},\, 1 - R)$	veredicto
gpt2-medium	-0.15	dirección opuesta
gpt2-large	-0.23	dirección opuesta
gpt2-xl	$+0.05$	signo esperado, pequeño

La hipótesis escrita predice una correlación positiva con $1 - R$. Así, $+.05$ tiene el signo esperado, mientras que los dos coeficientes negativos no lo hacen. Las correlaciones con muestras pequeñas no establecen independencia estadística, ortogonalidad geométrica, causalidad por cabeza inductiva o superioridad de gamma como resumen de información posicional. Los tres números permanecen informes históricos.

NR-5: Pendientes de rango de contenido y sus nulos específicos de rango. Predijimos $\gamma_{\text{long}} < \gamma_{\text{short}}$ a partir de la fórmula instantánea $\gamma = 1 - d\sqrt{2}/\theta$: mayores rangos deberían dar una decaimiento más lento. Los ajustes empíricos de la ley de potencia sobre $d \leq 100$ versus $d \geq 200$ dan los siguientes valores redondeados históricos, cuyo proceso subyacente no se identifica:

Model	θ	γ_{short}	γ_{long}	Δ
pythia-70m	10K	0.661	0.856	$+0.195$
LLaMA-3-8B	500K	0.812	1.041	$+0.229$
Qwen2.5-7B	1M	0.904	0.854	-0.049

Las pendientes redondeadas de Qwen mostradas se restan para obtener $-.050$; la retención de $-.049$ fue la diferencia históricamente reportada, la cual no puede recuperarse exactamente a partir de esas entradas redondeadas. Diferentes rangos de ajuste y la longitud variable del nulo afectan las pendientes. Para las rejillas E4 $\{30, 50, 100\}$ y $\{200, 500, 1000, 2000\}$, las pendientes con nulo uniforme son 0.525737 y 0.913768, dando $\Delta_0 = 0.388031$. Restando ese contraste nulo de los deltas reportados se obtiene aproximadamente -0.193, -0.159 y -0.437. Condicionalmente en esas rejillas, todos los tres tienen el signo previamente predicho después de la resta del nulo; «dirección invertida» se retira por tanto. Esto no valida al candidato instantáneo.

Repetir los textos JSON identificados por separado P10, P16 y P13 en exactamente esas rejillas da contrastes brutos -0.074078, -0.056949 y -0.047435; sus contrastes nulos restados son -0.462109, -0.444979 y -0.435466. Estos archivos no reproducen la tabla redondeada histórica, por lo que los dos conjuntos no se fusionan silenciosamente. No hay intervalos de confianza pareados disponibles para ninguna comparación de rangos. P4 requiere residuos nulos específicos de rango antes de interpretar el signo de curvatura.

NR-6/7/8 — Tres registros de perturbaciones en tiempo de inferencia. Tres experimentos intentaron manipular γ en tiempo de inferencia:

(a) Escaneo de temperatura ($\tau \in [0.3, 2.0]$, pythia-70m, pesos de 4 bits, cálculo en bf16, longitudes aleatorias de token $d+200$, consulta $d+199$ y clave 199, coincidiendo con el protocolo de (b), $N = 30$ prompts): ajuste $\gamma(\tau) = \gamma_0\, \tau^{-\beta}$ da $\beta = -0.008$, $R^2 = 0.014$—tendencia no monótona. El mínimo reproducido es .674860, máximo .757621, y rango

completo .082761. El mayor cambio absoluto desde la baseline $\tau = 1$ .749359 es .074499; el rango no es un intervalo simétrico ±.083. Las ocho medias almacenadas en $\tau = 1$ son exactamente idénticas a las ocho medias en base 10000 en (b), no simplemente iguales después de redondear. Ambos scripts reinician el mismo generador aleatorio y utilizan la misma extracción. Esto explica la baseline compartida a nivel de datos; sin identificadores de ejecución no se establece si se adquirió una o dos veces. No es independiente Evidencia de replicación. Softmax es invariante ante la suma de una escalar común, no ante la escala de los logits. Esta estabilidad de pendiente registrada no establece la invarianza de la matriz de atención.

(b) Barrido de base en tiempo de inferencia. Las nueve configuraciones de Pythia-70M usan pesos de 4 bits y cálculo en bfloat16, con 30 secuencias de entrada de tokens aleatorios por distancia, todas las cabezas y capas, y longitudes $d + 200$. El programa guardado modifica los búferes reales de frecuencias inversas con rotación parcial y sus copias originales. No se reentrenan pesos. Los nueve ajustes se reproducen a partir de las medias conservadas para ocho distancias. Su rango es .478980–.772036 (amplitud .293056), no $\leq$ .08; la mayor desviación respecto al ajuste con base nominal .749359 es .270379. El peor ajuste de potencia tiene $R^2 = .610621$ con base 4000. La regresión de las pendientes observadas frente a las candidatas sigue dando .152860 con $R^2 = .144304$. El nulo emparejado de atención uniforme es .439001 para este instrumento $d + 200$. Estas observaciones contradicen el seguimiento fiel del candidato, pero no demuestran invariancia frente a cambios de base ni un mecanismo por el que los pesos codifiquen la base de entrenamiento. El panel completo se mantiene a continuación.

Cuadro 48: Barrido inicial de nueve puntos durante la inferencia. Este protocolo con tokens cuantizados aleatorios difiere del piloto de entrenamiento.

θ	γ_P	$\widehat{\gamma}$	R^2
2000	0.171573	0.714375	0.974191
4000	0.477592	0.478980	0.610621
6000	0.618513	0.539640	0.830007
8000	0.699558	0.744934	0.987758
10000	0.752201	0.749359	0.986877
15000	0.827684	0.753198	0.961704
25000	0.892920	0.772036	0.970149
50000	0.944987	0.705375	0.916144
100000	0.972110	0.700762	0.980592

(c) Escalares de banda aprendidos (192 parámetros log-escalares entrenables, Adam $3{\times}10^{-3}$, 300 pasos, targets $\gamma \in \{0.5, 0.65, 0.9, 1.0\}$): máximo $|\Delta\gamma| = 0.063$; 5 de 6 capas convergen exactamente a $\alpha = 1.000$; la perplexidad permanece en la base aleatoria de tokens (PPL $\approx e^{10}$). Escalar $QK^\top$ por α^2 cambia generalmente el softmax; el resultado de optimización reportado requiere una auditoría de implementación y objetivo.

Perturbación	Rango	$\lvert\Delta\gamma_{\text{obs}}\rvert$	Mecanismo de bloqueo
Temperatura τ	[0.3, 2.0]	$\leq$0.08	Mecanismo no establecido
θ-steering	$[0.2, 10] \times \theta_{\text{train}}$	0.2704	No mechanism established
Aprendidos escalares	192 parámetros, 300 pasos	0.063306	Ningún mecanismo establecido

Los tres protocolos no establecen invariancia de atención ni identifican un mecanismo común de bloqueo. Sus estimadores, objetivos de optimización y precisión difieren. El registro de falla en seguir a un candidato es evidencia negativa para ese candidato bajo las condiciones probadas, no confirmación de que describa un exponente estructural inaccesible.

NR-9 — La afirmación sobre la escala espectral depende del instrumento. La antigua corrección de Pythia-1B para la dimensión de cabeza cambió la proporción recíproca reportada de aproximadamente 1.38 a 1.111, pero no reparó el agrupamiento completo de cabeza, el emparejamiento rotatorio o la ambigüedad de agregación. Reproducir los cinco registros guardados de Pythia da una correlación muestral fuerte $r = -0.986580$ para la proporción recíproca con la dimensión de cabeza y $r = +0.987556$ para M_{hist}. Estas son correlaciones de Pearson ordinarias sobre cinco tamaños confundidos, no leyes causales o correlaciones parciales. La estadística de dispersión del campo tiene su propia cota de longitud de contexto y regla para datos faltantes (Sección 14.3). El episodio respalda la auditoría del instrumento, no una afirmación de no efecto arquitectónico o una puntuación semántica validada.

NR-10 — Los costos de poda de coordenadas no validan una frontera de capa. Sección 11.7 preserva los registros v1 y completo v2, sus configuraciones de evaluación y las convenciones exactas de regresión. Las coordenadas de salida seleccionadas por el script disponible están fuera del subespacio rotatorio en las configuraciones auditadas Pythia-1B y 2.8B. La observación es por tanto una ablación de coordenada, no una remoción establecida de bandas vivas RoPE. Para un conjunto de datos seleccionado S, el modelo descriptivo es

$$y_i = a_S + b_S x_i + e_i, \qquad x_i = 100\,\mathrm{frac}_{Q/K,i}, \quad y_i = \Delta\mathrm{PPL}_i. \tag{81}$$

El coeficiente común original 0.175 ± 0.005 y $r > 0.99$ se retiran. Las pendientes positivas de v2 son 0.189964 (1B, tres puntos) y 0.313581 (2.8B, cuatro puntos); las pendientes que pasan por el origen son 0.153883 y 0.167242. El ajuste v1 de dos puntos 1B es otro subconjunto. Un pequeño costo positivo en una sola capa no establece una frontera de poda inocua; aumentar la intervención produce mayores costos registrados. También no convierte una identidad de entropía en una prescripción de poda.

