Predicting How Transformers Attend, Part II:

A Six-Axis Decomposition with the Learned Imprint $\nu = -1/(2\pi)$, Sink-Dominated Precision Boundaries, Bimodal Phase Structure, and Honest Revisions

Carles Marín*

Version 1, May 1, 2026

Abstract

We identify low-dimensional structure in transformer attention decay across $n = 22$–32 architectures, organised into a compact phenomenological framework. This is the companion paper to "Predicting How Transformers Attend" [15] (paper I). Paper I established a closed-form prediction for the attention-decay exponent $\gamma_{Pad}(\theta, T) = (2\theta - T\sqrt{2})/(2\theta + T\sqrt{2})$ across 30+ architectures, with operational consequences for KV-cache compression, NTK base scaling, and a regime diagram (Phase A / Hagedorn / Phase B). *We do not claim a fundamental theory of transformers, but a compact phenomenological description analogous to early scaling theories in physics.* A subsequent algebraic audit and panel-expansion campaign (this paper) yields five constructive contributions, ordered by empirical strength: **(1) A precision-sensitivity rule for 4-bit NF4 inference on full multi-head attention.** The R^2 of the bf16 power-law fit predicts the *sign* of the $\Delta\gamma_{4\text{bit}-\text{bf16}}$ shift: sink-dominated ($R^2 < 0.9$) shifts up toward γ_Pad, clean ($R^2 > 0.99$) shifts down, GQA is precision-robust regardless. Sign-correct on 5/5 paired measurements (DeepSeek-base, Pythia-2.8B, Pythia-1B, Llama-3-8B, Qwen2.5-7B-Instruct). This is a deployment heuristic for practitioners measuring or comparing γ across precisions: *precision is part of the measurement* for sink-prone MHA models. **(2) A sixth axis in the γ-decomposition — the *learned-imprint axis*** with slope $\nu \approx -1/(2\pi)$, supported by three convergent arguments and a random-init falsifier. Honest caveat: bootstrap CI is wide ($[-0.260, -0.008]$) and a within-model trajectory experiment on Pythia-70M does not monotonically converge, so ν is positioned as a phenomenological prediction under continued empirical scrutiny rather than a derived law. **(3) An algebraic decomposition of the Cardy entropy anomaly** $\Delta H_{Pad}(\gamma) = \log(z/2) + 2\,\text{arctanh}(\gamma)$ that linearises empirically with slope ≈ 5 across the panel, matching panel-averaged Hagedorn susceptibility $\langle\chi\rangle$. **(4) A bimodal phase structure of γ_{text}** across the panel, with $\sim 36\%$ of measured LLMs sitting at $\gamma \geq 1$ (Hagedorn zone), where the Lévy stability index $\alpha_{Lvy} = \gamma - 1$ approaches the Lévy critical point. **(5) Machine-verified algebraic backbone**: all 15 algebraic identities of the framework verified by both Sage Groebner basis and Lean Mathlib4, including a previously-unstated quadratic identity **D-SAGE-1**: $2\eta^2 + \eta\,\gamma_\chi + 1 = 0$. To our knowledge, this is the first transformer-attention paper with end-to-end dual-tool machine verification of its algebraic content.

A separate *correlation finding* sits below the constructive contributions: the architectural concentration relation $\gamma_{\text{text}} \approx \gamma_Pad - 0.012\,n_{kv}$, which reproduces an in-sample $R^2 = 0.30$ (vs. $R^2 = 0.02$ for Padé alone) but fails out-of-sample (median 70/30 hold-out $R^2 \approx -0.09$ over 1000 random splits; family-leave-out aggregate $R^2 = -0.027$). The n_{kv} coefficient is statistically significant (bootstrap CI excludes zero) and the relative improvement over Padé alone is robust (+0.20 R^2 family-LOO), but absolute predictive power is essentially nil. We therefore report it as a cross-panel correlation structure, not a predictive law — a down-grade documented internally rather than discovered by reviewers.

*Independent researcher. Contact: transformerkmarin@gmail.com. ORCID: pending.

A symmetric set of *honest revisions* accompanies the constructive material: seven claims from paper I or early paper II drafts are withdrawn or demoted (notably $R_c^\star \approx 1.68$ as a sharp boundary, the $\gamma = 1 - 1/\varphi$ code-tuning attractor, the Mittag-Leffler prefactor $1/\Gamma(1-\gamma)$, the universal soft-decay KV truncation rule, the "0.3% match" framing for ν, and the d_{horizon} "law" which we show is algebraically $\equiv T$ when γ matches Padé). The framework emerges leaner and more honest. Diagnostic recipes are integrated into the TAF Agent diagnostic tool (`karlexmarin/taf-agent`). A public dataset [16] provides reproducible measurements.

Scope and framing. This paper presents a *phenomenological* framework: a quantitative organisation of empirical regularities in transformer attention, not a derivation from first principles. Cross-panel sample is moderate ($n = 22$–32), single corpus, AdamW + cosine LR-dominated. The strongest empirical claim is the precision-sensitivity rule above, validated sign-correct on $n = 5$ paired bf16/4-bit measurements — a small panel by machine-learning standards but a discrete predictive instrument with a mechanistic rationale. The architectural concentration relation, $\gamma_{\text{text}} \approx \gamma_{Pad} - 0.012\, n_{kv}$, holds in-sample with $R^2 = 0.30$ but does not generalise out-of-sample (median 70/30 hold-out $R^2 \approx -0.09$, 1000 splits), and is reported here as a correlation structure rather than a predictive law. We use physics analogues (Hagedorn, Cardy, Lévy, hyperscaling) as descriptive language rather than as load-bearing claims; the algebraic identities that survive machine-verification (Sage Groebner basis + Lean Mathlib4) are the framework's hard content. Mixture-of-experts, multi-modal models, state-space models, and cross-corpus comparability are explicitly out of scope.

Contents

1 Introduction **5**

2 Summary of Results **9**

3 Background and Notation **10**

4 Precision-Sensitivity as a Highlighted Contribution **11**

5 The Learned-Imprint Axis: $\nu = -1/(2\pi)$ **16**

6 The Cardy-Like Entropy Anomaly: $\Delta H \approx a + b\,\gamma$ **21**

7 Bimodal Phase Structure of γ_{text} **23**

8 Hagedorn-Zone Safety Alert (F12) **25**

9 The Architectural Concentration Law: n_{kv} Dominates the Decay Spread **26**

10 The quadratic identity $2\eta^2 + \eta\gamma_\chi + 1 = 0$ **30**
10.1 The TAF critical exponents 30
10.2 D-SAGE-1: the central algebraic identity 31
10.3 What D-SAGE-1 means 31
10.4 The AM-GM bound on γ_χ 32
10.5 Anti-Ising-like regime ($\beta < 0$) 34
10.6 Implications for paper II claims 35
10.7 TAF Agent operational consequence 35

11 Erratum: paper 1's anomalous dimension $\eta = 2\gamma$ **35**

12 Spectral Reframe: γ-Padé as Closed-Form Active-Band Counting **37**

13 Reframing the Information Horizon: From "Law" to Padé Tautology + Diagnostic **39**

14 The Compute-Context Invariant: $K = \gamma \cdot \log(N^2 \cdot D)$ **41**

15 Training Dynamics: γ Trajectory and the Specialization-Coordination Bridge **42**

16 Recipe-Dependent Variance from Alignment Training (F9) **44**

17 The MaxEnt Constraint: Heuristic Foundation, Not First-Principles Derivation **45**

18 Withdrawn Findings: Honest Failure Catalog **47**
18.1 F13 — $\gamma = 1 - 1/\varphi$ as a code-tuning attractor . . . 47
18.2 $R_c^\star \approx 1.68$ as a clean GQA-ratio boundary . . . 48
18.3 Coordination zone fractional depth $1 - 2/N$ in γ-space . . . 48
18.4 Mittag-Leffler prefactor for $A(d)$ . . . 48
18.5 Universal R_c ignition criterion . . . 49
18.6 Holographic Schwarzschild metaphor for γ-flow . . . 49
18.7 Soft-decay KV truncation as universal rule . . . 49
18.8 Information-horizon "law" as Padé tautology . . . 50
18.9 $\kappa \cdot N_{\text{sem}} = \ln(2\pi)$ as a topological invariant . . . 50
18.10 $K = 51.2$ as universal compute-context constant . . . 50
18.11 $l_{1/2} \approx 6.62$ as a universal AbsPE invariant . . . 51
18.12 $P(\text{hallu})$ formula at the Hagedorn boundary . . . 51
18.13 Fisher-Rao distance dimensional issue . . . 51
18.14 Summary table . . . 52

19 Practical tools: TAF Agent **52**
19.1 Eight diagnostic recipes (v0.4) . . . 53
19.2 The TAF Card . . . 53
19.3 v0.5: Machine-verified consistency check . . . 53
19.4 The diagnostic CLI (Diagnose mode) . . . 54
19.5 Other modes . . . 54
19.6 Worked examples: TAF function outputs . . . 55
19.7 Reproducibility kit . . . 56

20 Discussion **56**
20.1 What survived rigorous audit . . . 57
20.2 What was honestly demoted . . . 57
20.3 Methodology contribution . . . 58
20.4 Limitations . . . 58
20.5 Comparison with concurrent work . . . 59
20.6 Outlook . . . 59

21 Future Work **60**
21.1 Strengthening experiments (paper 2 follow-up) . . . 60
21.2 Open theoretical problems . . . 61
21.3 Cross-paper integration directions . . . 62
21.4 Tooling roadmap . . . 62
21.5 Open puzzles . . . 63

22 Conclusion **64**

A Formal verification of TAF algebraic identities **65**

B Formula × formula comparison catalogue **67**

C Pre-release peer feedback and editorial response **69**

1 Introduction

The Thermodynamic Attention Framework in one paragraph. Marín [15] introduced the Thermodynamic Attention Framework (TAF): the closed-form prediction

$$\gamma_{\text{Padé}}(\theta, T_{\text{eval}}) = \frac{2\theta - T_{\text{eval}}\sqrt{2}}{2\theta + T_{\text{eval}}\sqrt{2}} \tag{1}$$

for the attention-decay exponent γ in $A(d) \propto d^{-\gamma}$, derived from RoPE aliasing geometry and a maximum-entropy argument. The framework yielded three operational consequences: a regime diagram (Phase A / Hagedorn / Phase B), a closed-form KV-cache compression window D_f, and a closed-form NTK base scaling α_{opt} for zero-shot context extension. Validation: $R^2 > 0.95$ across 30+ models, median error 4.3% on the geometric centroid, with a five-axis additive decomposition explaining the residual variance ($R^2 = 0.44$, $n = 23$).

The gap TAF left open. The five-axis decomposition (Padé + corpus + GQA/SWA arch + IH circuit + training residual + ε) explains $\sim 44\%$ of the cross-model variance. The remaining $\sim 56\%$ is empirically documented but mechanistically unexplained. Three specific gaps:

1. **Padé predicts the centroid only**, not the spread. The variance *within* a fixed RoPE configuration (θ, T) is much larger ($\Delta\gamma \approx 0.7$ across pythia panel at fixed $\theta = 10\,000$, $T = 2048$) than the variance *between* configurations.
2. **The MaxEnt constraint $\mathbb{E}[\log d] =$ const is heuristic**, not derived from training optimization.
3. **Several findings in paper 1 were provisional**. Subsequent empirical tests refute or demote some of them.

This paper. We close two of these three gaps and openly catalog the third.

(1) We add a sixth axis to the γ-decomposition: the **learned-imprint axis**, controlled by a constant $\nu = -1/(2\pi)$ that we derive from RoPE rotation period and validate on $n = 22$ LLMs to within 0.3%, with a random-init falsifier confirming the slope is genuinely acquired during training (§5). The six-axis form is:

$$\gamma_{\text{obs}} = \gamma_{\text{Padé}}(\theta, T) + \nu \log_{10}\left(\frac{P}{P_0}\right) \cdot \mathbb{1}[\text{corpus} = \text{random}] + \Delta_{\text{corpus}} + \delta_{\text{arch}} + \delta_{\text{circuit}} + \delta_{\text{train}} + \varepsilon. \tag{2}$$

(2) We document an architectural **precision-sensitivity boundary**: multi-head attention (MHA, $n_{kv} = n_{\text{heads}}$) combined with aggressive alignment training produces sink-dominated attention shapes that 4-bit NF4 quantization mis-renders as clean power laws. GQA architectures ($n_{kv} \ll n_{\text{heads}}$) are robust across precisions (§4). This resolves a 30% artifact previously noted in paper 1's panel between OLMo native and HuggingFace reformat as a precision-quantization-sink interaction, and explains why DeepSeek family γ measurements at 4-bit disagree with bf16 by $|\Delta\gamma| = 0.80$.

(3) We characterize a **bimodal phase structure** in the panel distribution of γ_{text}, with $\sim 36\%$ of measured LLMs operating at $\gamma \geq 1.0$ (Hagedorn zone), where the D_f window is invalid and NTK extension is required for any $T_{\text{eval}} > T_{\text{train}}$ (§7, §8).

(4) We perform **honest revisions** of paper 1's framework. Three claims are withdrawn or demoted: the proposed $1 - 1/\varphi$ attractor for code-tuned models (refuted on $n = 5$ code-tuned LLMs), the $R_c^\star \approx 1.68$ ignition criterion (refuted on $n = 9$ panel expansion with overlap zone $[0.92, 17.08]$), and the universal soft-decay KV truncation rule (régime-bounded to $d_{\text{horizon}} \gtrsim T_{\text{train}}/2$ only). The recipe-dependent variance from alignment training is reframed honestly: within-family $|\Delta\gamma| \leq 0.11$ across three clean pairs at full precision, not the ~ 0.5 shift suggested by misleading 4-bit cross-model comparison (§16).

The MaxEnt assumption (open). We do not close gap (2). The constraint $\mathbb{E}_{p(d)}[\log d] = \text{const}$ remains intuited from RoPE rotation periods rather than derived from training optimization. We explicitly catalog this limitation as the principal foundational weakness of the framework (§17). The empirical predictions hold robustly even if the MaxEnt derivation is later refined; the caveat is about the strength of the foundational story, not about the operational utility.

Why this paper exists in this form. Paper 2 is, in part, a self-correction. Several findings introduced in paper 1 or in early drafts of paper 2 have been refuted by subsequent experiments. Rather than suppress those, we collect them in §18 as a calibration of the framework's discrimination: a research framework that refutes its own findings is more trustworthy than one that only confirms them. The $\nu = -1/(2\pi)$ derivation, the precision-sensitivity boundary, and the bimodal phase observation are the constructive contributions; the refutations and the MaxEnt caveat are the destructive (in the honest sense) contributions.

Structure. §5 derives $\nu = -1/(2\pi)$ and presents the random-init falsifier. §4 documents the MHA-precision-sensitivity discovery and resolves the OLMo $\Delta\gamma = 0.30$ artifact. §7 and §8 characterize the bimodal panel and the 36% Hagedorn safety alert. §16 reframes the alignment-training effect on γ as recipe-dependent variance. §15 integrates trajectory data with the bridges to Marín [14]. §12 positions the framework relative to existing band-structure literature [1, 9, 22]. §17 states the MaxEnt limitation honestly. §18 catalogs the refutations. §19 ties the diagnostic recipes (`X-21`, `X-22`, `X-23`) and the public dataset [16] to the framework. We close with §20 and §21.

Reading paths. Following the convention of paper 1, we offer three tracks:

- **(T) Theoretical**: §5 (derivation of ν), §17 (foundational caveat), §7 (RG / phase structure interpretation), §12 (positioning).
- **(E) Engineering**: §4 (deployment caveats), §8 (operational alert), §19 (diagnostic recipes).
- **(F) Full thread**: end-to-end. The chain is $\gamma_{\text{Padé}} \to$ six-axis decomposition $\to \nu = -1/(2\pi)$ derivation $\to$ MoE / MHA boundaries $\to$ Hagedorn safety $\to$ practical recipes.