NR-12 — La comparación de los tres modelos con decaimiento suave no tiene una frontera validada. La tabla completa de las cuatro políticas se encuentra en la Sección 11.6. Baseline tiene el menor puntaje PPL registrado para 2.8B, decaimiento suave para 1.4B y duro D_f para 1B. La regla propuesta anteriormente se mantiene solo como una hipótesis no validada:

$$h \gtrsim T_{\text{train}}/2 \quad \text{(historical cutoff heuristic, not an iff criterion)}. \tag{82}$$

El mejoramiento de 2.8B en comparación con el PPL con corte rígido es 1.155 %; 20.729 % utiliza el exceso sobre la base como denominador. La truncación dura en exactamente T_{train} no fue probada en este panel, ni se estableció masa retentora igual o un mecanismo de suavidad causal. Los casos no-op y fallidos permanecen parte del registro. Se necesitan los resultados por fragmento para la incertidumbre emparejada.

NR-13: la comparación V1/V2 presenta un error de entrada contextual. La comparación de veinte filas guardada define $g_{V2} = g_{V1}(1 - \log 2/d_{\text{head}})$. Recalcular sus columnas de error relativo reproduce errores medios 30.112029 % → 29.780277 %, errores mediana 15.463669 % → 15.997564 % y 13/20 de victorias para V2. Por lo tanto, el error medio mejora ligeramente mientras que el error mediano empeora; "no mejora en MAE" era incorrecto. El subconjunto solo Pythia $d_{\text{head}} \leq 80$ contiene seis filas, no cuatro.

Más fundamentalmente, todos los valores almacenados en V1 reproducen la fórmula racional cuando su slot de contexto es llenado por el correspondiente registro maestro's `T_attn`, con precisión numérica. Ese campo es la coordenada redondeada de temperatura de atención cerca de $1/g_{\text{obs}}$, no una longitud de contexto de evaluación. Por ejemplo, Pythia-1B utiliza $T = 1.0740$, obteniendo $g_{V1} = .999848125$, en lugar de un contexto de token como 2000. Esto confunde tanto observables como introduce una coordenada derivada de la respuesta en la predicción. El generador de este archivo de validación no fue reconstruido, pero la coincidencia exacta del valor almacenado identifica la discrepancia de entrada de forma independiente. Las veinte filas congeladas y el registro maestro están incluidas en el paquete de auditoría.

Por lo tanto, este panel no puede validar ni rechazar una corrección de cabeza finita en los contextos intencionales, establecer causas residuales o hacer canónica a V1. La derivación supuesta de continuo ya ha sido retirada (Apéndice B). Una nueva comparación requiere entradas reales de contexto, dimensiones rotatorias válidas, poblaciones de ajuste prespecifiadas y un resultado independiente; la antigua ajuste de respuesta de cinco ejes con fugas no proporciona ninguna de estas.

Interfaz de entrada de contexto correcta y recalculation ilustrativa. El reemplazador de calculadora acepta un entero positivo nombrado `context_tokens` una base positiva y una definición explícita de ese contexto. No acepta la antigua `T_attn` Esquema o inferir el contexto a partir de un exponente medido. En base 10000 y dimensión de cabeza 256, da:

Contexto declarado	racional V1	propuesto finito-cabeza V2
512	0.930122115	0.927603711
1024	0.864962305	0.862620328
2000	0.752201314	0.750164649
4096	0.550831282	0.549339848

Estos son ejemplos aritméticos condicionales, no nuevas mediciones o una validación de veinte modelos reparada. Los antiguos protocolos E4 utilizan longitudes de secuencia variables; no se puede inferir ningún contexto único a partir de sus pendientes agregadas. Una comparación científica debe declarar su contexto convencionalmente de forma independiente y alinear el protocolo de observación. El multiplicador propuesto V2 aún necesita justificación empírica. La validación de entrada previene el error específico de esquema/unidades; no puede certificar la procedencia científica de un entero suministrado por el usuario.

G Compilado de hallazgos críticos

Los veinte identificadores originales se conservan para hacer el testimonio trazable. Esta tabla establece su alcance actual, en lugar de reproducir las antiguas estrellas de

novedad como certificación. Los paneles numéricos históricos permanecen explícitamente identificados en las secciones citadas; el compendio completo antiguo se almacena en el archivo separado de predecesores, indexado por el Anexo. J No se establece ninguna afirmación de novedad mediante una identidad simbólica o una prueba formal sola.

Cuadro 49: Estado actual de F1–F20. Las observaciones empíricas conservadas no son nuevas repeticiones.

ID	Tema original	Conclusión actual	Ubicación
F1	Padé predictor	Candidato algebraico. Derivación universal retirada; Los paneles iniciales y de octubre tienen protocolos y denominadores diferentes.	§5.3
F2	Geometría rotativa	Conteos exactos de ciclos requieren el subespacio rotativo real y convención de emparejamiento. Finalizar un ciclo no demuestra la pérdida de información.	§4.2
F3	Límite en uno	Cota de sumabilidad con soporte infinito. Sumas finitas son analítico; $(\log N)^2/12$ es la varianza continua, no el valor discreto.	§7
F4	Ley de entropía	Cota de entropía finita y asintóticas distintas con exponente fijo. El antiguo factor dominante $(1-\gamma)$ es incorrecto.	§7.4
F5	Reservorio semántico	No respaldado por las máscaras inspeccionadas: B3 elimina conjuntos de coordenadas desiguales; el script de amplificación también cambia V.	§13
F6	Desvinculación de banda	Escaneo completo de B3 conservado. La máscara de coordenadas más grande tiene un PPL más alto en todas las seis ventanas; no hay disociación doble semántica/posicional.	§7.4
F7	ventana KV	Los cuantiles exactos discretos y continuos de masa son diferentes. Los ensayos de truncamiento registrados no establecen un rango universal que preserve la precisión.	§11
F8	escalado NTK	El multiplicador base difiere del factor de contexto. B1 razones hid Los costos PPL cortos; los conteos de clasificación de candidatos no establecen un Pareto dominancia o un óptimo universal.	§12
F9	Diseño base	Tabla corregida generada a partir de la ecuación inversa. En $T = 4096$, los objetivos 0.5 y 0.85 dan aproximadamente 8689 y 35721; Las asignaciones de tareas no están validadas.	§12.1
F10	presión GQA	hipótesis de asociación arquitectónica. Configuraciones mixtas Y los corpus no aíslan un efecto causal del número de cabezas KV.	§8.4
F11	Criterio de ebullición	Umbral para olvido o picos de pérdida son Una familia de capacidad térmica no es una probabilidad de riesgo medida.	§7.2

ID	Tema original	Conclusión actual	Ubicación
F12	umbral de inducción	una transición universal cerca de 400M parámetros no se respalda; el ajuste de soportes y mecanismos alternativos son relevantes.	§8.4
F13	confinamiento SWA	La máscara afecta las claves disponibles, pero el reportado Los contrastes entre modelos no identifican un desplazamiento universal del exponente.	§16
F14	Desplazamientos de alineación	Pares base/instrucción históricos conservan su diferencias medidas. No aíslan al RLHF del SFT, DPO u otros cambios.	§15
F15	Holografía	Una cota de entropía de Shannon no establece entrelazamiento geometría, una identificación Ryu–Takayanagi o carga central universal.	§7.4
F16	identidad de Fisher	$C_V = \gamma^2 I$ es exacta en la familia finita. La división por γ^2 excluye al cero, y los extremos no necesitan coincidir.	§7.3
F17	flujo RG	La ecuación diferencial con coordenadas racionales y su solución son álgebra condicional, no una trayectoria de entrenamiento medida.	§20.2
F18	Recuperación del controlador	Pools de ganchos históricos adaptativos para posiciones futuras y fallan la causalidad prefix; sus puntajes no certifican causal PPL. Los denominadores de recuperación difieren y la predicción de Fisher se retira. Pruebas de noviembre son evidencias separadas.	§19.2
F19	Decaimiento suave	Las comparaciones de políticas registradas permanecen específicas del protocolo. El sesgo del puntaje es $\log w(d)$, no $w(d)$; no se sigue una mejora general del PPL.	§11.6
F20	Costo de poda	Las pendientes en dos experimentos de modelos son ajustes descriptivos no es una ley universal de costo ni un intervalo de incertidumbre trans-modelo. La propuesta El corte de capa fue retirado.	Ap. F

H La Alpha-S Histórica: Definición y Fallo

Este apéndice define la propuesta de puntaje y vuelve a calcular todos los cuatro ejemplos a partir de sus entradas declaradas. La tabla original permanece en el archivo separado. Su interpretación semántica-ignición y las reglas de pre-entrenamiento go/no-go se retiran. La escala $N_0 = 4 \times 10^8$ es una suposición, no un umbral de cabeza de inducción desde Olsson et al. [34]; su evidencia incluye modelos de atención solo de dos capas. El símbolo N_{IH} en las ecuaciones históricas que siguen denota solo esa referencia asumida.

H.1 Definición conservada

$$\boxed{\alpha_S \;=\; F_{R_c} \cdot F_{\text{Chin}}} \tag{83}$$

donde los dos factores son:

$$F_{R_c} \;=\; \tfrac{1}{2}\big(1+\tanh\big(3(R_c - R_c^\star)\big)\big), \qquad R_c^\star \approx 1.68 \tag{84}$$

$$\text{(provisional, } n{=}4\text{; see caveat below and §16.4)}, \tag{85}$$

$$F_{\text{Chin}} \;=\; \log\frac{N_{\text{params}} \cdot D_{\text{tokens}}}{N_{\text{IH}} \cdot D_{\text{IH}}}, \qquad N_{\text{IH}} = 4{\times}10^8, \;\; D_{\text{IH}} = 2{\times}10^{11}. \tag{86}$$

El coeficiente 3, centro 1.68 y escala de token 2×10^{11} fueron elecciones heurísticas. El factor logarítmico no es una derivación de entrenamiento óptimo para el cálculo ni una probabilidad de formación de circuito; puede ser negativo por debajo de su referencia de producto elegida. No aparece una susceptibilidad Hagedorn en la definición real. La factorización no establece una unificación Chinchilla–Hagedorn.

H.2 Aritmética en los inputs históricos declarados

La siguiente tabla utiliza la mostrada R_c y valores con factor logarítmico según los inputs especificados. Recalcula ambos puertos y su producto directamente; los productos originales redondeados y las etiquetas del circuito se almacenan en el archivo de predecesores separado. Los presupuestos de tokens por modelo faltantes no se reconstruyen.

Modelo	R_c	estableció F_{Chin}	recalculó F_{R_c}	producto
Yi-6B	1.28	5.43	0.083173	0.451628
Llama-3.2-3B	1.33	5.89	0.109097	0.642580
Qwen2.5-3B	2.03	6.54	0.890903	5.826507
Qwen2.5-1.5B	3.09	5.85	0.999788	5.848761

La puntuación no es una probabilidad y estos inputs no son un conjunto de validación. Las señales de contraste corpus no suministran etiquetas directas para el circuito. En consecuencia, no se asigna ninguna afirmación sobre precisión de clasificación ni recomendación de presupuesto de entrenamiento a estos cuatro ejemplos aritméticos.

H.3 ¿Por qué se retira la regla de diseño?

La proporción de auditoría de nueve modelos en la Sección 16.4 presenta contraejemplos al puerta propuesto monótono. Su intersección rango-hull es $[0.924358, 3.085057]$, no una región de incertidumbre calibrada. El anteriormente citado fracaso de Pythia-2.8B "observado post-IH" también mezcla protocolos: el resumen de la ronda 3 tiene contraste positivo +0.041 mientras que otra comparación de corpus tiene un signo diferente. Ninguna establece una etiqueta de inducción única.

La propuesta alternativa $\alpha_S^{\text{v2}} = F_{\text{Chin}}[1 + c\tanh(3(R_c - R_c^\star))]$ fue una nueva hipótesis, no un reparo validado. Reajustar un puntaje después de observar fallos requiere evaluación separada en datos de prueba con resultados directos. Ninguna arquitectura, presupuesto o decisión de entrenamiento debería usar cualquiera de las claves como evidencia de que la inducción o la recuperación surgirá.

I Glosario Bilingüe

Estos términos se refieren a la familia estadística especificada a menos que se indique explícitamente una medición empírica. No identifican un equilibrio físico ni una propiedad de nivel de tarea de un modelo entrenado.

Término	Definición o alcance	Español
Exponente γ	Parámetro en $p_\gamma(d) \propto d^{-\gamma}$; distinguir de un ajustado $\widehat{\gamma}$.	Exponente
Coordinada de temperatura	$1/\gamma$, definida para $\gamma \neq 0$; distinta de la temperatura de muestreo de tokens.	Coordenada de temperatura
Normalizador Z_N	$\sum_{d=1}^{N} d^{-\gamma}$; no el normalizador de una fila de atención arbitraria.	Función de partición ón de partición ón
Coordenadas de energía libre	$\Phi = -\log Z_N$; $F = \Phi/\gamma$ para $\gamma \neq 0$.	Coordenadas de energía libre ía
Log moment U_N	$\mathbb{E}_{p_\gamma}[\log d]$.	Momento logarítmicoít-mico
Entropía S_N	Entropía de Shannon de la distribución de distancia declarada.	Entropía de distancias
Información de Fisher I	$\mathrm{Var}_{p_\gamma}(\log d)$ sobre soporte fijo.	Información de Fisher
Capacidad calorífica formal C_V	$\gamma^2 I$; distinta de la legada recíproca $1/\lvert 1-\gamma\rvert$.	Capacidad calorífica for-malífica
Legado recíproco χ	$1/\lvert 1-\gamma\rvert$, para $\gamma \neq 1$; no la información Fisher de familia finita.	Recíproco propuesto
Límite de sumabilidad	$\gamma = 1$ para soporte infinito; sumas finitas permanecen analíticas.	Frontera de sumabilidad
Dimensión de cabeza d_{head}	Número de coordenadas en una consulta/clave de cabeza; $d_{\text{rot}} \leq d_{\text{head}}$.	Dimensión de cabeza ón de cabeza
Contexto inverso d_{inv}	$\theta\sqrt{2}(1-g)/(1+g)$ para $\theta > 0, -1 < g < 1$ Una coordenada candidata.	Contexto inverso
Context T	Longitud del token declarada o coordenada de comparación; nunca inferida $T_{\text{attn}} = 1/\gamma$.	Contexto declarado
Ventana de masa retenida	Un cuantil en un soporte continuo o discreto especificado.	Ventana de masa retenida

Agradecimientos

Reconocemos a los autores de la investigación citada, a los contribuyentes del modelo y del conjunto de datos, así como a los mantenedores del software científico utilizado en el trabajo archivado. El paquete de corrección registra los módulos locales de Curiosity, Excovery, Socrates Audit y Triangulum, junto con sus hashes de origen y dependencias. Sus comprobaciones simbólicas y numéricas tienen el alcance indicado en el Apéndice A; los backends compartidos no son pruebas independientes. La responsabilidad de las definiciones, interpretaciones y hipótesis empíricas restantes de este manuscrito recae en su autor.

J Registro de Corrección y Archivo Histórico Separado

Este apéndice sustituye la anterior reproducción de treinta páginas de argumentos obsoletos. El completo documento previo PDF, su fuente y evidencia ZIP se incluyen sin cambios en el paquete de corrección. La fuente desplazada exacta también se registra en el log de cambios. El presente manuscrito contiene las declaraciones corregidas; las recetas obsoletas y los ejemplos numéricos inconsistentes no se presentan como un resultado actual alternativo.

Cuadro 50: Resolución de errores identificados. Las verificaciones de implementación no generan observaciones históricas faltantes.

ID	Issue	Resolución en esta edición
1	Predicción geométrica de la atención entrenada	Descomposición del puntaje conservada; premisas estadísticas faltantes explicitadas. Universalidad solo basada en configuración y la derivación inválida son retiradas. Secciones 5.2–5.4.
2	Confundimiento de medición de longitud variable	Las E4 y los nulos de campo utilizan sus protocolos reales. Se distingue la QL de consulta fija, con procedencia y observaciones faltantes declaradas. Secciones 9.1 y 14.3.
3	Termodinámica finita y continua	Separar normalizadores, momentos, información de Fisher y cuantiles; dominios correctos y valor finito en uno. Secciones 6.4–7.4.
4	garantías de rendimiento de KV	Presupuestos exactos de masa y cotas para filas congeladas reemplazan las afirmaciones de precisión basadas únicamente en configuración; las políticas experimentales y los denominadores permanecen distintos. Sección 11.
5	Interpretación de banda semántica/rotatoria	Se reportan el subespacio rotatorio actual, el emparejamiento y las máscaras de coordenadas adyacentes. Las abolliciones de coordenadas ya no identifican las bandas de frecuencia semántica. Secciones 11.7 y 13.
6	Base inversa y garantías NIAH	Protocolos de prueba correctos e inversos; las identidades algebraicas no son predicciones de recuperación independientes. Sección 12.
7	Corpus, etiquetas léxicas y arquitectónicas	Se indican observables coincidentes y registros de consulta/contador faltantes. Los signos no identifican cabezas de inducción, SWA o GQA. Secciones 14.3–16.
8	El control utiliza posiciones futuras	Una nueva referencia por consulta pasa por los controles de prefijo y partición. Los puntajes históricos defectuosos no se relabelan como causal PPL. Sección 19.3.
9	Límite incorrecto de proyección de cabeza	Nuevo referente actúa antes de la proyección de mezcla NeoX inspeccionada, con comprobaciones de disposición declaradas. Los bloques posteriores a la proyección no se consideran cabezas. Sección 19.3.
10	Rotaciones no unitarias, incorrectamente emparejadas	Las nuevas tablas utilizan frecuencias únicas y emparejamiento explícito; las comprobaciones de norma, rotación inversa y cola no rotada pasan. Sección 19.3.