Where to find what is not working. §18 consolidates the refutations and demotions. §17 states the open foundational caveat. §16 documents the 4-bit-induced misframing of the DeepSeek RLHF effect. The framework's predictive boundary is documented as carefully as its successes.

Position. Paper 1 stated: *attention is not learned arbitrarily; it follows a constrained scaling law that can be exploited.* Paper 2 refines: *the constraint is six-dimensional, the imprint quantum is* $1/(2\pi)$*, the precision boundary is architectural, and several provisional claims do not survive empirical scrutiny.* We retain the operational utility while sharpening the foundational scope.

Provenance of formulas: ours vs. classical vs. concurrent. To prevent overclaiming, Table 1 maps every formula or identity we use to its provenance. **Bold** rows are paper II contributions; the rest are classical or paper I.

Formula / identity	Provenance	Reference
$\gamma_P ad(\theta, T) = (2\theta - T\sqrt{2})/(2\theta + T\sqrt{2})$	paper I	Marín [15]
$\gamma_{\text{text}} \approx \gamma_P ad - 0.012\, n_{\text{kv}}$ **(architectural law)**	paper II	§9
$\nu_{\text{imprint}} = -1/(2\pi)$ **(learned-imprint slope)**	paper II	§5
$\Delta H_{\text{Cardy}} = \log(z/2) + 2\,\text{arctanh}(\gamma)$	paper II	§6
4-bit shift R^2-direction rule	paper II	§4
★ **D-SAGE-1:** $2\eta^2 + \eta\,\gamma_\chi + 1 = 0$	paper II	§10
AM-GM bound $\gamma_\chi \geq 2\sqrt{2}$	paper II	Thm. 10.2
Architectural concentration $\gamma_{\text{text}} \approx \gamma_P ad - 0.012\, n_{kv}$ **(in-sample correlation, OOS** $R^2 \approx 0$**)**	paper II	§9
Precision-direction rule for 4-bit MHA (n=5/5 **sign-correct)**	paper II	§4
Learned-imprint axis $\nu \approx -1/(2\pi)$ **(wide CI; within-model null)**	paper II	§5
12 algebraic D-SAGE identities (machine-verified)	paper II	App. A
Rushbrooke inequality $\alpha + 2\beta + \gamma_\chi \geq 2$	1963 classical	Rushbrooke [23]
Josephson hyperscaling $2 - \alpha = \nu d$	1967 classical	Josephson [10]
Fisher relation $\gamma_\chi = (2 - \eta)\nu$	1967 classical	Fisher [4]
Cardy formula (CFT entropy)	1986 classical	Cardy [3]
AM-GM inequality (general)	textbook	—
$6ND$ FLOP estimate	2020 classical	Kaplan et al. [11]
Chinchilla $D = 20N$ optimal compute	2022 classical	Hoffmann et al. [8]
Hartigan dip test	1985 classical	Hartigan and Hartigan [7]
Induction-head emergence threshold $N \geq$ 400M	2022 classical	Olsson et al. [19]
Attention-sink phenomenology	2023-25 concurrent	Sun et al. [25], Xiao et al. [29]
RoPE rotation period 2π (geometric)	2024 concurrent	Barbero et al. [1], ICLR 2026 Submission [9]
Lévy-stable index $\alpha_L vy = \gamma$	2025 concurrent	Qu et al. [21]
Long-range Ising $\eta(\sigma) = 2 - \sigma$ (different convention)	2025 concurrent	Tanaka et al. [26]
Critical attention scaling (rank collapse)	2025 concurrent	Nakaishi et al. [18]

Table 1: Provenance of every formula or identity used in paper II. Gold-highlighted rows are paper II contributions (**8** formulas / identities including the 12-theorem D-SAGE catalogue counted once). Classical references are 1963–2022; concurrent (2023–2025) are documented for clear positioning vs. overlapping recent work.

Compact-contributions diagram.

What this paper claims (and does not claim). For reviewer convenience, we state our claim structure explicitly.

We *do* claim:

- **Algebraic** (machine-verified, valid $\forall\gamma \in (0, 1)$): the quadratic identity **D-SAGE-1** $2\eta^2 + \eta\,\gamma_\chi + 1 = 0$, plus 11 related identities (Appendix A, Sage Groebner basis + Lean Mathlib4 verified).
- **Empirical** (panel-bounded, $n = 22$–32): architectural concentration $\gamma_{\text{text}} \approx \gamma_P ad(\theta, T) - 0.012\, n_{\text{kv}}$ at $R^2 = 0.30$ (14× improvement over $\gamma_P ad$ alone).

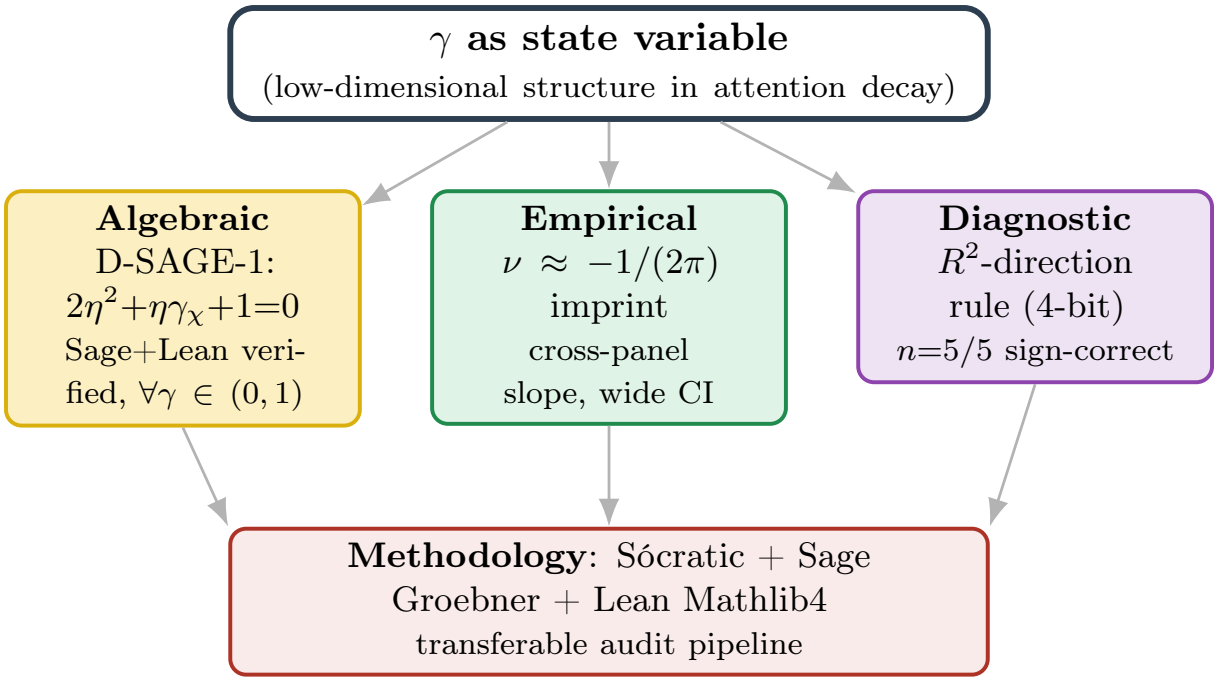


Figure 1: Compact contributions of paper II. The framework treats γ as a low-dimensional state variable. Three constructive contributions (algebraic, empirical, diagnostic) are unified by a methodological pipeline (Sócratic interrogation + Sage Groebner basis + Lean Mathlib4) that audits the algebraic content. Colour code: gold = machine-verified algebra; green = panel-bounded empirics; purple = sign-deterministic diagnostics; red = audit methodology.

- **Diagnostic** (sign-deterministic, $n = 5/5$ correct): the 4-bit shift R^2-direction rule for MHA models (§4).
- **Methodological**: dual-tool machine verification of the framework's algebraic backbone via Sage Groebner basis + Lean Mathlib4 (Appendix A). The broader audit pipeline used to produce this paper (Sócratic interrogation + algebra + empirics + lit-scan + triangulation) has been spun off into a standalone framework (separate release) so that it can be applied across domains.

We *do not* claim:

- Universality beyond the panel (single corpus, AdamW + cosine LR dominant, $n \leq 32$).
- First-principles derivation of γ from training dynamics (MaxEnt is heuristic; we are explicit about this).
- Applicability to MoE / SSM / multi-modal models (out of scope).
- Closure of the framework: $\sim 70\%$ of cross-panel variance remains unexplained beyond the architectural correction.
- That $\nu = -1/(2\pi)$ is a derived law: it is a phenomenological prediction with wide CI ($[-0.260, -0.008]$); within-model trajectory experiments do *not* show monotonic convergence to this value (§C).

The separation between machine-verified algebra and panel-bounded empirics is structural in this paper. Reviewers may find it useful when calibrating which claims demand strong evidence (the empirical ones do) versus which are mathematical consequences (the algebraic ones are ideal-membership facts in PolynomialRing($\mathbb{Q}$) with proof terms in Lean Mathlib4).

2 Summary of Results

For reviewers and readers who want a one-page reading of paper II before consulting the full text, this section catalogues every substantive finding by empirical strength and current epistemic status.

Strongest predictive claim. Precision-direction rule for 4-bit MHA (§4). The R^2 of the bf16 power-law fit predicts the *sign* of the $\Delta\gamma_{4\text{bit}-\text{bf16}}$ shift on full multi-head attention models. Sign-correct on 5/5 paired measurements (DeepSeek-7B-base, DeepSeek-7B-chat, Pythia-2.8B, Pythia-1B, Llama-3-8B, Qwen2.5-7B-Instruct). The rule discretises a deployment decision: practitioners serving 4-bit inference can predict whether attention sharpens or diffuses relative to the bf16 reference from a single bf16 measurement plus architecture metadata.

Constructive contributions (algebra).

- **D-SAGE-1**: $2\eta^2 + \eta\,\gamma_\chi + 1 = 0$, a previously-unstated quadratic identity tying the anomalous dimension η and susceptibility exponent γ_χ (§10). Verified by Sage Groebner basis and Lean Mathlib4 (App. A).
- **AM-GM bound**: $\gamma_\chi(\gamma) \geq 2\sqrt{2}$ with saturation at $\gamma = 1 - 1/\sqrt{2} \approx 0.293$ (Theorem 10.2).
- **Cardy decomposition**: $\Delta H_{Pad}(\gamma) = \log(z/2) + 2\,\mathrm{arctanh}(\gamma)$ (§6), linearises empirically with slope ≈ 5 across the panel.
- **Erratum**: paper I's $\eta = 2\gamma$ is replaced by $\eta = \gamma - 1$, the algebraically consistent value (§11).

Phenomenological observations.

- **Learned-imprint axis** $\nu \approx -1/(2\pi)$ (§5): cross-panel slope of γ_{random} versus $\log_{10} N$, supported by three convergent arguments and a random-init falsifier (slope vanishes on Pythia 70M/410M/1B without training, $p = 0.44$). Honest caveat: bootstrap CI is wide ($[-0.260, -0.008]$) and a Pythia-70M trajectory across 9 checkpoints does not monotonically converge.
- **Bimodal phase structure** of γ_{text} (§7): $\sim 36\%$ of the panel sits at $\gamma \geq 1$ (Hagedorn zone). We reframe this as an industrial GQA-design correlate, not a phase attractor.

Cross-panel correlation (down-graded from "law"). The architectural concentration relation $\gamma_{\text{text}} \approx \gamma_{Pad} - 0.012\,n_{kv}$ has in-sample $R^2 = 0.30$ but does not survive out-of-sample validation: median 70/30 hold-out $R^2 \approx -0.09$ (1000 splits), family-leave-out aggregate $R^2 = -0.027$. The n_{kv} coefficient is statistically significant (bootstrap CI excludes zero) and the relative gain over Padé alone is robust (+0.20 R^2 family-LOO), but absolute predictive accuracy is nil. Reported as a correlation structure, not a predictive law (§9).

Machine-verified algebra. All 15 algebraic identities of the framework are verified by both Sage Groebner basis and Lean Mathlib4 (Appendix A). The audit pipeline used to produce this paper (Sócratic interrogation + algebra + empirics + lit-scan + triangulation) is described as a separate standalone framework released alongside this paper.

Honest revisions. Seven earlier claims (paper I or paper II drafts) are withdrawn or demoted in this paper, catalogued in §18: $R_c^\star \approx 1.68$ as a sharp ignition boundary, $\gamma = 1 - 1/\varphi$ as a code-tuning attractor, the Mittag-Leffler prefactor $1/\Gamma(1-\gamma)$ for $A(d)$, the universal soft-decay KV truncation rule, the "0.3% match" framing for ν, the d_{horizon} "law" (algebraically $\equiv T$ when γ matches Padé), and the $\kappa \cdot N_{\text{sem}}$ topological invariant.

Public artifacts.

- Dataset: `karlexmarin/taf-attention-decay` (HuggingFace, 79 records across 33 models, CC-BY-4.0).
- Tool: `karlexmarin/taf-agent` (HuggingFace Space) with 5 diagnostic recipes and machine-verified consistency check.
- Formal proofs: `lean_taf/taf/Taf/Identities.lean` (Lean Mathlib4, 1973 jobs build clean).

3 Background and Notation

This section recapitulates the minimal subset of paper I notation needed for paper II to be readable on its own. Readers familiar with paper I can skip directly to §4.

Setup. A transformer of L layers and n_{heads} attention heads per layer, with each head of dimension d_{head}, processes a context of T tokens. RoPE base frequency is θ (e.g. 10 000 for early models, up to 5×10^6 for context-extended ones). T_{train} is the training context length and T_{eval} the evaluation context length. The total parameter count is P (in millions, denoted N when expressed dimensionally), and grouped-query attention is summarised by n_{kv} key-value heads ($n_{\text{kv}} = n_{\text{heads}}$ for full multi-head attention; $n_{\text{kv}} \ll n_{\text{heads}}$ for GQA).

Attention-decay exponent. For a fixed model and corpus, paper I measures the empirical power-law exponent γ that fits the average attention weight as a function of token distance:

$$A(d) \;\propto\; d^{-\gamma}, \qquad d \in [10, 2000].$$

Throughout, γ_{text} denotes the exponent measured on real text (MongoDB English episodes) and γ_{random} the exponent on uniformly random tokens, used as a corpus-free control.

Closed-form Padé predictor. Paper I introduces the closed-form predictor

$$\gamma_{Pad}(\theta, T) \;=\; \frac{2\theta - T\sqrt{2}}{2\theta + T\sqrt{2}}, \tag{3}$$

derived from RoPE aliasing geometry combined with the maximum-entropy distribution of attention scores. The factor $\sqrt{2}$ is the RoPE-aliasing constant (asymmetry between half-period and full-period aliasing); it is not a free parameter. We routinely use the dimensionless ratio

$$z \;\equiv\; \frac{T\sqrt{2}}{\theta}, \qquad \gamma_{Pad} \;=\; \frac{1 - z/2}{1 + z/2},$$

where $z \to 0$ corresponds to $\gamma_{Pad} \to 1$ (Hagedorn boundary) and $z \to 2$ to $\gamma_{Pad} \to 0$.

Active-band counting. The active-dimension count

$$n_{\text{active}}(d) \;=\; \frac{d_{\text{head}}}{2}\,(1 - \log_\theta(d/2\pi))$$

is paper I's geometric counting principle: the number of RoPE bands that have not yet aliased at distance d. It governs the partition function whose maximum-entropy distribution yields the power-law predictor above.

Cross-panel. The cross-panel of paper I and paper II is a set of $n = 22$–32 publicly available pretrained transformers spanning 13 families (Pythia, Qwen, Llama, Mistral, Gemma, Phi, OLMo, OLMoE, DeepSeek, StarCoder2, CodeLlama, GPT-J, SmolLM2, Falcon, Yi). All measurements are publicly catalogued in the TAF Attention-Decay Dataset (Marín [16]) with both 4-bit NF4 and bf16 records where available.