ID	Issue	Resolución en esta edición
11	Resúmenes de controladores y denominadores	Todas las filas guardadas se conservaron; denominadores de recuperación y comparaciones de capas corregidas. Ningún reclamo de eficacia basado en Fisher. Sección 19.2.
12	Potencial, estabilidad y afirmación en tiempo finito	Potencial correcto y trayectoria exacta; punto final positivo repela y punto final negativo atrae en la dirección indicada. Sección 20.2.
13	Möbius dominio y prueba de punto fijo	Mapeo del semiplano derecho y determinante corregido; 0/23 agrupamiento es estructuralmente no informativo. Sección 20.3.
14	Identificaciones de fraccionario/CFT/BTZ	Algebra correcta y condiciones de dominio; equivalencias físicas no respaldadas retiradas. Sección 20.
15	Demostración completa y afirmaciones de Taylor	Pasos de promedio e normalización inválidos eliminados; composiciones inversas y errores numéricos corregidos. Apéndice B.
16	Zipf, grokking, signos y rangos	Paneles completados, dominios de correlación adecuados y soporte real de la sonda reemplazan a las resúmenes anteriores. Anexos D y F.
17	desajuste entre entrada de contexto V1/V2	Panel antiguo rechazado como validación; la interfaz de reemplazo requiere un contexto de token declarado de forma independiente. Se proporcionan cálculos ilustrativos explícitos en el Anexo F.
18	Catálogo, software y glosario	Conteos correctos, dominios de cálculo y terminología; la captura de pantalla del veredicto obsoleto fue reemplazada por un flujo de evaluación. Sección 21 y Anexo I.
19	Productos Alpha-S y etiquetas de circuitos	Productos recalculados a partir de entradas declaradas. Marcas de clasificación y decisiones de presupuesto de entrenamiento no respaldadas eliminadas. Anexo H.
20	Errores bibliográficos	Locadores incorrectos, títulos y autores corregidos; blogs identificados; citación sensible a versiones fijada; notas argumentativas eliminadas. Metadatos de la fuente y cambios exactos archivados.
21	Prosa contradictoria y notación archivística	Prosa errónea movida al archivo sucesor separado. d_{head} y d_{inv} ahora denotan objetos distintos en toda su extensión.

Evidencia que no puede recuperarse mediante la aritmética. Ninguna corrección puede reconstruir observaciones no registradas o revisiones inmutables en tiempo de ejecución. Las ausencias de registros D3 y de octubre, las poblaciones de consultas lexicográficas y las revisiones históricas de modelos/entornos permanecen como brechas explícitas en la procedencia. Las afirmaciones que requieren estas son no se sostienen. Las nuevas intervenciones causales, comparaciones V1/V2, nulos de recuperación o garantías de calidad de tarea necesitan experimentos nuevamente especificados antes de poder convertirse en conclusiones empíricas. Esta distinción no deja una fórmula errónea conocida como resultado actual; previene que la evidencia faltante sea reemplazada por un resultado inventado.

K Ocho Grupos de Resultados Verificados Formalmente

Este apéndice integra los ocho grupos de la nota técnica del 19 de noviembre en el manuscrito. Cada enunciado que sigue tiene hipótesis explícitas y una prueba legible; las fuentes Isabelle/HOL adjuntas proporcionan las derivaciones verificadas por máquina. La sesión completa auditó 300 hechos nombrados de las teorías del proyecto. Este es un recuento de hechos intermedios y finales, no 300 teoremas independientes. Los ocho lemas auxiliares de Lean tienen un alcance más restringido. Los resultados establecen implicaciones matemáticas, no que un modelo entrenado satisfaga sus premisas. No se introducen nuevas mediciones de modelos.

Convenciones. Todos los logaritmos son naturales. Los soportes de probabilidad son finitos y no vacíos a menos que se muestre explícitamente una integral continua. Utilizamos $\mathrm{TV}(p,q) = \frac{1}{2}\sum_j |p_j - q_j|$ y la divergencia KL habitual con valores reales extendidos, con los términos de masa cero omitidos y una masa positiva dividida por cero igual a $+\infty$ El exponente real g denota el mismo parámetro de familia que γ en otro lugar; es distinto de la pendiente de regresión observada. Para valores finitos v_j en un espacio vectorial normado real, $\mathrm{diam}\{v_j\} = \max_{i,j} \|v_i - v_j\|$.

Cuadro 51: Grupos verificados y su rol en el texto principal. Cada teorema se aplica solo bajo sus hipótesis establecidas.

Grupo	Resultado	Conexión principal de texto
1	Nulo instrumental general; resta exacta de pendientes logarítmicas; controles de consulta fija y diagonal	Sección 9
2	Límites estrictos, monotonicidad y límites para exposición afín	Sección 9.1
3	Pendiente local, curvatura y cotas de regresión para mezclas positivas	Sección 10
4	Orden discreto de cuantil en exponente y tamaño de soporte	Sección 11
5	Cálculo de familias suaves, entropía, límites KL y TV	Secciones 6.4–7.3
6	Error de truncamiento exacto, TV, KL y límites de salida precisos	Sección 11.2
7	Estabilidad del softmax aguda y el puente de puntuación logarítmica	Sección 5.2
8	Integral del coseno, resto explícito RoPE y aproximación condicional de fila	Sección 5.2

K.1 Grupo 1: el nulo instrumental y la calibración logarítmica

Fijar distancias $d_k > 0$, $x_k = \log d_k$, pesos de regresión positivos w_k con $\sum_k w_k = 1$, y $V_x = \mathrm{Var}_w(x) > 0$. Para cualquier perfil positivo A, incluyendo una intersección ajustada, definir

$$G_w(A) = -\frac{\mathrm{Cov}_w(x, \log A)}{V_x}. \tag{87}$$

Los puntos y pesos se mantienen fijos cuando se comparan dos pendientes.

Teorema K.1 (Calibración uniforme nula y exacta)**.** *En cada distancia d_k, la sonda promedia entradas de atención con pesos $\alpha_{kr} \geq 0$, $\sum_r \alpha_{kr} = 1$. Si la fila r tiene $n_{kr} \geq 1$ claves permitidas, su atención uniforme nula es*

$$A_{0,k} = \sum_r \frac{\alpha_{kr}}{n_{kr}}, \qquad G_w(A_0) = -\frac{\mathrm{Cov}_w(x, \log A_0)}{V_x}.$$

Para cada perfil positivo en la misma rejilla y pesos de regresión,

$$G_w(A/A_0) = G_w(A) - G_w(A_0). \tag{88}$$

Una constante nula tiene pendiente cero. Mantener las mismas consultas, tamaños de soporte y pesos promedio en cada distancia es suficiente para una constante nula.

Demostración. Una entrada permitida en una fila uniforme es igual a $1/n_{kr}$. El promedio da como resultado la nula mostrada, la cual es positiva. Sustituye $\log(A_k/A_{0,k}) = \log A_k - \log A_{0,k}$ en la ecuación (87); la covarianza es lineal en su segundo argumento. La covarianza con una constante es cero. □

Consultas fijas y diagonales completas. Para un contexto causal fijo T, índices de consulta basados en cero $i \in Q$ y pesos fijos α_i, supongamos que cada distancia ajustada es elegible para cada consulta. Entonces $A_{0,\mathrm{QL}}(d) = \sum_{i \in Q} \alpha_i/(i+1)$ es constante. Por el contrario, promediando cada entrada de la diagonal causal da

$$A_{0,\mathrm{ALL}}(d) = \frac{1}{T-d} \sum_{i=d}^{T-1} \frac{1}{i+1}, \qquad d \in \{1, \ldots, T-1\}.$$

Este nulo disminuye estrictamente entre distancias elegibles sucesivas: eliminar el mayor miembro de la secuencia estrictamente decreciente $1/(d+1), \ldots, 1/T$ disminuye su media. Por lo tanto, su exponente OLS es positivo en cualquier red con al menos dos distancias distintas y pesos positivos. La identidad de covarianza pairwise en Grupo 2 prueba el signo estricto.

Alcance y contraejemplos. La resta produce una pendiente relativa a un nulo declarado; no demuestra estimación insesgada ni atribución causal. El control debe reproducir la agrupación en intervalos, las exclusiones y los pesos, pues el promedio y el logaritmo no conmutan. Fijar las consultas es suficiente, pero no necesario: una fila con seis claves permitidas y el promedio de igual peso de filas con cuatro y doce claves permitidas dan ambos 1/6. Asimismo, una covarianza nula en una cuadrícula no implica que el perfil sea constante. En atención por producto escalar sin sesgo aditivo dependiente de la clave, anular la proyección de consultas deja todas las consultas rotadas a cero y produce un softmax uniforme. Una sonda de consultas fijas tiene entonces exponente cero para cualquier base RoPE. Esto refuta una ley geométrica no nula válida para todos los pesos posibles; no caracteriza un procedimiento de entrenamiento concreto.

K.2 Grupo 2: cotas para rejillas finitas en exposición afín

La identidad secante. Para los pesos anteriores y observaciones reales y_k

$$\mathrm{Cov}_w(x,y)=\sum_{i<j} w_i w_j (x_j-x_i)(y_j-y_i), \qquad V_x=\sum_{i<j} w_i w_j (x_j-x_i)^2.$$

Expandir la correspondiente mitad-suma sobre todos los pares ordenados prueba ambas identidades. Por lo tanto, la pendiente de regresión es una combinación convexa de secantes sobre pares con valores distintos de x, con pesos $w_i w_j (x_j-x_i)^2/V_x$.

Teorema K.2 (Sesgo de exposición afín)**.** *Para $A_0(d)=c/(ad+b)$ con $a,b,c>0$, en la malla no degenerada del Grupo 1*

$$\frac{ad_{\min}}{ad_{\min}+b} < G_w(A_0) < \frac{ad_{\max}}{ad_{\max}+b}. \tag{89}$$

Con la cuadrícula y los pesos fijos, esta pendiente disminuye estrictamente con b. Sustituyendo todas las distancias por λd_k, $\lambda>0$, hace que la pendiente aumente estrictamente con λ; sus límites cuando $\lambda\downarrow 0$ y $\lambda\to\infty$ son respectivamente cero y uno. La amplitud c no tiene efecto.

Demostración. La pendiente es la regresión de $f(x)=\log(ae^x+b)$ sobre x. Su derivada $ae^x/(ae^x+b)$ aumenta estrictamente. El teorema del valor medio coloca cada secante no trivial estrictamente entre las derivadas en los puntos extremos; la combinación convexa prueba los límites. Derivando la covarianza finita se obtiene

$$\partial_b G_w=\frac{\mathrm{Cov}_w(x,1/(ae^x+b))}{V_x}<0, \qquad \partial_\lambda G_w=\frac{\mathrm{Cov}_w(x,ad/(a\lambda d+b))}{V_x}>0.$$

Los signos se derivan de la identidad por pares y el orden estrictamente opuesto o común. Sumar $\log\lambda$ a cada x_k no cambia ni las coordenadas centradas ni la varianza. En una cuadrícula finita, el perfil logarítmico tiende a una constante a escala cero, mientras que $\log(a\lambda d+b)=\log(a\lambda)+\log d+o(1)$ a escala infinita. Las constantes desaparecen de la covarianza, incluyendo $\log c$. □

Para el nulo E4 inicial, $a=1, b=50$. Su pendiente dependiente de escala está por tanto presente sin aprendizaje. Los casos límite, fuera de las hipótesis estrictas, dan pendiente uno para $b=0, a>0$, y cero para $a=0, b>0$.

K.3 Grupo 3: los perfiles de mezcla cambian la pendiente y la curvatura

Teorema K.3 (Mezclas positivas de potencias)**.** *Sea $J\ge 1$, $a_j>0$, $g_j\in\mathbb{R}$, y $A(d)=\sum_{j=1}^J a_j d^{-g_j}$ para $d>0$, con coeficientes y exponentes independientes de la distancia. Para $x=\log d$, ponga $\pi_j(x)=a_j e^{-g_j x}/\sum_\ell a_\ell e^{-g_\ell x}$. Luego*

$$g_{\mathrm{loc}}(x)=-\frac{d}{dx}\log A(e^x)=\sum_j \pi_j(x) g_j, \qquad g'_{\mathrm{loc}}(x)=-\mathrm{Var}_{\pi(x)}(g)\le 0. \tag{90}$$

El exponente de la red finita satisface

$$\min_j g_j \le g_{\mathrm{loc}}(x_{\max}) \le G_w(A) \le g_{\mathrm{loc}}(x_{\min}) \le \max_j g_j.$$

Si dos exponentes componentes difieren, la pendiente local disminuye estrictamente, y la mezcla no puede ser una potencia exacta en ningún intervalo abierto.

Demostración. La diferenciación logarítmica da como resultado la media ponderada. Dado que $\pi_j' = \pi_j(-g_j + \sum_\ell \pi_\ell g_\ell)$, diferenciarlo da como resultado la varianza negativa. Cada π_j es positiva, por lo que la varianza se anula exactamente cuando todos los exponentes componentes coinciden. Aplicando la identidad de la secante a $-\log A(e^x)$, se acota la regresión por sus derivadas extremas. Una potencia exacta tendría pendiente local constante. □

Los normalizadores finitos constantes pueden absorberse en a_j. Cambiar las cabezas, sus pesos o sus soportes con la distancia puede violar las hipótesis. Esta identidad log-sum-exp no establece que las cabezas individuales sean potencias ni identifica sus roles computacionales.

K.4 Grupo 4: ordenamiento de cuantiles discretos de masa

Teorema K.4 (Exponente y orden de soporte)**.** *Sea $N \geq 2$ un entero y $p_g(d) = d^{-g}/Z_N(g)$ sobre $\{1, \ldots, N\}$, con $g \in \mathbb{R}$. Para $1 \leq k < N$, la masa de prefijo $F_g(k) = \sum_{d \leq k} p_g(d)$ es estrictamente creciente en g. Para $0 < f < 1$, el cuantil entero bien definido*

$$D_f(g, N) = \min\{k \in \{1, \ldots, N\} : F_g(k) \geq f\}$$

es no creciente en g y, para g, f fijo, es no decreciente en N. Estas conclusiones incluyen exponentes negativos, cero y uno.

Demostración. El conjunto definitorio es finito y contiene N. Pongamos $X = \log d$ y $L_k = \{1, \ldots, k\}$. Derivando las sumas finitas da

$$\partial_g F_g(k) = -\operatorname{Cov}_{p_g}(\mathbf{1}_{L_k}, X) = P(L_k)P(L_k^c)\big(\mathbb{E}[X \mid L_k^c] - \mathbb{E}[X \mid L_k]\big) > 0.$$

Ambos grupos tienen masa positiva y cada logaritmo en el segundo excede a cada logaritmo en el primero. Aumentar g puede solo avanzar la primera intersección de f. Aumentar N deja inalterado el numerador de un prefijo antiguo y aumenta su denominador; no puede producir una intersección anterior. □

El cuantil entero es una función escalonada, por lo tanto su monotonía no necesita ser estricta. Esta propiedad de familia proporciona un presupuesto de masa, no una garantía de tarea KV.

K.5 Grupo 5: cálculo de familia finita y sensibilidad

Teorema K.5 (Límites de suavidad, entropía y divergencia)**.** *Para enteros $N \geq 1$ y reales g, h, se definen $\psi(g) = \log Z_N(g)$, $U(g) = \mathbb{E}_{p_g} \log d$, $I(g) = \operatorname{Var}_{p_g}(\log d)$ y $S(g) = -\sum_d p_g(d) \log p_g(d)$. Todas estas funciones son suaves de cualquier orden para familia finita g, y*

$$\psi' = -U, \quad \psi'' = I, \quad U' = -I, \quad S = \psi + gU, \quad S' = -gI, \quad C_V = g^2 I.$$

Escribiendo $L = \log N$,

$$0 \leq I(g) \leq L^2/4, \qquad 0 \leq C_V(g) \leq g^2 L^2/4.$$

Además,

$$\begin{aligned}\mathrm{KL}(p_g\|p_h) &= (h-g)U(g)+\psi(h)-\psi(g)\\ &= (h-g)^2\int_0^1 (1-t)I(g+t(h-g))\,dt\\ &\le (h-g)^2L^2/8, \qquad (91)\\ \mathrm{TV}(p_g,p_h) &\le \tanh\left(\frac{|h-g|L}{4}\right) \le \min\{1,|h-g|L/4\}. \qquad (92)\end{aligned}$$

Para $N \ge 2$, $I(g) > 0$ y S tiene su máximo único en $g = 0$.