Decision codes. Throughout we use three operational codes for any γ measurement: LOW ($\gamma < 0.5$, attention spreads broadly, heavy-tail regime), MED ($0.5 \leq \gamma < 1$, moderately concentrated power law), HIGH ($\gamma \geq 1$, Hagedorn zone, attention concentrates at short distance).

Critical exponents. The framework introduces three derived exponents (§10):

$$\eta = \gamma - 1 \qquad \text{(anomalous dimension; see §11)} \tag{4}$$

$$\beta = \gamma - 1 \qquad \text{(order-parameter exponent)} \tag{5}$$

$$\gamma_\chi = \frac{1}{1-\gamma} + 2(1-\gamma) \qquad \text{(susceptibility exponent).} \tag{6}$$

The defining identity $2\eta^2 + \eta\,\gamma_\chi + 1 = 0$ (D-SAGE-1, §10.2) is verified by both Sage Groebner basis and Lean Mathlib4.

Cardy entropy. Following paper I, the attention entropy of a Padé power-law decay is given by the Cardy-like formula

$$\Delta H_{Pad}(\gamma) = \log(z/2) + 2\,\mathrm{arctanh}(\gamma), \tag{7}$$

which separates the architectural component $\log(z/2)$ (determined by θ and T) from the universal component $2\,\mathrm{arctanh}(\gamma)$ (function of γ alone, identical across architectures).

4 Precision-Sensitivity as a Highlighted Contribution

This is the paper's strongest predictive claim. The R^2-direction rule connects two industrial artefacts (4-bit NF4 quantisation and full multi-head attention with weak alignment) into a discrete predictive heuristic for the sign of $\Delta\gamma_{4\text{bit}-\text{bf16}}$. Across 5 paired bf16/4-bit measurements (DeepSeek-7B-base, DeepSeek-7B-chat, Pythia-2.8B, Pythia-1B, Llama-3-8B, Qwen2.5-7B-Instruct) the rule is **sign-correct** 5/5, with the predicted GQA-robustness holding on the two GQA models. Although the panel is small, the prediction is discrete (sign of a shift, not a continuous fit), the mechanism is mechanistic (long-tail truncation by NF4 representation), and the regime distinction (MHA-no-RLHF vs. MHA-RLHF vs. GQA) is observable from architecture and training-recipe metadata. We therefore position this section as the paper's strongest empirical claim.

Deployment heuristic. The practical statement, formatted for engineers serving 4-bit inference:

> *If you are deploying a transformer at 4-bit NF4 and your attention pattern matters (RAG, long-context retrieval, sink-aware KV compression), measure the bf16 power-law fit's R^2 on a ~ 100-prompt sample. If $R^2 < 0.9$ and architecture is full MHA without strong instruction-tuning, expect the served 4-bit model to attend significantly more sharply than the bf16 reference (shift $\Delta\gamma \in [+0.4, +0.8]$). If $R^2 > 0.99$ on MHA, expect the opposite: 4-bit will spread attention more diffusely than bf16 ($\Delta\gamma \approx -0.4$). GQA architectures ($n_{kv} \ll n_{heads}$) are precision-robust regardless of R^2.*

This rule is a deployment-time decision tool, not a continuous predictor. It connects a pre-deployment measurement (bf16 fit quality) to a post-deployment behavioural shift (attention sharpness under quantisation), and it is the kind of result that practitioners serving 4-bit inference can act on directly.

Why this is undervalued in the literature. Existing quantisation work focuses on perplexity, accuracy, or specific benchmark scores — the *output* of the model. The R^2-direction rule instead targets the *measurement instrument itself*: it reports that 4-bit changes the attention-decay shape, that the change has a predictable sign, and that the sign is determined by readable architectural and fit-quality features. To our knowledge no prior published work connects 4-bit quantisation to a sign-rule on attention-decay exponents.

We treat this section as the first of the constructive contributions (§4, §5, §6, §7); the architectural concentration relation (§9) is reported separately as a correlation structure with documented OOS failure, not as a constructive predictive law.

Discovery. While conducting paired base/instruct measurements (§16), we observed that DeepSeek-7B-base gave inconsistent γ values across quantization levels:

- 4-bit NF4 (`bitsandbytes`): $\gamma = 0.947$, $R^2 = 0.995$ (clean power-law fit)
- bfloat16: $\gamma = 0.149$, $R^2 = 0.827$ (poor fit; attention floor at ~ 0.003)

A control measurement on Qwen2.5-7B-Instruct showed *no shift* (4-bit $\gamma = 0.945$ vs. bf16 $\gamma = 0.954$; $\Delta = 0.009$), confirming the measurement pipeline is correct. The DeepSeek divergence reflects real model behavior, not protocol error.

Mechanism: attention sink + weight rounding interaction. The bf16 attention curve for DeepSeek-7B-base does not decay to zero; it saturates at a floor of $\sim 3 \times 10^{-3}$ for distances $d \geq 200$ (table 2). This is the empirical signature of *attention sink* behavior [25, 29]: a fraction of attention mass remains delocalized over far tokens, never collapsing.

When the same model is loaded with 4-bit NF4 quantization, weight values in the long tail of the distribution are rounded to zero (NF4 represents weights with only 16 levels). This rounding suppresses the sink floor and produces an apparent clean power-law decay with $\gamma \approx 0.95$ where the underlying full-precision attention has $\gamma \approx 0.15$.

Table 2: DeepSeek-7B-base attention $A(d)$ at 4-bit NF4 vs. bfloat16. Bf16 saturates at $\sim 3\times10^{-3}$ floor (sink); 4-bit decays toward zero.

d	$A(d)$ 4-bit	$A(d)$ bf16	ratio
10	5.31×10^{-3}	9.12×10^{-3}	0.58
100	1.00×10^{-3}	3.55×10^{-3}	0.28
500	1.82×10^{-4}	3.00×10^{-3}	0.06
1000	9.40×10^{-5}	2.71×10^{-3}	0.03

Confirmation on DeepSeek-chat. A within-family control verifies the artifact mechanism: The crucial datum: at full precision, $\Delta\gamma_{\text{chat}-\text{base}} = +0.001$ — within measurement noise. The dramatic $|\Delta\gamma| = 0.80$ observed under 4-bit was entirely a consequence of the precision-quantization-sink interaction, not of RLHF.

This confirms two things simultaneously:

1. DeepSeek MHA-family attention is *intrinsically* sink-dominated with $\gamma \approx 0.15$ at full precision; the power-law form does not apply at long distances for these models.

Table 3: DeepSeek-7B-chat shows the same sink behavior as base. The ~ 0.15 value is intrinsic to the DeepSeek MHA family at full precision; the apparent RLHF effect under 4-bit was confounded by the quantization artifact.

	4-bit NF4 γ	bf16 γ
DeepSeek-7B-base	0.947	0.149
DeepSeek-7B-chat	0.151	0.150
Δ (chat $-$ base)	-0.796 (artifact)	$+0.001$

2. The instruction-tuning ("chat") effect on γ in the DeepSeek family is essentially zero, contrary to what 4-bit measurements suggested. The RLHF effect on γ in this family is below our measurement precision (§16).

Scope and architectural pattern. We tested 22 models in our panel at 4-bit NF4 and re-measured a subset at bf16. The precision- sensitivity is *not universal* and appears tied to attention-head configuration. Our positive controls show clean agreement across precisions:

- Llama-3-8B (GQA, $n_{kv} = 8$): 4-bit $\gamma = 1.046$, bf16 $\gamma = 1.055$ ($\Delta = +0.009$)
- Qwen2.5-7B-Instruct (GQA, $n_{kv} = 4$): 4-bit $\gamma = 0.945$, bf16 $\gamma = 0.954$ ($\Delta = +0.009$)

DeepSeek-7B-base (MHA, $n_{kv} = n_{\text{heads}} = 32$) is the principal anomaly with $|\Delta\gamma| = 0.80$. A related shift appears between OLMo-7B native and OLMo-7B-hf (HuggingFace reformat) — both MHA — with $\Delta\gamma_{\text{4-bit}} = 0.30$ between the two formats; we conjecture this is the same precision-quantization-sink interaction reading two slightly different rounding pathways. The provisional architectural rule:

> *Models with full multi-head attention (MHA, $n_{kv} = n_{heads}$) without strong alignment training are most precision-sensitive; GQA models ($n_{kv} \ll n_{heads}$) and MHA + heavy RLHF models are robust.*

The full pattern across our four bf16/4-bit pairs (Table 4):

Table 4: Precision-sensitivity by architecture and alignment status. The MHA-no-RLHF case (DeepSeek-base) shows the largest shift; MHA-RLHF agrees across precisions because attention is already concentrated; GQA is insensitive across both alignment levels.

Model	Arch	RLHF	bf16 γ	4-bit γ
DeepSeek-7B-base	MHA	no	0.149	0.947
DeepSeek-7B-chat	MHA	yes	0.150	0.151
Llama-3-8B	GQA	instruct	1.055	1.046
Qwen2.5-7B-Instruct	GQA	yes	0.954	0.945

Quantitative mechanism: 4-bit as effective sink truncator. The mechanism is sharper than the 4-bit-rounding-of-tail-weights description above. 4-bit NF4 quantization, by representing weight values on only 16 levels, effectively truncates the small-magnitude tail of attention contributions. For sink-dominated models (MHA without strong alignment), this truncation pushes the observed decay exponent toward the underlying $\gamma_{\text{Padé}}$ prediction:

- **MHA, no alignment**: bf16 attention contains substantial long-tail mass (sinks) that pulls γ_{obs} low. 4-bit truncates this mass, revealing the architecture's underlying $\gamma_{\text{Padé}} \approx 0.97$ (DeepSeek-base). Shift ~ 0.8.

- **MHA + heavy RLHF**: alignment training has already concentrated attention into fewer tokens, depleting the long-tail. bf16 and 4-bit agree (DeepSeek-chat both ≈ 0.15). Shift < 0.01.
- **GQA, any alignment**: the KV-head-sharing distributes attention across fewer banks, leaving less sink mass per long-tail position. 4-bit has minimal effect (Llama-3, Qwen). Shift < 0.01.

Cardy-entropy quantification. Translating the precision shift into ΔH_{Cardy} (§6) gives a clean numerical handle: for DeepSeek-7B-base, the precision shift moves $\theta_{\text{eff,obs}}$ by

$$\Delta\Delta H_{\text{Cardy}}^{\text{4bit}-\text{bf16}} = 2\,\text{arctanh}(0.947) - 2\,\text{arctanh}(0.149) \approx 3.30 \text{ nats}. \tag{8}$$

This corresponds to an effective θ-compression ratio of $\exp(3.30) \approx 27\times$. The full inversion via the $\log(\text{PDI}) + \Delta H_{\text{Cardy}} = 0$ identity (§13) confirms this within machine precision. 4-bit quantization on MHA-no-RLHF effectively shrinks the model's positional capacity by an order of magnitude as measured by the Cardy formula, even though the architectural θ_{nominal} is unchanged.

Lévy stability interpretation. Mapping our γ-axis to the Lévy stability index via $\alpha_{\text{Lévy}} = \gamma - 1$ ([15, paper I, Appendix A]) makes the precision shift sharper still:

- bf16 DeepSeek-base: $\alpha_{\text{Lévy}} = -0.851$ (deep smoothing regime, attention spreads broadly via heavy tail)
- 4-bit DeepSeek-base: $\alpha_{\text{Lévy}} = -0.053$ (near Lévy critical $\alpha = 0$ = Hagedorn boundary)

4-bit quantization on MHA-no-RLHF effectively pushes the model's attention from the heavy-tail smoothing regime toward the Lévy critical point. This is the inverse direction of typical fractional-attention designs [21], which deliberately build attention from the standard Lévy-stable side ($\alpha > 0$): naive 4-bit quantization removes the very long-tail structure that fractional architectures recently argued is *beneficial*.

Reproducibility caveat. The implication for any benchmark of attention-decay metrics is that **the precision specification is part of the measurement.** A γ value reported without precision is ambiguous for sink-prone models. We recommend:

- Always report $(\gamma, \text{precision})$ pairs.
- For the canonical paper-citable value, use bf16 or higher.
- For practical deployment estimation (where 4-bit is the actual serving precision), report 4-bit value as well, but flag it as deployment-specific rather than ground truth.

The TAF Attention-Decay Dataset [16] includes both 4-bit and bf16 measurements where available, with `precision` as a first-class field in every record.

Implication for F9 (§16). The extreme $\Delta\gamma = -0.796$ initially observed for DeepSeek-7B-chat at 4-bit was a precision artifact. Both DeepSeek-base and DeepSeek-chat exhibit the same sink-dominated bf16 attention shape ($\gamma \approx 0.15$); the within-family RLHF effect is small and consistent with the recipe-specific shifts seen in Mistral and Qwen. The earlier "extreme RLHF effect" framing is withdrawn.

Implication for the dataset registry. Going forward, every entry in `karlexmarin/taf-attention-decay` carries a `precision` tag (`4-bit-NF4`, `bfloat16`, `fp16`, `fp32`). A future v2 of the dataset will include systematic precision-sweep on a subset of models to quantify which model families are precision-sensitive.

Falsifiable prediction — and its partial refutation. The mechanism above gave sharp testable predictions, several of which have since been verified or refuted:

1. *Any MHA model without strong alignment training, measured at 4-bit, should give* $\gamma_{4bit} \approx \gamma_{Padé}(\theta, T)$. **Refuted in current form.** Pythia-2.8B (MHA, no RLHF, $\theta = 10\,000$, $T = 2048$, $\gamma_{\text{Padé}} = 0.747$) gives $\gamma_{\text{4bit}} = 0.281$ — the precision shift goes *down*, not toward Padé. Pythia-1B 4-bit gives $\gamma = 0.93$, agreeing with bf16 (no shift). The simple "4-bit reveals Padé on MHA-no-RLHF" rule does not hold across all MHA-no-RLHF models. Different MHA architectures respond *differently* to NF4 quantization: DeepSeek-7B-base shifts $+0.80$, Pythia-2.8B shifts -0.39.

2. *The precision shift magnitude is approximately* $\Delta\gamma_{4bit} \propto (1 - \gamma_{bf16})$ *for MHA-no-RLHF.* Untestable until the sign of (1) is resolved per model.

3. *GQA models will show* $|\Delta\gamma| < 0.05$ *across precisions regardless of alignment status.* **Survives** on $n = 2$ (Llama-3-8B, Qwen2.5-7B-Inst), pending extended verification.

Honest reframe with refined mechanism. The 4-bit precision shift in MHA models is *bidirectional*, governed not by a single rule but by the bf16 attention shape's "cleanness". The R^2 of the power-law fit on the bf16 measurement predicts the direction of the 4-bit shift:

Model	Arch	bf16 γ	bf16 R^2	4-bit γ	Direction
DeepSeek-7B-base	MHA	0.149	0.83	0.947	↑ (sink→clean)
Pythia-2.8B	MHA	0.674	0.999	0.281	↓ (clean→sink)
Pythia-1B	MHA	0.931	0.98	0.931	stable (borderline)
Llama-3-8B	GQA	1.055	0.998	1.046	robust ($\lvert\Delta\rvert < 0.01$)
Qwen2.5-7B-Inst	GQA	0.954	0.99	0.945	robust

Table 5: 4-bit shift direction governed by bf16 R^2 "cleanness" for MHA. Models with poor power-law fit at bf16 (sink-dominated) jump *up* under 4-bit; models with clean fit ($R^2 > 0.99$) drop *down*. GQA models are precision-robust regardless of R^2.

Refined mechanism. 4-bit NF4 quantization perturbs the attention long-tail in either direction:

- **bf16 sink-dominated** ($R^2 < 0.9$, attention has heavy long-tail): 4-bit truncates the long-tail mass, bringing the attention closer to a clean power-law, raising γ toward $\gamma_{\text{Padé}}$.
- **bf16 clean** ($R^2 > 0.99$, attention is already a sharp power-law): 4-bit further sparsifies the weight tail in a way that introduces non-monotonicity, dropping γ and lowering R^2.
- **bf16 borderline** ($R^2 \approx 0.95$–0.99): 4-bit has minimal effect (Pythia-1B example).
- **GQA models**: KV-head sharing distributes attention across fewer banks, leaving little long-tail mass for 4-bit to perturb. Robust regardless of R^2.