Demostración. Diferencia las sumas finitas usando $p'_g = -p_g(\log d - U)$. La positividad de Z_N hace que su logaritmo y todas las cocientes sean suaves. Para $X \in [0, L]$, $\mathbb{E}[X(L - X)] \ge 0$ da $\mathrm{Var}\, X \le L\mathbb{E}X - (\mathbb{E}X)^2 \le L^2/4$. La log-ratio de probabilidades da la identidad KL. El resto integral de Taylor para ψ da la siguiente igualdad, también para $h < g$; integrando la cota de varianza se obtiene $L^2(h-g)^2/8$. Las dos distribuciones de potencia son softmaxes cuya diferencia de puntaje tiene oscilación $|h-g|L$. El Grupo 7 prueba la cota TV, y $\tanh u \le \min\{1, u\}$ para $u \ge 0$ da su forma lineal. Para $N \ge 2$ al menos dos log-distancias distintas tienen masa positiva; así que $I > 0$ y $S' = -gI$ prueba el máximo único. Para $N = 1$ todas las cantidades mostradas y divergencias son cero. □

Función generadora y medida de referencia. En general, para pesos positivos fijos a_j y energías reales E_j, la función $\psi(t) = \log \sum_j a_j e^{-tE_j}$ cumple $-\psi' = \mathbb{E}_t E$ y $\psi'' = \mathrm{Var}_t(E)$. Los Grupos 3 y 5 son casos con energías g_j y $\log d$, respectivamente. La identidad de Shannon $S = \psi + tU$ presupone pesos de referencia unitarios; para pesos arbitrarios a_j aparece además el término $-\mathbb{E}_t \log a_j$. En soporte finito, $g = 1$ no es singular. El valor continuo $(\log N)^2/12$ no debe sustituirse por la varianza discreta en uno.

K.6 Grupo 6: qué garantiza la truncación de fila renormalizada

Teorema K.6 (Divergencias exactas y cotas precisas de salida congelada)**.** *Sea p una probabilidad sobre un conjunto finito y no vacío de claves y conserve un subconjunto K con $p(K) = 1 - \epsilon > 0$. Establezca $q_j = p_j/(1-\epsilon)$ en K y $q_j = 0$ fuera.*

$$\mathrm{TV}(p,q) = \epsilon, \qquad \mathrm{KL}(q\|p) = -\log(1-\epsilon).$$

Si $\epsilon > 0$, entonces $\mathrm{KL}(p\|q) = +\infty$. Para valores fijos v_j, $o = \sum_j p_j v_j$ y $o_K = \sum_j q_j v_j$ satisfacen

$$\|o - o_K\| \le \epsilon\, \mathrm{diam}\{v_j\} \le 2\epsilon \max_j \|v_j\|. \qquad (93)$$

Para $\epsilon > 0$, si o_R es la media condicional sobre claves eliminadas, $o - o_K = \epsilon(o_R - o_K)$. Las constantes uno y dos son óptimas en general.

Demostración. En K, las diferencias no negativas $q_j - p_j$ suman a ϵ; fuera, las p_j también suman a ϵ. Esto prueba la TV. En el soporte positivo de q, su razón de verosimilitud respecto a p es $1/(1-\epsilon)$, dando KL inversa. La masa eliminada positiva crea una p_j positiva con $q_j = 0$, dando KL hacia adelante infinita. Para $\epsilon > 0$, descomponga $o = (1-\epsilon)o_K + \epsilon o_R$. La diferencia de las dos medias condicionales es una combinación convexa de diferencias $v_r - v_k$ y está acotada por el diámetro, que es a lo sumo el doble de la norma máxima. Para $\epsilon = 0$ las dos distribuciones y las salidas coinciden. Un valor retenido $-Bu$ y un valor eliminado Bu, con $\|u\| = 1$, $B > 0$ y masas $1 - \epsilon, \epsilon$, alcanzan ambos límites. □

Las puntuaciones y valores están congelados. Una proyección lineal posterior contribuye con su norma operacional; la propagación a través de una red cambiante requiere suposiciones adicionales de estabilidad. No se sigue ninguna garantía de perplexidad o calidad de respuesta por sí sola.

K.7 Grupo 7: de errores en puntajes a errores tras el softmax

Teorema K.7 (Estabilidad óptima en función de la oscilación)**.** *En un soporte finito no vacío, sean* $q = \mathrm{softmax}(s)$ *y* $p = \mathrm{softmax}(s + r)$ *para puntajes reales finitos. Con* $\omega = \max_j r_j - \min_j r_j$ *,*

$$\mathrm{TV}(p, q) \le \tanh(\omega/4). \tag{94}$$

El límite es óptimo sobre distribuciones positivas y errores con esta oscilación. Si $|r_j| \le \eta$*, entonces la distancia total es a lo sumo* $\tanh(\eta/2)$*. Para valores fijos,*

$$\left\| \sum_j (p_j - q_j) v_j \right\| \le \mathrm{diam}\{v_j\} \tanh(\omega/4).$$

Demostración. Un desplazamiento constante de los puntajes se cancela, lo que prueba el caso $\omega = 0$. En otro caso, desplace los errores de modo que $Y_j = e^{r_j} \in [1, M]$, $M = e^{\omega} > 1$, y sea $\mu = \mathbb{E}_q Y$. Entonces $p_j = q_j Y_j/\mu$, de modo que $\mathrm{TV}(p, q) = \mathbb{E}_q|Y - \mu|/(2\mu)$. La cuerda de la función convexa $|y - \mu|$ en $[1, M]$ da

$$\mathrm{TV}(p, q) \le \frac{(\mu - 1)(M - \mu)}{\mu(M - 1)} \le \frac{\sqrt{M} - 1}{\sqrt{M} + 1} = \tanh(\omega/4).$$

La segunda desigualdad sigue de $(\mu - \sqrt{M})^2 \ge 0$. La igualdad se alcanza con $Y = 1, M$ y probabilidades respectivas $\sqrt{M}/(1 + \sqrt{M})$ y $1/(1 + \sqrt{M})$, lo cual da $\mu = \sqrt{M}$. Estas probabilidades estrictamente positivas se realizan mediante puntajes softmax finitos. La cota de error absoluto da $\omega \le 2\eta$. Finalmente, las partes positiva y negativa de $p - q$ cada una tienen masa TV; normalizarlas expresa la diferencia de salida como TV multiplicado por una diferencia de combinaciones convexas, acotada por el diámetro. Si TV es nulo, la diferencia de salida es cero directamente. □

Corolario del puntaje logarítmico. Si los puntajes sobre un conjunto declarado finito no vacío de distancias positivas son $s_d = b - g \log d + r_d$, su distribución por fila está dentro de $\tanh((\max r - \min r)/4)$ en la distancia total variacional de $d^{-g}/\sum_{u \in \mathcal{D}} u^{-g}$ sobre el mismo soporte. Un resto constante proporciona igualdad exacta. Esto se sigue al aplicar el teorema a los puntajes no perturbados $b - g \log d$. La hipótesis se refiere a los puntajes individuales por fila. Una afirmación sobre los puntajes promedio no permite intercambiar la esperanza y el softmax.

K.8 Grupo 8: una integral coseno con un residuo explícito RoPE

Teorema K.8 (Aproximación de red coseno y residuo logarítmico)**.** *Para enteros* $m \ge 1$*,* $\theta > 1$*,* $L = \log \theta$ *y reales* $d > 0$*, definir*

$$S_m(d) = \sum_{i=0}^{m-1} \cos(d\theta^{-i/m}), \qquad J_m(d) = \frac{m}{L}\big[\,\mathrm{Ci}(d) - \mathrm{Ci}(d/\theta)\big],$$

donde $\operatorname{Ci}(x) = -\int_x^\infty \cos t/t\, dt$ *para* $x > 0$. *Luego*

$$|S_m(d) - J_m(d)| \le \int_{d/\theta}^{d} |\sin u|\, du \le d(1 - 1/\theta), \tag{95}$$

$$\left|S_m(d) - \frac{m}{L}(L - \gamma_E - \log d)\right| \le B_m(d), \tag{96}$$

$$B_m(d) = d(1 - 1/\theta) + \frac{m}{L}\left(\frac{2}{d} + \frac{d^2}{4\theta^2}\right),$$

con la constante de Euler γ_E. *A valor fijo de* d, θ, *la aproximación de Riemann normalizada tiene un error a lo sumo de* $d(1 - 1/\theta)/m$, *que tiende a cero cuando* $m \to \infty$.