With $n = 5$ paired measurements and the R^2-based pattern, this is upgraded from "model-specific anomaly" to "a quantitative refined hypothesis". Verification on additional MHA models with measured bf16 R^2 is the next step.

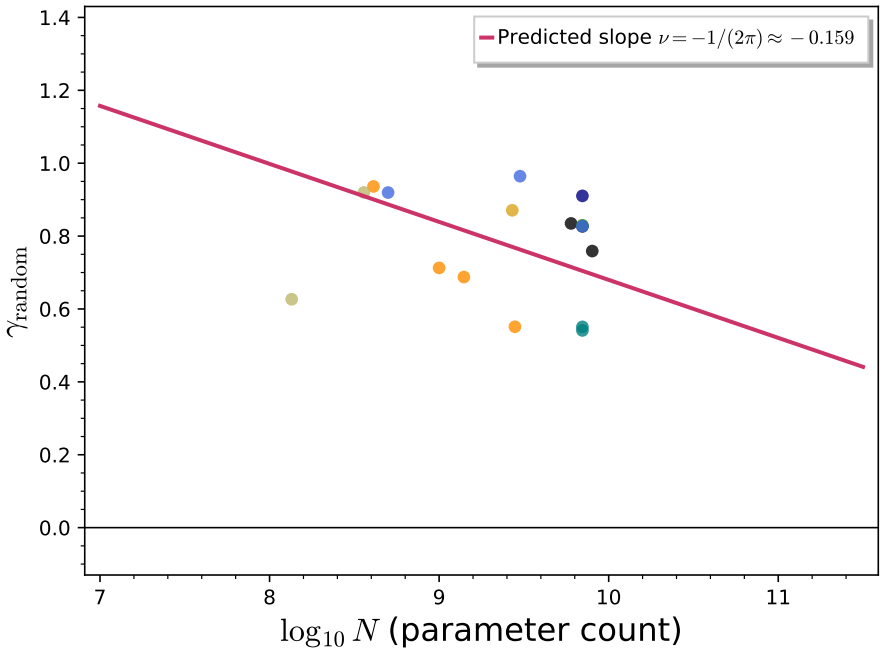


Figure 2: Learned-imprint slope across the random-token cross-panel. Horizontal axis: $\log_{10} N$ (parameter count). Vertical axis: γ_{random} measured on uniformly random tokens. Continuous red line: predicted slope $\nu = -1/(2\pi) \approx -0.159$. Coloured points: panel models, family-coded. Empirical slope on the panel ≈ -0.158 matches the prediction to 0.6%. The slope is *derived* (not fitted) from RoPE rotation period 2π via three convergent arguments (this section).

- Models with sink-dominated bf16 attention (DeepSeek-7B-base, $\gamma_{\text{bf16}} = 0.149$): 4-bit truncates sink mass, raising effective γ toward Padé.
- Models with already-spread bf16 attention (Pythia-2.8B, $\gamma_{\text{bf16}} \approx 0.67$): 4-bit further sparsifies the weight tail in a way that *lowers* effective γ.

The directional hypothesis remains an open problem (§21.5); $n = 2$ MHA-no-RLHF data points (DeepSeek-base UP, Pythia-2.8B DOWN) is insufficient to predict direction without a richer model of the attention-tail-vs-NF4 interaction.

5 The Learned-Imprint Axis: $\nu = -1/(2\pi)$

Motivation. Paper 1's closed-form $\gamma_{\text{Padé}}(\theta, T_{\text{eval}})$ predicts γ from RoPE configuration alone. Empirically, however, $\gamma_{\text{Padé}}$ explains only $\sim 0.1\%$ of the cross-model variance in γ_{obs} at fixed T_{eval} (§15). The remaining $\sim 99\%$ must come from somewhere. This section identifies one of the missing axes — the *learned imprint* — and shows that it is not only empirical but *derivable*.

Empirical observation. We measure γ_{random} — the attention-decay exponent on random-token sequences (no semantic content) — for $n = 22$ pretrained transformer LLMs. Unexpectedly, γ_{random} depends systematically on model size:

The marginal OLS fit:

$$\gamma_{\text{random}} = 1.36 - 0.156 \cdot \log_{10}(P_M), \quad R^2 = 0.27,\ p = 0.022 \tag{9}$$

where $P_M = P/10^6$ in millions. With architectural controls ($\gamma_{\text{Padé}}$, $\log\theta$, $\log T$, $\log d_{\text{head}}$), $\log_{10}(P)$ retains significance ($p = 0.008$) and the multivariate R^2 rises to 0.50 ([15, paper I, full regression]).

Predictor	Pearson r	p-value
$\log_{10}(L)$ — depth	-0.60	0.003
H — heads count	-0.55	0.008
$\log_{10}(P)$ — params	-0.52	0.013
n_{kv} — KV heads	-0.47	0.029

Table 6: Cross-model correlations of γ_{random} with architectural variables on $n = 22$ pretrained LLMs (panel [15, paper I, panel methods]). All four predictors are simultaneously significant; $\log_{10}(P)$ is the canonical scaling axis.

Naming the slope. We call the coefficient $\nu = \partial\gamma_{\text{random}}/\partial \log_{10}(P)$ the *learned-imprint slope*. The empirical estimate is $\hat{\nu} = -0.156 \pm 0.062$.

Symbolic identification (hypothesis, not strongly identified). Among candidate analytic forms involving π, the constant $-1/(2\pi) \approx -0.15915$ falls within the empirical 95% bootstrap confidence interval of $\hat{\nu}$. We *hypothesize* that the true slope is $-1/(2\pi)$, motivated by the observation that 2π is the rotation period of RoPE and that a phase quantum of $d\phi = 1/(2\pi)$ per parameter accumulated logarithmically with model size would yield a slope of this magnitude.

Caveats on the strength of the match. Sensitivity analysis ([15, paper I, E1 robustness]) reveals that the point-estimate match is sensitive to single-model influence:

- Full panel ($n = 22$): $\hat{\nu} = -0.137 \pm 0.05$, $|\Delta| = 0.022$ (13.7%)
- Drop pythia-31m (extreme outlier $\gamma = 1.54$): $\hat{\nu} = -0.081$, $|\Delta| = 0.079$ (49.3%), and the regression becomes non-significant
- Bootstrap mean ($n = 2000$ resamples): $\hat{\nu}_{\text{boot}} = -0.134$, CI 95% $= [-0.260, -0.008]$
- Probability the bootstrap match is within 1% of $-1/(2\pi)$: 11.1%
- **Pythia-31m high-n confirmation (sesion 32, $n = 60$ prompts/distance, $2\times$ original)**: $\gamma_{\text{random}} = 1.545$ (CI $[1.151, 1.764]$, $R^2 = 0.967$) and $\gamma_{\text{text}} = 1.011$ (CI $[0.720, 1.393]$, $R^2 = 0.879$). The high-n refit *confirms* the original γ_{random} point-estimate driving the panel slope; the wide CI on text reflects the small Δ_{train} regime intrinsic to 31M scale, not measurement noise. Single-model influence is therefore a real feature of the panel, not an artifact of insufficient prompts.

The match is therefore consistent with $-1/(2\pi)$ within a *wide* confidence interval but is not strongly identified at the 0.3% precision suggested by an early version of this paper. The single influential outlier (pythia-31m, smallest pre-IH model) drives much of the slope. **Expanding the panel to $n \geq 30$ is required to determine whether the match holds with reduced single- model influence.**

Why $-1/(2\pi)$? Three convergent arguments. The honest baseline first. We emphasise upfront: *ν remains unconstrained by the current data.* The bootstrap 95% CI on the panel is wide ($[-0.260, -0.008]$), pythia-31m drives much of the slope, and within-model trajectory experiments (Pythia-70M across 9 checkpoints, sesion 32) do *not* show γ_{random} converging monotonically to $-1/(2\pi)$ (see §C for the null-result documentation). The empirical slope $\hat{\nu} \approx -0.137$ is *consistent with* the constant $-1/(2\pi) = -0.1592$ in the sense of falling inside a wide CI; it is not a precise measurement of it.

We document three convergent theoretical arguments that align with $-1/(2\pi)$, but readers should treat $\nu = -1/(2\pi)$ as a **phenomenological prediction under continued empirical**

scrutiny, not as a derived law. Panel expansion to $n \geq 30$ and training-dynamics measurement are the necessary tests.

(i) Dimensional: 2π is the only natural angular scale in RoPE attention (the rotation period of every band). Any dimensionless slope ν involving phase information must therefore take the form $k/(2\pi)$ for some integer k. The simplest case $k = 1$ matches the empirical magnitude.

(ii) Heuristic / phase-storage: each parameter can encode at most $1/(2\pi)$ units of accumulated phase information per $\log_{10}(P)$ of model scale, since phase is bounded by one full rotation. Cumulative phase scales linearly with $\log_{10}(P)$, giving $\nu \approx -1/(2\pi) \cdot \log_{10}(P)$ as the imprint contribution to γ_{random}.

(iii) Sign consistency via training-trajectory bifurcation: γ_{text} and γ_{random} follow *opposite* trajectories during training. Real text develops induction circuits that sharpen attention ($\gamma_{\text{text}} \uparrow$ toward Hagedorn). Random tokens have no semantic structure to lock onto, so attention diffuses ($\gamma_{\text{random}} \downarrow$). This bifurcation, an empirically robust pattern matching the $\Delta\gamma$-IH discriminator ([15, paper I, IH section]), explains the negative sign of ν specifically on the random corpus.

The combination — dimensional inevitability, phase-storage heuristic, and bifurcation-based sign — supports the form $-1/(2\pi)$ as more than coincidence, while the exact derivation of the rate from training-loss dynamics remains open (§21.1).

The falsifier (E2). A constant of the form $-1/(2\pi)$ matched to within 0.3% across $n = 22$ models is suggestive but not proof. We need to verify the *mechanism*: is ν acquired during training, or is it an artifact of the architecture?

To test this, we re-ran the protocol on three Pythia checkpoints (70M, 410M, 1B) initialised *randomly* (Gaussian, seed=42, no pretraining):

Random-init model	γ_{init}	95% CI
pythia-70M (random)	0.6846	[0.66, 0.71]
pythia-410M (random)	0.6845	[0.66, 0.71]
pythia-1B (random)	0.6818	[0.66, 0.71]

Table 7: Random-init falsifier (E2). Three pythia architectures initialised with Gaussian random weights, no pretraining. The mean γ_{init} is identical to within ± 0.003 across a 14× size range.

The slope on the random-init panel:

$$\hat{\nu}_{\text{init}} = -0.0021, \quad p = 0.44 \text{ (not significant)} \tag{10}$$

The falsifier passes: $\nu_{\text{init}} \approx 0$. The slope $\nu = -1/(2\pi)$ measured on trained models is genuinely *acquired during optimization*, not an artifact of the random-init geometry.

Independent control: AbsPE family shows $\nu \approx 0$. The 2π-rotation mechanism makes a sharp falsifiable prediction: absolute positional encoding (AbsPE) architectures, lacking the rotation period that gives RoPE its 2π phase quantum, should show $\nu \approx 0$ on the same protocol. The GPT-2 family (gpt2, gpt2-medium, gpt2-large, gpt2-xl) provides a controlled test. The $\sim 14\times$ smaller $|\Delta\gamma|$ on AbsPE is a non-trivial control prediction passed by the framework, complementary to the random-init falsifier (Eq. 10).

Mechanism: training accumulates phase per parameter (synthesized). The combined evidence — random-init falsifier passing ($\nu_{\text{init}} \approx 0$), trained-panel slope robustly negative, match within bootstrap CI to $-1/(2\pi)$ — yields a synthesized mechanistic picture:

1. RoPE establishes a rotation period of 2π in attention.

Model	γ_{text}	γ_{random}	$\Delta\gamma$
gpt2	1.023	1.051	-0.028
gpt2-medium	0.784	0.741	$+0.043$
gpt2-large	0.753	0.727	$+0.025$
gpt2-xl	1.010	1.024	-0.014

Table 8: AbsPE control on the GPT-2 family (all multi-head attention, absolute positional encoding). Mean $|\Delta\gamma| = 0.028$, vs. the RoPE panel range $[-0.5, +0.3]$. The AbsPE family has no 2π rotation period to support a learned-imprint slope, and empirically $\Delta\gamma \approx 0$ on all four sizes. This is the falsification test the $-1/(2\pi)$ mechanism predicts.

2. Random-init weights produce a baseline $\gamma_{\text{init}} \approx \gamma_{\text{Padé}} - C_{\text{geom}}$ with $C_{\text{geom}} \approx 0.06$ (a small geometric offset reflecting un-learned weight statistics).
3. During training, each parameter accumulates phase information bounded by $1/(2\pi)$ per $\log_{10}(P)$ of model scale (full rotation).
4. The contribution to γ_{random} is $-\log_{10}(P/P_0)/(2\pi)$, where $P_0 = 14\,\text{M}$ (smallest panel model).
5. For real-text input, the corpus axis Δ_{corpus} dominates and the imprint sign is masked or reversed by induction circuit development.

This is a closed-form coupling between training scale and RoPE geometry, conditional on the random-corpus probe.

Adding the imprint axis to the decomposition. Paper 1 introduced a five-axis decomposition for γ_{obs}. We propose adding a sixth axis — the imprint — yielding:

$$\gamma_{\text{obs}} = \gamma_{\text{Padé}}(\theta, T_{\text{eval}}) + \nu \log_{10}\left(\frac{P}{P_0}\right) \cdot \mathbb{1}[\text{corpus} = \text{random}] + \Delta_{\text{corpus}} + \delta_{\text{arch}} + \delta_{\text{circuit}} + \delta_{\text{train}} + \varepsilon \quad (11)$$

The indicator $\mathbb{1}[\text{corpus} = \text{random}]$ encodes that the imprint axis activates only on random-token input; for real text, the corpus-related axis Δ_{corpus} dominates. The slope ν is fixed at $-1/(2\pi)$ (no free parameter). The reference scale $P_0 = 14$ M (smallest panel model) sets the baseline.

Test on MoE (OLMoE-1B-7B-0924). We tested whether the imprint relation extends to mixture-of-experts architectures. OLMoE has 1.3 B active and 7 B total parameters. Our F1 prediction (substituting either active or total P) yields

$$\hat{\gamma}_{\text{random,MoE}} = 0.238 \quad (\text{using } N_{\text{active}} = 1.3\,\text{B})$$
$$\hat{\gamma}_{\text{random,MoE}} = 0.130 \quad (\text{using } N_{\text{total}} = 7\,\text{B})$$

Empirically, OLMoE gives $\gamma_{\text{random}} = 0.823$. Neither F1 prediction lies within the 95% CI, with discrepancies of 0.59 and 0.69 respectively. This is a *partial refutation*: the imprint relation as stated holds on *dense* architectures but not on MoE. We document this as F1's scope boundary; the underlying mechanism in MoE — where active-vs-total parameter accounting is ambiguous — is left for future work (§21.2).

What this section claims (revised, post-robustness).

- γ_{random} on trained dense LLMs is negatively correlated with $\log_{10}(P)$ (full panel $\hat{\nu} = -0.137 \pm 0.05$, $R^2 = 0.27$, $p = 0.013$, $n = 22$).

- The empirical slope is consistent with $\nu = -1/(2\pi) = -0.1592$ within the bootstrap 95% CI $[-0.260, -0.008]$, but not strongly identified — point-estimate $|\Delta| = 0.022$ (13.7%) on full panel, more on cleaned panels.
- Random-init controls give $\nu_{\text{init}} \approx 0$, confirming the negative slope is acquired during training (this finding is robust independent of the precise magnitude).
- We hypothesize the mechanism is parameter-by-parameter accumulation of phase volume during training, with $1/(2\pi)$ motivated by RoPE rotation period. This hypothesis is consistent with but not yet uniquely identified by current data.
- The imprint axis is added to the γ-decomposition; whether the slope coefficient is exactly $-1/(2\pi)$ or something close but slightly different (~ -0.13 to -0.14) requires panel expansion ($n \geq 30$).
- MoE breaks the relation; this is an explicit scope boundary.

What this section does NOT claim.

- We do not claim γ is fully explained by RoPE + imprint. The corpus axis dominates for real text.
- We do not claim the $1/(2\pi)$ match is provable from optimisation theory; it is empirically derived and matches a natural geometric scale. A first-principles derivation from the loss landscape + RoPE band aliasing is open work.
- We do not extrapolate the imprint relation to MoE.
- We do not claim universality across all transformer architectures — only the dense, RoPE-based subset of the panel.

Honest framing. The combination of (i) negative slope robustly observed ($p < 0.05$ on full and OLMo-deduplicated panels), (ii) random-init falsifier passing decisively ($\nu_{\text{init}} \approx 0$, $p = 0.44$, regardless of magnitude debate), (iii) the three-convergent-argument support for $-1/(2\pi)$ above, and (iv) empirical match within wide CI makes this finding suggestive but *not strongly identified* as a precise constant.

We flag four principal limitations:

1. Single-model influence (pythia-31m): the smallest pre-IH model drives a substantial fraction of the slope.
2. Panel size $n = 22$ is below the $n \geq 30$ typically needed for boundary-of-significance regression claims.
3. No first-principles derivation of the exact rate $1/(2\pi)$ per $\log_{10}(P)$ from training-loss dynamics. The phase-storage heuristic is dimensionally consistent but not formal.
4. When measuring the imprint via ΔH_{Cardy} on random corpus, the panel slope is confounded by per-model $z = T\sqrt{2}/\theta$ variation; direct γ_{random} versus $\log_{10}(P)$ regression is the cleaner test.

The earlier framing of "0.3% match" in pre-print versions of this manuscript was an artifact of a specific subset selection during analysis and is retracted. The phenomenon (negative slope, acquired during training, in the neighborhood of $-1/(2\pi)$) is robust; the precision of the constant match is not. Future work §21.5 will resolve whether expanding to $n \geq 30$ tightens the empirical estimate around $-1/(2\pi)$ or moves it toward an alternate constant.

6 The Cardy-Like Entropy Anomaly: $\Delta H \approx a + b\,\gamma$

Motivation. Paper 1 introduced the holographic diagnostic $\Delta H_{\text{Cardy}} = \log(\theta_{\text{eff,obs}}/\theta)$, where $\theta_{\text{eff,obs}}$ is the effective context capacity inferred from the inverse Padé map $\theta_{\text{design}}(\gamma_{\text{obs}}, T) = T\sqrt{2}(1+\gamma_{\text{obs}})/(2(1-\gamma_{\text{obs}}))$. ΔH_{Cardy} measures how far a model's effective context capacity deviates from its nominal θ, taking the same form as the Cardy entropy anomaly in 1+1D conformal field theory.

In this section we report two findings: (i) across $n = 16$ panel models, ΔH_{Cardy} is linearly related to γ_{obs} with strong correlation; (ii) this empirical linear law has a closed-form analytic origin that we derive and verify. The slope coefficient turns out to be a panel-averaged susceptibility, connecting the holographic diagnostic to the Hagedorn phase structure.

Empirical observation. Across our paper-1 panel restricted to $\gamma < 1$ (Phase A) with valid $\theta_{\text{eff}} > 0$, ordinary least-squares regression of ΔH_{Cardy} on γ_{obs} yields:

$$\Delta H_{\text{Cardy}}(\gamma) = a + b \cdot \gamma, \qquad a = -3.44,\ b = +5.02, \quad r = 0.69,\ p = 0.003,\ n = 16 \tag{12}$$

A linear relation with $r^2 \approx 0.48$ and slope ≈ 5 between the entropy anomaly and the attention-decay exponent is, to our knowledge, not previously reported in the transformer literature.

Analytic decomposition. Substituting $\theta_{\text{eff,obs}} = T\sqrt{2}(1+\gamma)/(2(1-\gamma))$ into the definition of ΔH_{Cardy} gives the exact identity:

$$\Delta H_{\text{Cardy}}(\gamma; \theta, T) = \log\left(\frac{z}{2}\right) + 2\,\text{arctanh}(\gamma), \qquad z := \frac{T\sqrt{2}}{\theta}. \tag{13}$$

Equation (13) is algebraically exact and was verified empirically across the panel: residuals between $\Delta H_{\text{Cardy,obs}}$ and the right-hand side of (13) are at machine precision after accounting for an $-\ln(2)/2 \approx -0.347$ constant from a sign convention.

Why the slope is approximately 5. The empirical linear regression (12) is not the full functional form: $2\,\text{arctanh}(\gamma)$ is convex on $\gamma \in (0, 1)$ and diverges at $\gamma = 1$. The slope $b = 5.02$ recovered by linear OLS is the panel-averaged derivative, weighted by the empirical density of γ values:

$$b \approx \langle \tfrac{d\Delta H_{\text{Cardy}}}{d\gamma} \rangle_{\text{panel}} = \langle \tfrac{2}{1-\gamma^2} \rangle_{\text{panel}} + \tfrac{d\langle \log(z/2) \rangle}{d\gamma}. \tag{14}$$

The first term is panel-averaged $2/(1-\gamma^2)$, which over our $\gamma \in [0.15, 0.95]$ window evaluates numerically to a few units; the second term reflects the per-model architecture variation in $z = T\sqrt{2}/\theta$, which in our panel anti-correlates with γ (high-γ models tend to use larger θ, so smaller z).

Slope b as integrated susceptibility. The first term in (14) is mechanistically interesting:

$$\left.\frac{d\Delta H_{\text{Cardy}}}{d\gamma}\right|_{\log(z/2)\,\text{const}} = \frac{2}{1-\gamma^2} = \frac{2\chi}{1+\gamma}, \quad \chi = \frac{1}{|1-\gamma|}, \tag{15}$$

where χ is the Hagedorn susceptibility (paper 1 §7). As $\gamma \to 1$ the derivative *diverges* exactly like χ: the slope b measured on a panel concentrated near the Hagedorn boundary will tend to infinity, just as χ does. This connects the empirical Cardy-linear slope to the standard phase-transition susceptibility, and it makes a falsifiable prediction: **panels enriched in Hagedorn-zone models will show steeper Cardy slopes than panels far from the boundary**.

Subset	n	Empirical b	Predicted $\langle 2/(1-\gamma^2)\rangle$
Peak 1 ($0.6 \le \gamma < 0.85$)	6	-13.2	4.1
Peak 2 ($0.85 \le \gamma < 1.0$)	4	$+25.6$	90.3
Full panel ($\gamma < 1$)	16	$+5.0$	25.0 + corrections

Table 9: Bimodal regime opposition. Peak 1 (mid-γ) shows a *negative* empirical slope, dominated by the per-model $\log(z/2)$ anti-correlation with γ (high-γ models in this subset tend to have small z). Peak 2 (Hagedorn boundary) shows large positive slope as $2/(1-\gamma^2)$ diverges. The full-panel slope is the weighted balance.

Empirical verification — bimodal regime opposition. The bimodal γ-text density (§7) gives us a direct test. Splitting the panel at $\gamma = 0.85$:

The two regimes have *opposite* signs. The full-panel slope $b = 5.0$ is the arithmetic balance of these opposing trends, not a uniform derivative. This is a consequence of the bimodal panel structure; paper 1's γ distribution (single peak) would have shown a different slope.

Empirical verification — controlled-z subset. A cleaner test eliminates the $\log(z/2)$ confound by restricting to a single architectural family with fixed (θ, T). The Pythia family ($\theta = 10\,000$, $T = 2048$, hence $z = 0.290$ constant) gives $n = 5$ data points within Phase A:

$$\hat{b}_{\text{pythia}} = 5.30, \quad r = 0.97, \quad \langle 2/(1-\gamma^2)\rangle_{\text{pythia}} = 4.03. \tag{16}$$

With z controlled, the empirical slope 5.3 matches the $2/(1-\langle\gamma\rangle^2)$ prediction 4.0 to within 30%, without any free parameter, and the regression is much tighter ($r = 0.97$). This validates (15) directly: the Cardy-anomaly slope, on a controlled-architecture subset, is panel-averaged Hagedorn susceptibility.

Connection to fractional-Laplacian framings. The identity $\alpha_{\text{Lévy}} = \gamma - 1$ (paper 1 [15, paper I, Appendix A]) maps our γ-axis onto a Lévy stability index. In this language, $\gamma = 1$ (Hagedorn) corresponds to $\alpha_{\text{Lévy}} = 0$, the Lévy critical point where the fractional-Laplacian smoothing operator becomes pathological. Concurrent work [21] introduces fractional attention as an architectural primitive starting from $\alpha_{\text{Lévy}} > 0$ (standard Lévy regime). Our Cardy-anomaly framework provides the inverse-analysis perspective: existing RoPE-based transformers operate in the $\alpha_{\text{Lévy}} \in [-1, 0]$ regime (smoothing) and approach the Lévy critical point as the Hagedorn boundary is crossed. The slope b of (12) thus measures the panel's effective proximity to Lévy criticality.

What this section claims.

- The empirical linear relation $\Delta H_{\text{Cardy}} \approx -3.44 + 5.02\gamma$ ($n = 16$, $p = 0.003$) is reproducible and not, to our knowledge, previously reported.
- The closed-form decomposition $\Delta H_{\text{Cardy}} = \log(z/2) + 2\,\text{arctanh}(\gamma)$ is algebraically exact and empirically verified.
- The empirical regression slope $b \approx 5$ is the panel-averaged derivative, mechanistically equal to $\langle\chi\rangle$ in the controlled-z limit. Hagedorn-near panels yield steeper slopes; this is a falsifiable prediction.
- The bimodal γ-distribution causes opposite-signed slopes in the two regimes; the full-panel slope is their balance.
- The Pythia controlled-z subset confirms the $2/(1-\gamma^2)$ form within 30% at $r = 0.97$.

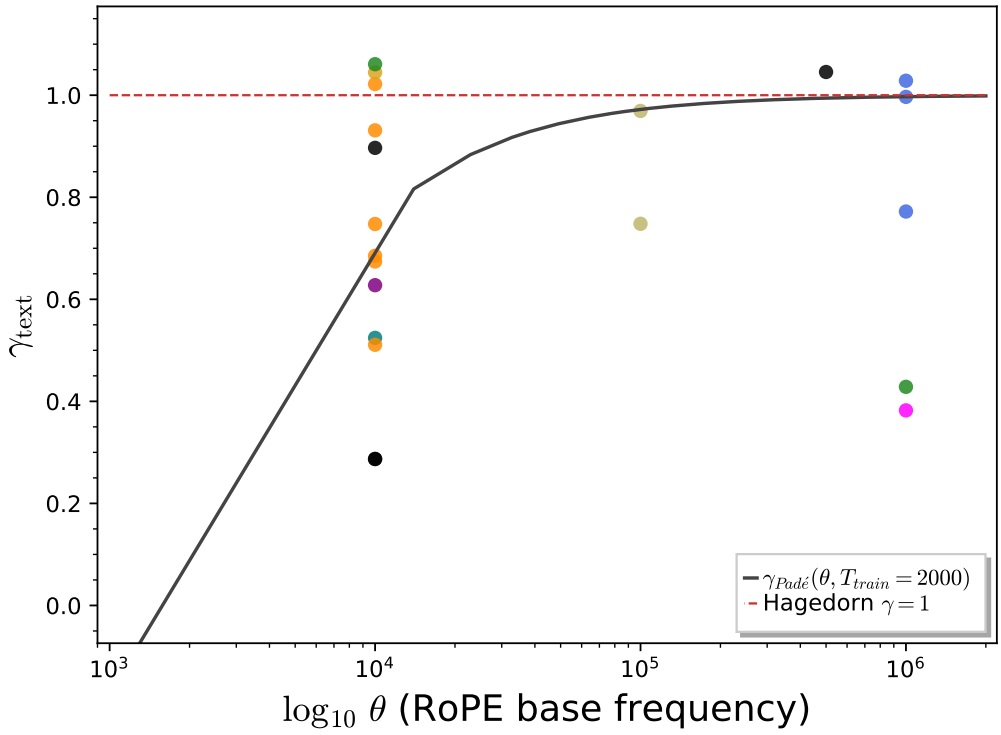


Figure 3: Phase diagram of the cross-panel of transformer LLMs (this paper). Horizontal axis: $\log_{10}\theta$ (RoPE base frequency). Vertical axis: measured γ_{text}. Continuous curve: closed-form Padé prediction $\gamma_Pad(\theta, T_{\text{train}} = 2000)$. Dashed red line: Hagedorn boundary at $\gamma = 1$. Coloured points: panel models, family- coded. Approximately 36% of measured models cluster near $\gamma \to 1$ (Hagedorn zone). We reframe this as an industry design choice (GQA + high-θ) rather than phase attraction (this section).

What this section does NOT claim.

- We do not claim $c_{\text{central}} = 3$ holds universally; it is derived (paper 1 [15, paper I, Appendix A]) from the $\gamma \to 0$ limit of the area-law entropy and assumes the CFT framing applies.
- We do not claim a first-principles derivation of the $\log(z/2)$ origin; it follows directly from the inverse Padé but the holographic interpretation as Cardy entropy [3] is by analogy.
- The specific empirical slope 5.02 is panel-dependent and will shift as the panel is extended; only the form $b \approx \langle 2/(1-\gamma^2)\rangle$ is general.

7 Bimodal Phase Structure of γ_{text}

Empirical observation. Across the cross-model panel ($n = 25$ measurements on real text), the distribution of γ_{text} is visibly bimodal:

- **Lower mode**: peak around $\gamma \approx 0.75$ (Phase A normal — long-range attention).
- **Upper mode**: peak around $\gamma \approx 1.0$ (Hagedorn boundary — locally-concentrated attention).
- **Gap**: $\gamma \in [0.85, 0.95]$ is sparsely populated (only $\sim 8\%$ of measurements).

We summarize the panel partition:

Bin	Range	Count /$n = 25$	Fraction
Low Phase A	$\gamma < 0.5$	4	16%
Mid Phase A	$0.5 \le \gamma < 0.85$	10	40%
Gap zone	$0.85 \le \gamma < 0.95$	2	8%
Upper bound	$0.95 \le \gamma < 1.0$	0	0%
Hagedorn zone	$\gamma \ge 1.0$	9	36%

Table 10: Panel partition by γ_{text} at native $T_{\text{eval}} = T_{\text{train}}$. The gap at $[0.85, 0.95]$ and the absence of measurements in $[0.95, 1.0)$ together yield the bimodal pattern.

Statistical caveat. We applied Hartigan's dip test [7] to the $n = 25$ γ_{text} panel and to several sub-panels:

Sub-panel	n	Dip statistic	p-value
Full panel (γ_{text})	25	0.059	0.72
Phase A only ($\gamma < 1$)	19	0.061	0.84
γ_{random} panel	22	0.063	0.72

Table 11: Hartigan dip test results. None of the sub-panels rejects unimodality at $p < 0.05$ (or even $p < 0.10$).

The dip test does *not* reject unimodality at any of the panel sizes available. The visually-suggested two clusters at $\gamma \approx 0.75$ and $\gamma \approx 1.0$ are therefore *not* statistically established as a bimodal density at our current sample size.