Demostración. Definamos $f(t) = \cos(de^{-Lt})$. La sustitución da $m\int_0^1 f(t)\, dt = J_m(d)$. En cada celda $I_i = [i/m, (i + 1)/m]$,

$$\left|f(i/m) - m\int_{I_i} f(t)\, dt\right| \le m\int_{I_i}\int_{i/m}^{t} |f'(u)|\, du\, dt \le \int_{I_i} |f'(u)|\, du.$$

Sumar y cambiar variables prueba la cota de seno absoluto. La integración por partes en la integral de cola da $|\operatorname{Ci}(d)| \le 2/d$. La normalización estándar de Euler es

$$\operatorname{Ci}(y) = \gamma_E + \log y + \int_0^y \frac{\cos t - 1}{t}\, dt.$$

La desigualdad $|1 - \cos t| \le t^2/2$ acota el último integral en magnitud por $y^2/4$. Aplica estas cotas en d y $y = d/\theta$, multiplícalas por m/L y suma el error de Riemann. Dividiendo la ecuación (95) por m se demuestra la convergencia normalizada. El desarrollo formal establece también el vínculo con la función especial normalizada de Euler, en lugar de asumir una constante de integración no especificada. □

Transferencia condicional a una fila de atención. Sean $h > 0$, $a, c \in \mathbb{R}$ y $\rho \ge 0$, bajo las condiciones del teorema. En un soporte finito no vacío de distancias positivas, supongamos que las puntuaciones reales de la fila satisfacen $s_d = c + aS_m(d)/\sqrt{h} + e_d$, $|e_d| \le \rho$. Definamos

$$g = \frac{am}{\sqrt{h}\log\theta}, \qquad \eta = \frac{|a|}{\sqrt{h}} \max_{d\in\mathcal{D}} B_m(d) + \rho. \tag{97}$$

Su softmax está dentro de $\tanh(\eta/2)$ en TV de la familia de leyes de potencia con exponente g en el mismo soporte. De hecho, el puntaje es $b - g\log d + r_d$, donde $b = c + g(L - \gamma_E)$ y $|r_d| \le \eta$; se aplica el grupo 7. El soporte finito no vacío hace que el máximo sea válido.

Equivalentemente, coloca

$$\delta_m(d) = \frac{L}{m}B_m(d) = \frac{Ld}{m}(1 - 1/\theta) + \frac{2}{d} + \frac{1}{4}(d/\theta)^2.$$

Entonces $\eta = |g| \max_d \delta_m(d) + \rho$. Los tres términos separan la discretización, los errores a pequeña distancia y los errores a gran distancia relativa. El límite puede ser grande y no informativo. La convergencia normalizada hacia la integral coseno no elimina su error de aproximación logarítmica. Las amplitudes desiguales en el plano, componentes seno, coordenadas no rotadas y la dependencia del contenido deben satisfacer la hipótesis del puntaje o ser controlados en e_d. La distancia propia $d = 0$ está fuera de este soporte logarítmico y debe ser tratada por separado si se incluye en una fila de atención completa. No existe fórmula para $\gamma_P(\theta, T)$ o para el exponente aprendido sin premisas adicionales.

K.9 Prueba de procedencia, reproducibilidad y alcance restante

El snapshot verificado completo se suministra con las fuentes del manuscrito en `research_program/theorems_20261118/verified_ci_300_control/`Contiene 38 archivos teóricos, la sesión `ROOT` El recibo de construcción exitosa, la auditoría de dependencias y el registro de control negativo. Todos los 39 hashes de archivos de origen y de sesión fueron verificados contra ese recibo al integrar este apéndice. La sesión completa utiliza Isabelle2026/HOL y la entrada no modificada de la AFP. `Exp_Log_Integral` con su cierre de dependencias del 6 de noviembre de 2026. Las pruebas de las funciones especiales upstream no se consideran como resultados propios del autor.

Cuadro 52: Anclas formales representativas. El mapa adjunto `research_program/part1_formal_20261119/formal_coverage.json` asigna todos los grupos a las teorías suministradas y a los nombres de hechos auditados.

Grupo	Teoría y resultado nombrado
1	`TAF_Probe.log_slope_ratio`; `TAF_All_Null.diagonal_null_positive_slope`
2	`TAF_Affine_Bounds.affine_regression_strict_bounds`; `TAF_Affine_Limits.affine_slope_limits`
3	`TAF_Regression_Calculus.mixture_regression_bounds`; `TAF_Strict_Family.non_affine_log_partition`
4	`TAF_Power_Family.quantile_nonincreasing`; `TAF_Power_Support.power_quantile_support_order`
5	`TAF_Smooth_Family.finite_family_all_orders`; `TAF_KL_Remainder.kl_integral_remainder`; `TAF_Sensitivity.family_tv_oscillation`
6	`TAF_Truncation.truncation_reverse_kl`; `TAF_Transport.frozen_output_diameter`
7	`TAF_Softmax.softmax_oscillation_bound`; `TAF_Log_Scores.logarithmic_score_stability`; `TAF_Log_Scores.softmax_sharp_binary`
8	`TAF_RoPE_Ci.rope_cosine_logarithmic_remainder`; `TAF_RoPE_Scores.conditional_rope_power_row`

Nombres en la Tabla 52 omiten cualificadores de localización para mayor legibilidad; los nombres completamente cualificados exactos se conservan en el mapa legible por máquina. Los 300 hechos auditados de proyecto no tienen dependencias de oráculo; se utilizó `quick_and_dirty=false` en la construcción. Su control negativo agregó lemas falsos a una copia y requirió la rechazo. La sesión anterior de 271 hechos es un hito de subconjunto, no un conjunto adicional de 271 resultados independientes. Ocho lemas algebraicos de Lean verificados por separado son auxiliares, no una sustitución de estos ocho grupos completos. La integración preserva estas fuentes verificadas y recibos; no afirma que los asistentes de prueba fueran reejecutados para el formato.

Para la reproducción con Isabelle2026 y AFP 2026-11-06 instalados, ejecute `isabelle build` con un trabajo, un hilo, quick-and-dirty desactivado, las teorías AFP como directorio de sesión y la instantánea proporcionada como sesión a construir. El comando exacto y la procedencia de las dependencias se proporcionan en las instrucciones del paquete fuente.

Las pruebas escritas y los objetos formales establecen los resultados referidos a la familia finita, instrumento y aproximación. No certifican todas las afirmaciones del manuscrito, la novedad bibliográfica, la falta de procedencia empírica, la unicidad de un máximo de

capacidad térmica para cada tamaño de soporte, las etiquetas aprendidas de circuitos, o la calidad de la tarea después de las intervenciones. El histórico 20-lema de auditoría en el Apéndice A sigue siendo un registro separado con su alcance original.

Referencias

[1] Máximo Bañados, Claudio Teitelboim, and Jorge Zanelli. The black hole in three dimensional space time. *Physical Review Letters*, 69:1849–1851, 1992. doi: 10.1103/PhysRevLett.69.1849. URL https://arxiv.org/abs/hep-th/9204099.

[2] Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and Round We Go! What makes Rotary Positional Encodings useful? *arXiv preprint arXiv:2410.06205*, 2024. URL https://arxiv.org/abs/2410.06205.

[3] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. *arXiv preprint arXiv:2304.01373*, 2023. URL https://arxiv.org/abs/2304.01373.

[4] Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Yucheng Li, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Junjie Hu, and Wen Xiao. PyramidKV: Dynamic KV Cache Compression based on Pyramidal Information Funneling. *arXiv preprint arXiv:2406.02069*, 2024. URL https://arxiv.org/abs/2406.02069.

[5] Pasquale Calabrese and John Cardy. Entanglement entropy and quantum field theory. *Journal of Statistical Mechanics: Theory and Experiment*, 2004(06):P06002, 2004. URL https://arxiv.org/abs/hep-th/0405152.

[6] Byung Gyu Chae. Self-Organized Criticality from Protected Mean-Field Dynamics: Loop Stability and Internal Renormalization in Reflective Neural Systems. *arXiv preprint arXiv:2601.04450*, 2026. URL https://arxiv.org/abs/2601.04450.

[7] Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Critical attention scaling in long-context transformers. *arXiv preprint arXiv:2510.05554*, 2025. URL https://arxiv.org/abs/2510.05554.

[8] Ting-Rui Chiang and Dani Yogatama. The Rotary Position Embedding May Cause Dimension Inefficiency in Attention Heads for Long-Distance Retrieval. *arXiv preprint arXiv:2502.11276*, 2025. URL https://arxiv.org/abs/2502.11276.

[9] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006. ISBN 978-0-471-24195-9.

[10] Róbert Csordás, Christopher D. Manning, and Christopher Potts. Do Language Models Use Their Depth Efficiently? *arXiv preprint arXiv:2505.13898*, 2025. URL https://arxiv.org/abs/2505.13898.

[11] Hugo Cui, Freya Behrens, Florent Krzakala, and Lenka Zdeborová. A phase transition between positional and semantic learning in a solvable model of dot-product attention. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[12] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

[13] Zihan Gu, Ruoyu Chen, Han Zhang, Hua Zhang, and Yue Hu. Deconstructing Positional Information: From Attention Logits to Training Biases. *arXiv preprint arXiv:2505.13027*, 2025. URL https://arxiv.org/abs/2505.13027.

[14] Rolf Hagedorn. Statistical thermodynamics of strong interactions at high energies. *Nuovo Cimento Suppl.*, 3:147–186, 1965.

[15] Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024. URL https://arxiv.org/abs/2404.15255.

[16] Jonghyun Hong and Sungyoon Lee. Variance sensitivity induces attention entropy collapse and instability in transformers. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8360–8378, 2025. doi: 10.18653/v1/2025.emnlp-main.421. URL https://aclanthology.org/2025.emnlp-main.421/.

[17] Noah Hong and Tao Hong. Evidence of Phase Transitions in Small Transformer-Based Language Models. *arXiv preprint arXiv:2511.12768*, 2025. URL https://arxiv.org/abs/2511.12768.