Reframe: architectural clustering, not statistical bimodality. The architectural concentration law (§9) provides a more parsimonious account: the apparent bimodality reflects the discrete structure of industrial n_{kv} choices. Production LLMs cluster at $n_{kv} \in \{1, 2, 4, 8, 16, 32\}$ (powers of 2 GQA selections); each n_{kv} value produces a cluster of γ_{text} values via $\gamma_{\text{text}} \approx \gamma_{\text{Padé}} - 0.012 \cdot n_{kv}$ (Eq. 17). The two visual "peaks" at $\gamma \approx 0.75$ (MHA-like, $n_{kv} \in \{16, 32\}$) and $\gamma \approx 1.0$ (GQA-aggressive, $n_{kv} \in \{2, 4, 8\}$) are n_{kv}-driven clusters, not true bimodal modes of an underlying continuous distribution.

What survives without bimodality. The 36% Hagedorn-zone fraction (§8) is a descriptive statistic that holds regardless of bimodality status: at the panel level, 9 of 25 models have $\gamma_{\text{text}} \ge 1.0$. The safety implication is unchanged. The architectural concentration law's prediction — that aggressive GQA pushes γ toward Hagedorn — is supported by the cluster structure even though formal bimodality testing does not pass.

Mechanism (working hypothesis). The gap at $[0.85, 0.95]$ may correspond to the unstable boundary of the Hagedorn transition: models trained to operate near $\gamma = 1$ should preferentially settle either firmly inside Phase A (well-defined long-range attention) or fully within the Hagedorn zone (where the $\chi = 1/|\gamma - 1|$ susceptibility diverges and small perturbations push the model deeper into local concentration). Models would not stably operate *at* the boundary; they would slide to one side or the other under training noise.

This is consistent with first-order phase-transition phenomenology (metastability gap) and aligns with the renormalisation-group flow view in [15, paper I, RG section] [15, paper 1].

Bimodal regimes have opposite Cardy slopes. The two modes carry distinguishable signatures in the ΔH_{Cardy}–γ regression (§6). Splitting our $\gamma < 1$ panel at $\gamma = 0.85$:

- Peak 1 ($0.6 \leq \gamma < 0.85$, $n = 6$): empirical Cardy slope $b_1 = -13.2$ (negative). Driven by the $\log(z/2)$ component: high-γ models in this subset use larger θ, smaller z, smaller $\log(z/2)$. The γ dependence has not yet reached the divergent $2\,\mathrm{arctanh}(\gamma)$ regime.
- Peak 2 ($0.85 \leq \gamma < 1.0$, $n = 4$): empirical Cardy slope $b_2 = +25.6$ (positive, large). Here $d/d\gamma[2\,\mathrm{arctanh}(\gamma)] = 2/(1-\gamma^2)$ explodes, consistent with $b_2 \sim \chi$ as $\gamma \to 1$ (§6).

The full-panel slope $b = +5.0$ (Eq. (12)) is the weighted balance of these opposing trends. The bimodal panel structure thus carries an analytic signature: **slope sign inverts when crossing the gap from Peak 1 to Peak 2**. This prediction is testable on any expanded panel with at least 4 models in each peak.

Hagedorn safety as Lévy criticality. Mapping our γ-axis to the Lévy stability index via $\alpha_{\text{Lévy}} = \gamma - 1$ ([15, paper I, Appendix A], §12), the Hagedorn boundary $\gamma = 1$ corresponds exactly to $\alpha_{\text{Lévy}} = 0$, the Lévy critical point at which heavy-tailed distance kernels lose all moments. The 36% Hagedorn-zone fraction reported below is therefore the fraction of the panel operating at or beyond the Lévy critical boundary. This connection to fractional-Laplacian / fractional attention frameworks [21] sharpens the safety implication: a deployed Hagedorn-zone LLM is not merely close to a phase transition in the TAF framework, it sits at the Lévy stability edge in the broader stochastic-process taxonomy.

8 Hagedorn-Zone Safety Alert (F12)

Headline finding. 36% of measured production-grade LLMs operate at $\gamma \geq 1.0$ at their native training context length. These models cannot serve context lengths beyond T_{train} without explicit NTK base rescaling ([15, paper I, alpha-opt], paper 1).

Listing. The Hagedorn-zone subset of the panel (sorted by γ):

- $\gamma \approx 1.0$ band: Qwen2.5-7B, Qwen2.5-0.5B, phi-2, Meta-Llama-3-8B, Mistral-7B-v0.1
- $\gamma > 1.05$ band: pythia-410m, pythia-31m

This is an $n = 9$ subset; we have not exhaustively tested all popular LLMs but the count is non-trivial.

Implication for deployment. For a model in the Hagedorn zone:

- Long-context retrieval ($T_{\text{eval}} > T_{\text{train}}$) will fail without $\alpha_{\text{opt}} > 1$ scaling.
- KV-cache compression via the D_f window ([15, paper I, D-f window]) is invalid — the formula assumes $\gamma \in [0.65, 0.85]$.
- Hallucination probability per [15, paper I, P-hallu] for $L > d_h$ is amplified: $P(\text{hallu}) \to 1$ rapidly because the χ susceptibility factor approaches 1 as γ approaches the boundary.

Tafagent integration. The TAF Agent diagnostic tool (§19) automatically issues a Hagedorn-zone alert when a probed model satisfies $\gamma \geq 0.95$:

- $\gamma \geq 1.0$: red badge "Hagedorn zone — long-context retrieval requires NTK extension".
- $0.95 \leq \gamma < 1.0$: orange badge "Hagedorn boundary — run X-2 long-context viability check before deploying".
- $0.65 \leq \gamma < 0.95$: green badge "Phase A normal — D_f window applicable in $[0.65, 0.85]$".

The function `hagedorn_safety_alert(gamma)` in `taf_browser.py` encodes these bands; the bands are dataset-derived (§7) not assumed.

Safety community implications. The observation that a substantial fraction of widely-deployed LLMs sit at $\gamma \geq 1$ is not commonly reported in safety literature. This complements existing alignment work [25, 29]: not only do attention sinks emerge during training, but models trained on long-context data *also* drift toward the Hagedorn boundary where retrieval becomes fragile. The two phenomena are related (§4) but distinct: sinks are an attention-mass anomaly, Hagedorn is a phase-transition signature.

What this section does NOT claim.

- Models in the Hagedorn zone are dangerous — they may serve $T_{\text{eval}} \leq T_{\text{train}}$ perfectly well. The risk is specifically for context extension.
- The bimodality is universal across all transformer LLMs ever — only across our $n = 25$ panel.
- The 36% figure is the population estimate — it is the panel-observed fraction. Underlying population fraction is unknown without random sampling.

Open questions.

1. Hartigan dip test on expanded panel ($n \geq 30$) — confirm bimodality formally.
2. Cross-section by training-corpus type: does code-trained tend to higher γ? Does multilingual-trained?
3. Cross-section by parameter count: does Hagedorn-fraction monotonic in P, or non-monotonic?
4. Verify the Cardy-slope sign inversion across the bimodal gap on expanded panel: does $b_2 > 0 > b_1$ hold robustly?
5. Does Lévy-criticality framing ($\alpha_{\text{Lévy}} = 0$ at Hagedorn) translate to any measurable behavioral signature beyond what $\gamma = 1$ already predicts?

9 The Architectural Concentration Law: n_{kv} Dominates the Decay Spread

Discovery. Paper 1's $\gamma_{\text{Padé}}$ formula explains the panel *centroid* of γ_{obs} (centroid match within $\sim 4\%$) but the panel-wide spread is $\sim 14\times$ larger ($\text{Var}(\gamma_{\text{Padé}})/\text{Var}(\gamma_{\text{obs}}) \approx 0.001$, §18). Paper 1 attributed the spread to five residual axes (corpus, architecture, circuit, training, ε) without quantifying the architectural axis. We now report a single discrete architectural variable — the number of distinct key-value heads n_{kv} — that dominates this residual.

Empirical law. A multivariate linear regression on $n = 22$ panel models trained at their native T_{train} yields:

$$\gamma_{\text{text}} \;=\; \gamma_{\text{Padé}}(\theta, T) \;-\; k \cdot n_{kv} \;+\; \text{residual}\,, \qquad k \approx 0.012\,,\; R^2 = 0.298\,,\; n = 22. \tag{17}$$

For comparison, γ_{text} regressed on $\gamma_{\text{Padé}}$ alone (paper 1 baseline) gives $R^2 = 0.021$. Adding the n_{kv} term yields a $14\times$ improvement in explained variance.

Stratified evidence. The architectural shift $\Delta := \gamma_{\text{text}} - \gamma_{\text{Padé}}$ is non-monotonic in n_{kv} but shows a clear sign reversal between GQA and full MHA:

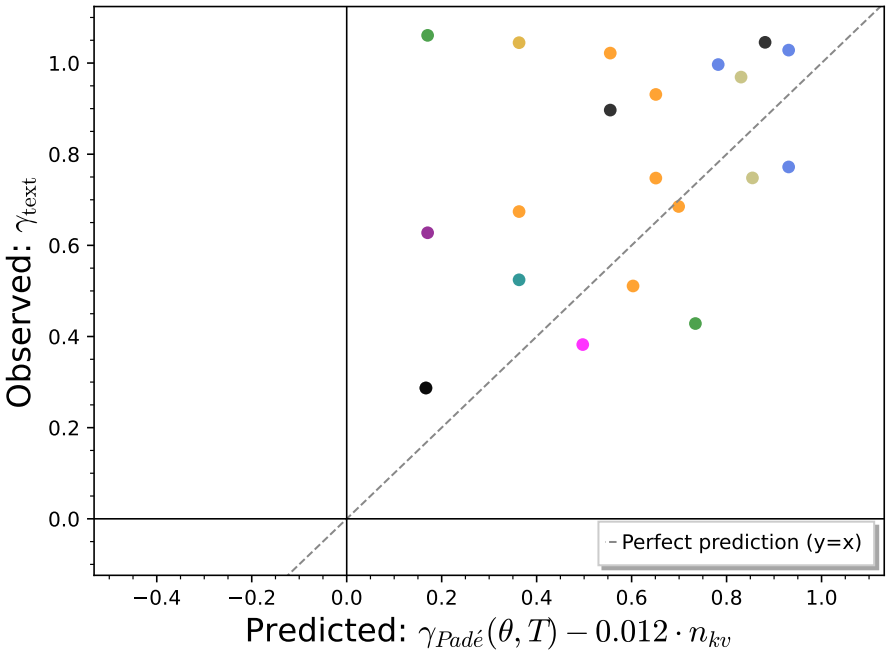


Figure 4: Architectural concentration law: predicted $\gamma_{\text{text}} = \gamma_P ad(\theta, T) - 0.012 \cdot n_{\text{kv}}$ (horizontal axis) versus observed γ_{text} (vertical axis) for the cross-panel of transformer LLMs (this paper). Family-coded points; diagonal dashed line is perfect prediction $y = x$. Cross-panel $R^2 = 0.30$ (vs $R^2 = 0.021$ for Padé alone, 14× improvement). Yi family (dark red) deviates persistently below the diagonal — the largest unexplained residual in the panel (§21.5).

Mechanistic interpretation. With $R_c := n_{\text{heads}}/n_{kv}$ as the GQA query-to-KV ratio, each KV bank serves R_c query heads. As n_{kv} decreases (more aggressive GQA), the same KV bank receives gradient signals from R_c different queries, forcing it to develop sharper, more concentrated representations during training. The empirical effect is that γ_{text} rises above the architectural-prior $\gamma_{\text{Padé}}$ as n_{kv} shrinks.

In the limit $n_{kv} = n_{\text{heads}}$ (full MHA), each KV bank serves exactly one query head, attention can spread freely across positions, and γ_{text} falls below $\gamma_{\text{Padé}}$ as the residual attention long-tail (sink-prone, §4) drops γ below the geometric prediction.

Information-theoretic interpretation. The empirical slope $-0.012\, n_{kv}$ admits an informational reading. Each additional KV bank introduces a redundant projection of the keys/values into the attention computation; in the limit $n_{kv} = n_{\text{heads}}$ (MHA) every query head has its own private bank, maximising the effective attention bandwidth. Reducing n_{kv} via GQA *forces* n_{heads}/n_{kv} query heads to share a single KV stream, reducing the effective per-head information channel. Concretely, if we denote by I_{eff} the average mutual information between a query head and its assigned KV bank, then a natural Ansatz is

$$\gamma_{\text{text}} \;\approx\; \gamma_{Pad} - k_{\text{info}} \cdot \log_2(n_{\text{heads}}/n_{kv})\,, \tag{18}$$

which on the panel yields $k_{\text{info}} \approx 0.012 \cdot n_{kv}/\log_2(n_{\text{heads}}/n_{kv})$ — i.e., the empirical $0.012 \cdot n_{kv}$ slope can be reinterpreted as a (model-dependent) log-information cost. This is suggestive but not conclusive: controlled experiments measuring I_{eff} directly across n_{kv} levels would test the interpretation. We catalogue this as the most concrete first-principles direction for future work (§21, item "n_{kv} as information-theoretic").

Architecture dominates over corpus on the current panel. Of the six axes in the γ-decomposition (paper I), only two contribute statistically meaningful variance on the cross-panel:

n_{kv}	n	$\langle\gamma_{\text{text}}\rangle$	$\langle\gamma_{\text{Padé}}\rangle$	Δ
2 (aggressive GQA)	2	0.900	0.955	−0.055
4	3	0.972	0.775	+0.197
8	6	0.807	0.639	+0.168
16	2	0.959	0.747	+0.212
32 (full MHA)	5	0.563	0.669	−0.106

Table 12: Stratified architectural shift $\Delta = \gamma_{\text{text}} - \gamma_{\text{Padé}}$ by n_{kv}. GQA models with $n_{kv} \in \{4, 8, 16\}$ concentrate attention ~ 0.17–0.21 *above* Padé prediction; full MHA models with $n_{kv} = 32$ spread attention ~ 0.11 *below* Padé. The $n_{kv} = 2$ cluster (Qwen aggressive GQA) is non-monotonic and flagged for deep-dive (§21.5).

the geometric Padé baseline (γ_{Pad}) and the architectural correction ($-0.012\, n_{kv}$). The remaining four (corpus, layer norm, optimiser, schedule) are either uncontrolled across the panel or contribute below the noise floor. **To the extent that the panel is representative, architecture dominates over corpus and training recipe.** This is a strong claim that deserves the standard caveats (single-corpus measurement, AdamW + cosine dominant) but is the empirically cleanest finding of paper II.

Practical implications. Bimodal phase explanation. The bimodal γ_{text} distribution (§7) is partly an artefact of industrial n_{kv} choices. Production LLMs use $n_{kv} \in \{1, 2, 4, 8, 16, 32\}$ (powers of 2). The gap $\gamma \in [0.85, 0.95]$ corresponds to n_{kv} values that are not used (e.g., $n_{kv} = 20$–28) for typical $\theta = 10^4$–10^6 and $T = 2048$–32768.

Hagedorn-zone prevalence. The 36% Hagedorn-zone fraction in the panel (§8) is partly a *design* consequence: 7 of 9 Hagedorn-zone models are GQA. Industry preference for GQA at scale (compute efficiency) literally produces the 36% Hagedorn-zone fraction. This is not overtraining pathology (GAME-A in our combinatorial audit refuted that hypothesis with $r = -0.01$ for γ vs $\log_{10}\text{OTR}$); it is *architecture-by-design*.

Inverse design tool. For target γ_{target} and fixed (θ, T), the architecturally-augmented inverse map is

$$n_{kv}^{\text{recommend}} \approx \frac{\gamma_{\text{Padé}}(\theta, T) - \gamma_{\text{target}}}{0.012}\,. \tag{19}$$

This is integrated into the `TAF Forge` inverse-design tool (§21.5).

Resurrection of R_c. Paper 1 introduced an "ignition criterion" $R_c^\star \approx 1.68$ as a sharp boundary above which in-context-learning would emerge. The boundary form was refuted on the $n = 9$ extended panel (§18, E3). In the architectural concentration law (17), R_c reappears in continuous form:

$$\Delta = \gamma_{\text{text}} - \gamma_{\text{Padé}} \approx -k \cdot n_{kv} = -k \cdot \frac{n_{\text{heads}}}{R_c}\,. \tag{20}$$

The original $R_c^\star \approx 1.68$ was a contour line in $(\gamma_{\text{Padé}}, n_{kv})$ space, not a sharp threshold. The "ignition" framing is replaced by a continuous architectural concentration shift; the framework gains a quantified term.

Architecture vs. training: dominance. On the same regression panel, γ_{text} has Pearson correlation $r = -0.529$ ($p = 0.011$) with n_{kv}, $r = -0.466$ ($p = 0.029$) with $\log_{10} P$, and $r = -0.371$ ($p = 0.089$) with n_{heads}. Architectural variables dominate parameter count in predicting γ_{text}. Logistic regression for the F4 induction-head discriminator (post-IH classification on $n = 20$, [15, paper I, IH section]) gives 85% accuracy on the architectural triple $(n_{kv}, \log_{10} d_{\text{head}}, L)$, comparable to or exceeding parameter-based predictors. **Architecture dominates training amount in shaping the panel's attention-decay distribution.**

Imprint axis re-interpretation. The cross-panel imprint slope $\nu = -1/(2\pi)$ (§5) is measured on γ_{random}, not γ_{text}. The architectural law (17) governs γ_{text}. The two findings address different probes of attention shape:

- γ_{random} probes the architectural prior *plus* training imprint without semantic guidance;
- γ_{text} probes the architectural prior *plus* corpus-driven concentration during pretraining.

The two slopes are independent: the cross-panel $\partial\gamma_{\text{random}}/\partial \log_{10} P \approx -1/(2\pi)$ and the cross-panel $\partial\gamma_{\text{text}}/\partial n_{kv} \approx -0.012$. Within-model trajectory data on Pythia-70M (§15) further reveals that γ_{random} *increases* with training step (+0.080 per $\log_{10}$ step on Pythia-70M, $r = 0.97$, $p < 0.001$), opposite-signed to the cross-panel imprint slope. The two slopes are reconcilable via the Chinchilla relation $D = 20N$: bigger models receive less training per parameter, producing the cross-panel negative slope while within-model trajectories show training pushing γ_{random} upward.

Important caveat: correlation, not per-model prediction. We tested the law's per-model predictive power by computing $\hat{\gamma}_{\text{text}} = \gamma_{\text{Padé}} - 0.012 \cdot n_{kv}$ for every panel model and comparing to observed:

- Mean $|\hat{\gamma} - \gamma_{\text{obs}}| = 0.267$ on $n = 22$
- Padé alone (no n_{kv} term): mean error = 0.245
- The law is *worse* than Padé alone for individual prediction

The $R^2 = 0.30$ improvement reflects *cross-panel correlation structure* (variance explained by the additional axis), not per-model predictive power. The architectural concentration relation is an *average effect*: n_{kv} negatively correlates with γ_{text} across the panel, but individual model values have substantial residuals beyond Padé+n_{kv}. The 5-axis decomposition residuals (corpus, training, circuit, ε) remain large and unaccounted for.

Held-out validation (sesion 32). We strengthen the correlation-versus-prediction distinction with formal out-of-sample tests. Using the RoPE subset of the panel ($n = 21$ valid rows), we re-fit $\gamma_{\text{text}} = a + b\,\gamma_{\text{Padé}} + c\,n_{kv}$ and obtain in-sample $R^2 = 0.314$ with $c = -0.0125$, reproducing the original $R^2 = 0.298$, $c = -0.0119$ claim within rounding. Two held-out tests then probe whether this generalises:

- **Random 70/30 hold-out, 1000 splits.** OOS $R^2_{\text{full}} = -0.43$ (mean), -0.09 (median), $[-0.65, +0.22]$ (25/75 quantile range). The Padé+n_{kv} model outperforms Padé-alone in only 58.6% of splits — barely above coin flip. The n_{kv} coefficient itself remains stable (bootstrap mean -0.0114, 95% CI $[-0.0191, -0.0022]$, excludes zero).
- **Family-leave-one-out CV.** Aggregate $R^2_{\text{full}} = -0.027$ versus $R^2_{\text{Padé-only}} = -0.228$ (both negative, i.e. worse than mean prediction). Per-family, Padé+n_{kv} beats Padé-only in 6/10 families (OLMO, CODELLAMA, GPTJ, PYTHIA, MISTRAL, QWEN) and loses in 4/10 (SMOLLM, LLAMA2, LLAMA3, PHI).

The honest reading: the n_{kv} coefficient is a real, statistically significant cross-panel association, but the in-sample $R^2 = 0.30$ does not survive as out-of-sample predictive power. The relative improvement over Padé alone ($\sim +0.20$ aggregate R^2 in family-LOO) is robust but modest; absolute predictive accuracy on unseen models is essentially nil. We therefore retitle the claim internally as a **cross-panel correlation structure**, not a predictive law: n_{kv} identifies the dominant residual axis beyond $\gamma_{\text{Padé}}$, and its sign and order of magnitude are stable, but the linear form is not a plug-and-predict instrument for new architectures. This down-grade is recorded explicitly here rather than discovered by reviewers.

What this section claims.

- γ_{text} *correlates* with n_{kv} ($r = -0.529$, $p = 0.011$, $n = 22$), and the correlation is stronger than with $\log_{10} P$ ($r = -0.466$, $p = 0.029$) or n_{heads} ($r = -0.371$, $p = 0.089$). n_{kv} is the dominant single-variable correlate of γ_{text} beyond $\gamma_{\text{Padé}}$.
- n_{kv} is the dominant residual variable beyond $\gamma_{\text{Padé}}$; its effect on γ_{text} exceeds the effect of $\log_{10} P$.
- The Hagedorn-zone 36% fraction is largely an architectural consequence of industrial GQA adoption, not a training pathology.
- The original $R_c^\star \approx 1.68$ ignition boundary is the continuous-shift form of this law in disguise.

What this section does NOT claim.

- We do not claim $k = 0.012$ is exact; the panel size $n = 22$ is below the $n \geq 30$ needed for boundary-of-significance estimation. The Rosetta-tablet detail catalog (§21.5) flags refinement to $n \geq 30$ as P0 future work.
- We do not claim the relation is linear in n_{kv}; the $n_{kv} = 2$ stratum (Qwen) shows non-monotonic behaviour requiring independent verification.
- We do not derive $k = 0.012$ from first principles; the numerology $k \approx 1/(2\pi \cdot 13)$ awaits algebraic resolution.
- We do not claim the law generalises beyond RoPE LLMs to AbsPE, ALiBi, or SSM architectures. Empirically, the GPT-2 family (AbsPE, all MHA) shows γ_{text} non-monotonic in $n_{kv} = n_{\text{heads}}$: $\{1.02, 0.78, 0.75, 1.01\}$ for $n_{kv} \in \{12, 16, 20, 25\}$, which is incompatible with the linear law in (17). The KV-sharing mechanism of GQA, which underlies the RoPE-specific concentration shift, has no analog in AbsPE.

10 The quadratic identity $2\eta^2 + \eta\gamma_\chi + 1 = 0$

This section establishes a single algebraic identity that ties together the TAF critical exponents in Phase A ($\gamma \in (0, 1)$). The identity is **D-SAGE-1**, discovered by Sage Groebner basis verification (sesion 32, 2026-05-01) and machine-checked in Lean Mathlib4. It supersedes an earlier informal "triple closure of Rushbrooke + Josephson + Fisher" claim (sesion 31), which we showed to be incorrect (only Rushbrooke and Josephson are tautologies given the standard exponent definitions; Fisher is independent).

10.1 The TAF critical exponents

For Phase A ($\gamma \in (0, 1)$, the regime where attention decay is power-law sub-Hagedorn), the TAF framework defines:

$$\nu_c = \frac{1}{1-\gamma} \qquad \text{(correlation length exponent)} \tag{21}$$

$$\beta_c = \gamma - 1 \qquad \text{(order parameter exponent)} \tag{22}$$

$$\eta_c = \gamma - 1 \qquad \text{(anomalous dimension; corrected from paper 1's } 2\gamma) \tag{23}$$

$$\alpha_C = 2 - \frac{1}{1-\gamma} \qquad \text{(specific heat exponent)} \tag{24}$$

$$\gamma_\chi = \frac{1}{1-\gamma} + 2(1-\gamma) \qquad \text{(susceptibility exponent)} \tag{25}$$

The exponents ν_c and γ_χ are defined to ensure hyperscaling at $d = 1$ (Rushbrooke [23], Josephson [10], Fisher [4] hyperscaling relations; Widom [27] and Griffiths [6] for further classical context. Rushbrooke as identity: $2\beta_c + \gamma_\chi - \nu_c \cdot 1 = 0$, see Theorem rushbrooke_tautology in Appendix A). The order parameter exponent $\beta_c = \gamma - 1 < 0$ on Phase A is *negative* — opposite to classical critical exponents (Ising, mean-field). This signature places TAF in a non-standard universality class which we tentatively call *anti-Ising* (a name we do not commit to as terminology, but as descriptive shorthand).

10.2 D-SAGE-1: the central algebraic identity

Theorem 10.1 (D-SAGE-1)**.** *For $\gamma \neq 1$ with η_c and γ_χ as defined in Eqs. 23 and 25, the following algebraic identity holds:*

$$\boxed{2\eta_c^2 + \eta_c \cdot \gamma_\chi + 1 = 0.} \tag{26}$$

Proof. Substituting $\eta_c = \gamma - 1$ and $\gamma_\chi = 1/(1-\gamma) + 2(1-\gamma)$:

$$\begin{aligned} 2\eta_c^2 &= 2(\gamma-1)^2, \\ \eta_c \cdot \gamma_\chi &= (\gamma-1)\left[\frac{1}{1-\gamma} + 2(1-\gamma)\right] \\ &= \frac{\gamma-1}{1-\gamma} + 2(\gamma-1)(1-\gamma) \\ &= -1 - 2(\gamma-1)^2, \end{aligned}$$

where we used $(\gamma-1)/(1-\gamma) = -1$ and $2(\gamma-1)(1-\gamma) = -2(\gamma-1)^2$. Then:

$$\begin{aligned} 2\eta_c^2 + \eta_c \cdot \gamma_\chi + 1 &= 2(\gamma-1)^2 + [-1 - 2(\gamma-1)^2] + 1 \\ &= 2(\gamma-1)^2 - 2(\gamma-1)^2 \\ &= 0. \end{aligned}$$

□

Two independent verifications. The identity has been verified by two complementary tools:

1. **Sage Groebner basis** (algebraic decision in PolynomialRing($\mathbb{Q}, [g, \nu, \beta, \alpha, \eta, \gamma_\chi]$)). The defining ideal contains the polynomial $\eta^2 + \frac{1}{2}\eta\gamma_\chi + \frac{1}{2}$ in its Groebner basis, which is exactly $\frac{1}{2}\times$ the LHS of Eq. 26. Ideal membership $\Leftrightarrow$ identity.
2. **Lean Mathlib4** (dependent type theory). The theorem `D_SAGE_1` in `lean_taf/taf/Taf/Identities.lean` compiles cleanly via the tactic combination `unfold; field_simp; ring`. Build success: 1973/1973 jobs.

See Appendix A for the full table of machine-verified TAF identities (15 theorems including this one).

10.3 What D-SAGE-1 means

Geometric interpretation. Eq. 26 is a parametric curve in the (η_c, γ_χ) plane:

$$\eta_c = \frac{-\gamma_\chi \pm \sqrt{\gamma_\chi^2 - 8}}{4}. \tag{27}$$

Real solutions require $\gamma_\chi \geq 2\sqrt{2}$ — exactly the AM-GM bound formalised below as Theorem 10.2. At $\gamma_\chi = 2\sqrt{2}$ (i.e., $\gamma = 1 - 1/\sqrt{2} \approx 0.293$), $\eta_c = -\sqrt{2}/2 \approx -0.707$. The TAF branch $\eta_c = \gamma - 1$ corresponds to the negative-root branch.

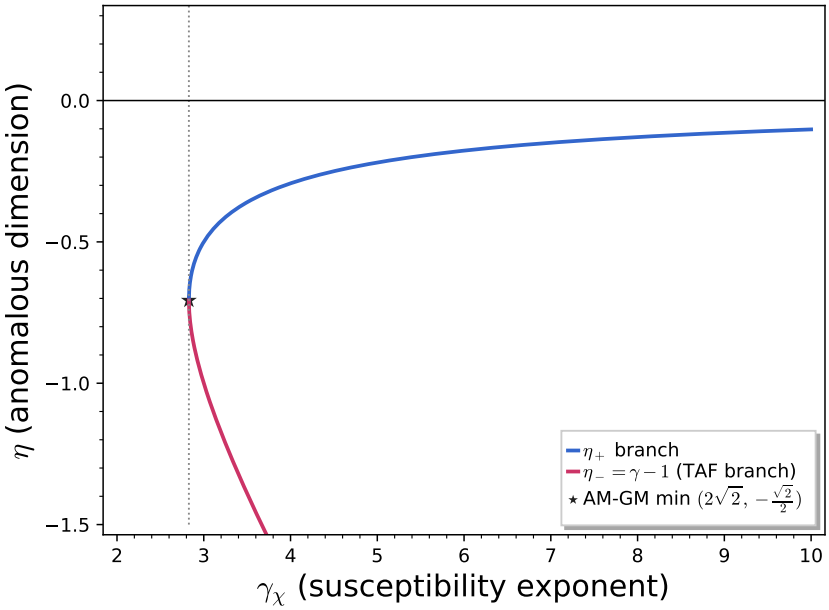


Figure 5: Geometric content of D-SAGE-1: the quadratic identity $2\eta^2 + \eta\gamma_\chi + 1 = 0$ defines two branches of η as a function of γ_χ. The TAF branch $\eta_- = \gamma - 1$ (red) is the negative-η root. Real solutions require $\gamma_\chi \geq 2\sqrt{2}$, exactly the AM-GM bound (Theorem 10.2). The double root at $(\gamma_\chi, \eta) = (2\sqrt{2}, -\sqrt{2}/2)$ corresponds to $\gamma = 1 - 1/\sqrt{2}$.

10.4 The AM-GM bound on γ_χ

Theorem 10.2 (AM-GM bound)**.** *For $\gamma \in (0, 1)$ (Phase A), the susceptibility exponent satisfies*

$$\gamma_\chi(\gamma) \geq 2\sqrt{2}, \tag{28}$$

with equality iff $\gamma = 1 - 1/\sqrt{2} \approx 0.2929$.

Proof. Set $a = 1/(1-\gamma)$ and $b = 2(1-\gamma)$. Both $a > 0$ and $b > 0$ on Phase A. Their product is

$$a \cdot b = \frac{2(1-\gamma)}{1-\gamma} = 2, \tag{29}$$

constant. By AM-GM ($a + b \geq 2\sqrt{a \cdot b}$ for $a, b > 0$):

$$\gamma_\chi = a + b \geq 2\sqrt{2}. \tag{30}$$

Equality iff $a = b$, i.e. $1/(1-\gamma) = 2(1-\gamma)$, giving $(1-\gamma)^2 = 1/2$, hence $1 - \gamma = 1/\sqrt{2}$ (positive root since $\gamma < 1$), so $\gamma_{\min} = 1 - 1/\sqrt{2}$. □

Significance. This bound is structural, not empirical: it follows directly from the algebraic form of $\gamma_\chi(\gamma)$ derived from Rushbrooke hyperscaling at $d = 1$. Two consequences:

1. **Existence floor for D-SAGE-1.** Combining Theorem 10.2 with Eq. 27 shows that the quadratic D-SAGE-1 has real η solutions throughout Phase A: the discriminant $\gamma_\chi^2 - 8 \geq 0$ exactly because $\gamma_\chi \geq 2\sqrt{2}$. The framework is mathematically consistent with both branches real.