[18] Edwin T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106 (4):620–630, 1957. doi: 10.1103/PhysRev.106.620.

[19] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. URL https://arxiv.org/abs/2401.04088.

[20] Gunn Kim. Thermodynamic isomorphism of transformers: A Lagrangian approach to attention dynamics. *arXiv preprint arXiv:2602.08216*, 2026. URL https://arxiv.org/abs/2602.08216v1. Version 1, 9 February 2026; later versions substantially revised the title and scope.

[21] Sejin Kim. Learning the Inverse Ryu–Takayanagi Formula with Transformers. *Journal of High Energy Physics*, 2026. URL https://arxiv.org/abs/2511.06387.

[22] Mateusz Kwaśnicki. Ten equivalent definitions of the fractional laplace operator. *Fractional Calculus and Applied Analysis*, 20(1):7–51, 2017. doi: 10.1515/fca-2017-0002. URL https://arxiv.org/abs/1507.07356.

[23] Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[24] Feilong Liu. Rotary Positional Embeddings as Phase Modulation: Theoretical Bounds on the RoPE Base for Long-Context Transformers. *arXiv preprint arXiv:2602.10959*, 2026. URL https://arxiv.org/abs/2602.10959.

[25] Minghui Liu, Aadi Palnitkar, Tahseen Rabbani, Hyunwoo Jae, Kyle Rui Sang, Dixi Yao, Shayan Shabihi, Fuheng Zhao, Tian Li, Ce Zhang, Furong Huang, and Kunpeng Zhang. Hold Onto That Thought: Assessing KV Cache Compression On Reasoning. *arXiv preprint arXiv:2512.12008*, 2025. URL https://arxiv.org/abs/2512.12008.

[26] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. In *Transactions of the Association for Computational Linguistics*, volume 12, 2024.

[27] Xiaoran Liu, Hang Yan, Shuo Zhang, Chenxin An, Xipeng Qiu, and Dahua Lin. Scaling Laws of RoPE-based Extrapolation. *arXiv preprint arXiv:2310.05209*, 2023. URL https://arxiv.org/abs/2310.05209.

[28] Ekdeep Singh Lubana, Kyogo Kawaguchi, Robert P. Dick, and Hidenori Tanaka. A Percolation Model of Emergence: Analyzing Transformers Trained on a Formal Language. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2408.12578.

[29] Carles Marín Muñoz. Universal specialization-coordination structure in language model training, 2026. URL https://zenodo.org/records/19199838.

[30] Xin Men, Mingyu Xu, Bingning Wang, Qingyu Zhang, Hongyu Lin, Xianpei Han, and Weipeng Chen. Base of RoPE Bounds Context Length. *arXiv preprint arXiv:2405.14591*, 2024. URL https://arxiv.org/abs/2405.14591.

[31] Xin Men, Mingyu Xu, Qingyu Zhang, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. ShortGPT: Layers in Large Language Models are More Redundant Than You Expect. *arXiv preprint arXiv:2403.03853*, 2024. URL https://arxiv.org/abs/2403.03853.

[32] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2202.05262.

[33] Kai Nakaishi, Yoshihiko Nishikawa, and Koji Hukushima. Phase transition in large language models and the criticality of natural languages. *arXiv preprint arXiv:2406.05335*, 2024. URL https://arxiv.org/abs/2406.05335.

[34] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context Learning and Induction Heads. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

[35] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient Context Window Extension of Large Language Models. *arXiv preprint arXiv:2309.00071*, 2023. URL https://arxiv.org/abs/2309.00071.

[36] Cheng Kevin Qu, Andrew Ly, and Pulin Gong. Fractional neural attention for efficient multiscale sequence processing. *arXiv preprint arXiv:2511.10208*, 2025. URL https://arxiv.org/abs/2511.10208.

[37] Enrique Queipo-de-Llano, Álvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention Sinks and Compression Valleys in LLMs are Two Sides of the Same Coin. *arXiv preprint arXiv:2510.06477*, 2025. URL https://arxiv.org/abs/2510.06477.

[38] Shinsei Ryu and Tadashi Takayanagi. Holographic derivation of entanglement entropy from the anti–de Sitter space/Conformal Field Theory correspondence. *Physical Review Letters*, 96:181602, 2006. doi: 10.1103/PhysRevLett.96.181602.

[39] Ning Shang, Li Lyna Zhang, Siyuan Wang, Gaokai Zhang, Gilsinia Lopez, Fan Yang, Weizhu Chen, and Mao Yang. LongRoPE2: Near-Lossless LLM Context Window Scaling. *arXiv preprint arXiv:2502.20082*, 2025. URL https://arxiv.org/abs/2502.20082.

[40] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. *Neurocomputing*, 568, 2024. URL https://arxiv.org/abs/2104.09864.

[41] Youran Sun and Babak Haghighat. Phase Transitions in Large Language Models and the $O(N)$ Model. *arXiv preprint arXiv:2501.16241*, 2025. URL https://arxiv.org/abs/2501.16241.

[42] Lorenzo Tiberi, Francesca Mignacco, Kazuki Irie, and Haim Sompolinsky. Dissecting the Interplay of Attention Paths in a Statistical Mechanics Theory of Transformers. *arXiv preprint arXiv:2405.15926*, 2024. URL https://arxiv.org/abs/2405.15926.

[43] Felipe Urrutia, Jorge Salas, Alexander Kozachinskiy, Cristian Buc Calderon, Hector Pasten, and Cristobal Rojas. Decoupling Positional and Symbolic Attention Behavior in Transformers. *arXiv preprint arXiv:2511.11579*, 2025. URL https://arxiv.org/abs/2511.11579.

[44] Yifei Wang, Yueqi Wang, Zhenrui Yue, Huimin Zeng, Yong Wang, Ismini Lourentzou, Zhengzhong Tu, Xiangxiang Chu, and Julian McAuley. FASA: Frequency-aware Sparse Attention. *arXiv preprint arXiv:2602.03152*, 2026. URL https://arxiv.org/abs/2602.03152.

[45] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient Streaming Language Models with Attention Sinks. *arXiv preprint arXiv:2309.17453*, 2023. URL https://arxiv.org/abs/2309.17453.

[46] Jing Xiong, Liyang Fan, Hui Shen, Zunhai Su, Min Yang, Lingpeng Kong, and Ngai Wong. DoPE: Denoising Rotary Position Embedding. *arXiv preprint arXiv:2511.09146*, 2025. URL https://arxiv.org/abs/2511.09146.

[47] Bowen Yang, Bharat Venkitesh, Dwarak Talupuru, Hangyu Lin, David Cairuz, Phil Blunsom, and Acyr Locatelli. Rope to Nope and Back Again: A New Hybrid Attention Strategy. *arXiv preprint arXiv:2501.18795*, 2025. URL https://arxiv.org/abs/2501.18795.

[48] Menglin Yang, Harshit Verma, Delvin Ce Zhang, Jiahong Liu, Irwin King, and Rex Ying. Hypformer: Exploring Efficient Transformer Fully in Hyperbolic Space. In *Proceedings of KDD 2024*, 2024. URL https://arxiv.org/abs/2407.01290.

[49] Jiayi Yuan, Cameron Shinn, Kai Xu, Jingze Cui, George Klimiashvili, Guangxuan Xiao, Perkz Zheng, Bo Li, Yuxin Zhou, Zhouhai Ye, Weijie You, Tian Zheng, Dominic Brown, Pengbo Wang, Markus Hoehnerbach, Richard Cai, Julien Demouth, John D. Owens, Xia Hu, Song Han, Timmy Liu, and Huizi Mao. BLASST: Dynamic BLocked

Attention Sparsity via Softmax Thresholding. *arXiv preprint arXiv:2512.12087*, 2025. URL https://arxiv.org/abs/2512.12087.

[50] Edward Zhang. Attention's Gravitational Field: A Power-Law Interpretation of Positional Correlation. *arXiv preprint arXiv:2603.04805*, 2026. URL https://arxiv.org/abs/2603.04805.

[51] Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H_2O: Heavy-Hitter Oracle for Efficient Generative Inference of Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2306.14048.

[52] Meizhi Zhong, Chen Zhang, Yikun Lei, Xikai Liu, Yan Gao, Yao Hu, Kehai Chen, and Min Zhang. Understanding the RoPE Extensions of Long-Context LLMs: An Attention Perspective. *arXiv preprint arXiv:2406.13282*, 2024. URL https://arxiv.org/abs/2406.13282.

[53] Yuhao Zhou, Sirui Song, Boyang Liu, Zhiheng Xi, Senjie Jin, Xiaoran Fan, Zhihao Zhang, Wei Li, and Xuanjing Huang. EliteKV: Scalable KV Cache Compression via RoPE Frequency Selection and Joint Low-Rank Projection. *arXiv preprint arXiv:2503.01586*, 2025. URL https://arxiv.org/abs/2503.01586.

[54] George Kingsley Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA, 1949.