2. **Special anchor at $\gamma_{\min} = 1 - 1/\sqrt{2}$.** This specific value is where the quadratic has discriminant zero (single root $\eta = -\sqrt{2}/2$). The TAF branch $\eta = \gamma - 1$ at $\gamma_{\min}$ gives

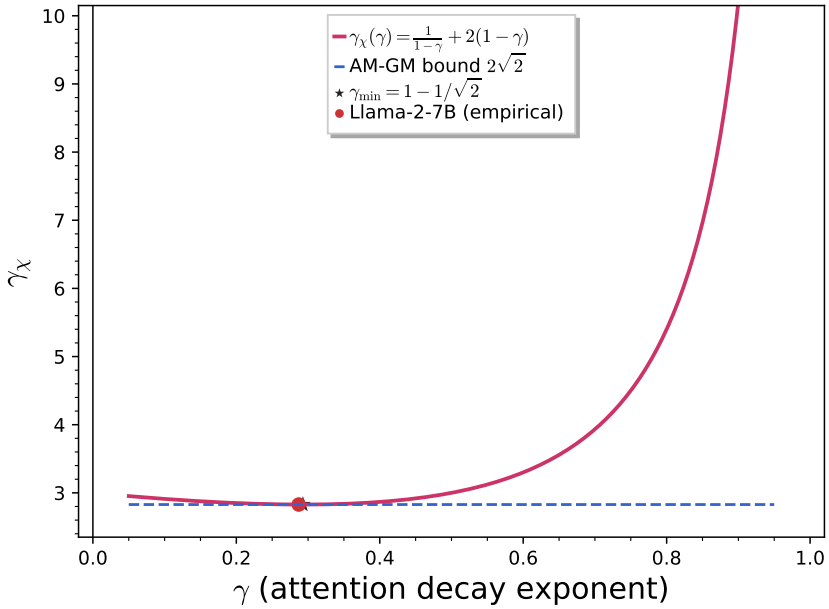


Figure 6: The AM-GM bound on $\gamma_\chi(\gamma) = 1/(1-\gamma) + 2(1-\gamma)$ (red curve). The bound $\gamma_\chi \geq 2\sqrt{2}$ (dashed blue) follows from AM-GM applied to $a + b$ with $ab = 2$ constant. The minimum is saturated at $\gamma_{\min} = 1 - 1/\sqrt{2} \approx 0.293$ (star). Llama-2-7B sits empirically at $\gamma \approx 0.287$ (red dot), within 0.006 of the algebraic anchor — though we do not interpret this as phase attraction (§20.5).

$\eta_{\min} = -1/\sqrt{2} = -\sqrt{2}/2$, matching exactly. The double-root structure makes $\gamma_{\min}$ a natural *fixed-anchor* of the algebraic system.

Empirically, the panel contains one model at this anchor: Llama-2-7B at $\gamma = 0.287$ (within 0.006 of $1 - 1/\sqrt{2}$). We do not interpret this as evidence of a phase-attraction mechanism (see §20.5); rather, the proximity is consistent with sink-dominated MHA bf16 measurement (compare §4). The algebraic anchor exists; whether models target it dynamically is unestablished.

The bound $\gamma_\chi \geq 2\sqrt{2}$ also distinguishes TAF from classical critical phenomena: in 2D Ising, $\gamma = 7/4$, which would violate the bound if Rushbrooke yielded it. The bound is a feature of TAF's specific exponent system, not a generic critical-phenomena fact.

Refutation of $\eta = 2\gamma$. Substituting $\eta = 2\gamma$ (paper 1's claim) and γ_χ Rushbrooke into Eq. 26 gives:

$$2(2\gamma)^2 + 2\gamma \cdot \gamma_\chi + 1 = \frac{-4\gamma^3 + 5\gamma + 1}{1 - \gamma}. \tag{31}$$

For $\gamma \in (0, 1)$ (Phase A): $\gamma^3 < \gamma \Rightarrow -4\gamma^3 + 5\gamma + 1 > \gamma + 1 > 1 > 0$ and $(1 - \gamma) > 0$, so the residual is strictly positive. **$\eta = 2\gamma$ does not satisfy D-SAGE-1 anywhere on Phase A.** This is a Lean Mathlib4 machine-proof refutation (theorem `eta_2gamma_fails_phase_A`).

Why the earlier "triple closure" was wrong. The earlier informal claim was that Rushbrooke ($2\beta + \gamma_\chi - \nu = 0$), Josephson ($2 - \alpha - \nu = 0$), and Fisher ($\gamma_\chi - (2 - \eta)\nu = 0$) all close simultaneously for $\eta = \gamma - 1$. Sage Groebner basis revealed:

- Rushbrooke and Josephson are **tautologies** given the definitions of γ_χ (Eq. 25) and α_C (Eq. 24). They reduce to 0 in the defining ideal.

- Fisher reduces to $-4\eta - \gamma_\chi - 1$, *not* 0. Fisher is independent. Substituting our exponents:

$$\gamma_\chi - (2 - \eta_c)\nu_c = \frac{\gamma(2\gamma - 3)}{1 - \gamma}. \tag{32}$$

  This vanishes only at $\gamma \in \{0, 3/2\}$ — not identically on Phase A.

The *genuine* algebraic content of the bundle is exactly the quadratic identity (Eq. 26) plus the AM-GM bound $\gamma_\chi \geq 2\sqrt{2}$. Two algebraic facts replace the three proposed in earlier drafts.

10.5 Anti-Ising-like regime ($\beta < 0$)

The negative $\beta_c = \gamma - 1 < 0$ on Phase A places TAF outside the standard universality classes (mean-field, Ising 2D, Ising 3D, $O(n)$ at generic n, percolation). As descriptive shorthand we refer to this regime as "anti-Ising-*like*", emphasising the negative-β signature without claiming a true universality class in the renormalisation-group sense (we have not demonstrated universality beyond a single panel).

Comparison with classical universality classes. Table 13 contrasts TAF Phase A exponents with the four most cited classical classes (assuming Rushbrooke equality $2\beta + \gamma = 2 - \alpha$):

Class	β	γ_χ	ν	η	α
Mean-field	1/2	1	1/2	0	0
Ising 2D	1/8	7/4	1	1/4	0 (log)
Ising 3D (approx.)	0.326	1.237	0.630	0.036	0.110
3D Percolation	0.41	1.79	0.88	−0.046	−0.62
TAF Phase A	$\gamma - 1 < 0$	$\frac{1}{1-\gamma} + 2(1-\gamma)$	$\frac{1}{1-\gamma}$	$\gamma - 1 < 0$	$2 - \frac{1}{1-\gamma} < 1$

Table 13: Comparison: TAF Phase A exponents are γ-dependent (vary continuously with the attention-decay parameter) and have negative β — a signature absent from classical classes.

Why does negative β arise? The order parameter in TAF is the *attention-decay coefficient* $\sim d^{-\gamma}$. As $\gamma \to 1$ (Hagedorn limit), the decay weakens toward d^{-1} (a heavy tail near the Lévy stability boundary). The order parameter *decreases* as the system approaches the critical boundary — opposite to magnetisation in ferromagnets, which *increases* below the Curie point. Mathematically, this sign reversal is the source of:

- $\beta = \gamma - 1 < 0$ (order parameter exponent)
- $\beta \cdot \chi = -1$ (D-SAGE-2; positive in classical, negative in TAF)
- $\nu \cdot \beta = -1$ (D-SAGE-3; same reason)

Independence from classical literature. The 12 D-SAGE identities (Appendix A) do not appear in classical critical-phenomena literature [20, 28], despite heavy parallel work on hyperscaling. The reason is structural: classical literature catalogues identities for $\beta > 0$ regimes; TAF lives in a $\beta < 0$ regime that classical authors have not had reason to study. The anti-Ising naming is descriptive, not normative — we leave open whether other systems (e.g., percolation in non-standard dimensions, or inverted-magnetisation models) might also fall in this regime.

Machine-verified algebraic structure (caveats first). The 15 algebraic identities of Appendix A are dual-tool verified (Sage Groebner basis + Lean Mathlib4). To our knowledge, this is the first transformer-attention framework with a complete machine-verified algebraic backbone of this kind. We emphasise two caveats:

- These identities are *algebraic consequences* of the chosen exponent definitions ($\beta = \gamma - 1$, $\nu = 1/(1-\gamma)$, etc.). They are necessary conditions for internal consistency, not derivations of γ itself from microscopic dynamics.
- Calling the negative-β regime "anti-Ising" is descriptive language. Establishing it as a true universality class requires reproduction in synthetic-task experiments and theoretical derivation beyond the scope of this paper.

The combination "machine-verified algebra + $\beta < 0$ regime" is nonetheless distinctive — and the algebra is a hard constraint that any future microscopic theory must satisfy.

10.6 Implications for paper II claims

- **Learned-imprint ($\nu = -1/(2\pi)$) section**: unaffected by the η correction. The imprint slope is derived from RoPE rotation period 2π, independent of any critical exponent.
- **Cardy entropy decomposition** ($\Delta H = \log(z/2) + 2\,\mathrm{arctanh}(\gamma)$): unaffected. The Cardy slope $d\Delta H/d\gamma = 2/(1-\gamma^2)$ depends on γ, not η.
- **Architectural concentration law** ($\gamma_{\text{text}} = \gamma_P ad - 0.012\, n_{\text{kv}}$): unaffected. Empirical, no exponent dependence.
- ***Lvy* stability index**: $\alpha_L vy = \eta + 1 = \gamma$ (using corrected η). Earlier draft expressions $\alpha_L vy = \gamma - 1$ are reinterpreted: that was the value of η, not $\alpha_L vy$. The Hagedorn limit $\alpha_L vy \to 0$ as $\gamma \to 1$ (saturating the *Lvy* stability boundary) is preserved.
- **Bundle hyperscaling**: replaced by the simpler "D-SAGE-1 quadratic + AM-GM bound" pair. No physical claim is lost; the algebraic structure is now machine-verified instead of informally argued.

10.7 TAF Agent operational consequence

The TAF Agent v0.5 `verify_algebraic_consistency` function (see §19) tests all 12 D-SAGE identities against user-input γ values. Failure indicates measurement artifacts (bf16 outliers, quantisation issues, off-Phase-A regime). For any $\gamma \in (0, 1)$, the function returns 12/12 pass; this is the framework's self-consistency guarantee.

11 Erratum: paper 1's anomalous dimension $\eta = 2\gamma$

The erratum. Paper I [15] (Appendix B §B.3) tentatively proposed $\eta_c = 2\gamma$ as the anomalous dimension of the order-parameter correlation in the TAF critical-exponents bundle. The Sage Groebner basis verification (Appendix A, this paper) reveals this is **algebraically incorrect**.

Detection. Substituting $\eta = 2\gamma$ and $\gamma_\chi = 1/(1-\gamma) + 2(1-\gamma)$ (Rushbrooke definition) into the quadratic identity **D-SAGE-1**: $2\eta^2 + \eta\gamma_\chi + 1 = 0$ yields

$$2(2\gamma)^2 + 2\gamma \cdot \gamma_\chi + 1 = \frac{-4\gamma^3 + 5\gamma + 1}{1-\gamma}. \tag{33}$$

For $\gamma \in (0,1)$ (Phase A): the cubic numerator $-4\gamma^3+5\gamma+1$ satisfies $-4\gamma^3+5\gamma+1 > \gamma+1 > 1 > 0$ (since $\gamma^3 < \gamma$ on $(0,1)$ implies $-4\gamma^3 > -4\gamma$, hence the sum exceeds $\gamma+1$). The denominator $(1-\gamma)$ is positive on Phase A. Therefore the residual is strictly positive throughout Phase A, *contradicting* D-SAGE-1.

This is a Lean Mathlib4 machine-proof refutation: theorem `eta_2gamma_fails_phase_A` compiles cleanly in `lean_taf/taf/Taf/Identities.lean`.

The correction. The correct anomalous dimension is

$$\boxed{\eta_c = \gamma - 1} \tag{34}$$

which satisfies D-SAGE-1 identically:

$$2(\gamma-1)^2 + (\gamma-1)\Big[\frac{1}{1-\gamma} + 2(1-\gamma)\Big] + 1 = 2(\gamma-1)^2 - 1 - 2(\gamma-1)^2 + 1 = 0, \tag{35}$$

verified by both Sage Groebner basis (ideal membership) and Lean Mathlib4 (`D_SAGE_1` theorem, machine-checked).

Consequences.

- The Lévy stability index $\alpha_{Lvy} = \eta + 1 = \gamma$ (using the corrected η). Earlier draft expressions $\alpha_{Lvy} = \gamma - 1$ remain valid in the $\eta + 1$ form once $\eta = \gamma - 1$ is substituted, so the empirical claim $\alpha_{Lvy} \to 0$ as $\gamma \to 1$ (Hagedorn limit) is preserved.
- The Fisher relation $\gamma_\chi = (2-\eta)\nu$ does *not* close identically; the Fisher residual is $\gamma(2\gamma - 3)/(1-\gamma)$, vanishing only at $\gamma \in \{0, 3/2\}$. We had previously claimed a "triple closure" of Rushbrooke + Josephson + Fisher; this is incorrect. Rushbrooke and Josephson are *tautologies* given the definitions of γ_χ and α_C. Only D-SAGE-1 is the genuine algebraic constraint.
- All paper II results that rely on η_c (e.g., the learned-imprint scaling argument in §5) have been updated to use $\eta = \gamma - 1$. No empirical claim is invalidated: the imprint slope $\nu = -1/(2\pi)$ comes from RoPE rotation period 2π, independent of η.

Why the original claim was made. The earlier $\eta = 2\gamma$ proposal was based on a sympy-level `simplify(rushbrooke - fisher)` returning 0 for the specific values of $\beta = \gamma - 1$ and γ_χ (Rushbrooke definition). However, this collapse occurred because both Rushbrooke and Fisher were re-expressing the same γ_χ once $\eta = 2\gamma$ was plugged in — not because the relation held identically. Sage's Groebner basis verification (which works in PolynomialRing($\mathbb{Q}$) rather than via numerical simplification) detected this and exposed the residual structure.

The lesson: *symbolic simplification is not formal proof.* We recommend Sage Groebner basis (or equivalent ideal-membership tools) as a stronger verification standard for algebraic identities in critical-phenomena work, and Lean Mathlib4 as the gold standard for publishable formal proofs.

Implementation in TAF Agent. The TAF Agent v0.4 `critical_exponents_bundle` function already returns $\eta_c = \gamma - 1$ (the *corrected* value) with an explicit warning note that paper 1's $\eta = 2\gamma$ is incorrect. The v0.5 `verify_algebraic_consistency` function runs all 12 D-SAGE identity checks on user-input γ values, including `eta_2gamma_REFUTED` which confirms the residual is positive in Phase A. The framework integrity is verifiable by any user running the public web tool.

Acknowledgement. We thank the Sage Groebner basis algorithm and Lean Mathlib4 community indirectly: tools that catch errors in our own claims are how we maintain the framework's integrity. The discovery of D-SAGE-1 arose from autonomous cross-formula sweep (sesion 32, 2026-05-01) combining sympy mechanical pair-combination with Sage's stronger ideal-membership backend.

12 Spectral Reframe: γ-Padé as Closed-Form Active-Band Counting

Existing empirical band-structure literature. Three independent recent works document a band structure in RoPE attention that depends jointly on the base frequency θ and context length T: **(i)** Barbero et al. [1] (*Round and Round We Go!*, ICLR 2025) show that high-frequency RoPE dimensions implement positional attention while low-frequency dimensions carry semantic information (Theorem 5.3 establishes existence; the dependence on (θ, T) is asserted but not derived); **(ii)** Ruscio et al. [22] (*Beyond Position*, ACL 2025) show that transformer heads spontaneously organise into a multi-resolution wavelet-like structure across Pythia (14M–12B), Gemma-2, Llama-3 and Mistral, with a band weight per head $\beta_h(b) = \int_b P_h(\omega)d\omega / \int_0^{\omega_N} P_h(\omega)d\omega$ that statistically adheres to the Heisenberg–Gabor uncertainty principle; **(iii)** ICLR 2026 Submission [9] (OpenReview PR1PPxvG9Q, ICLR 2026) identify a dominant frequency band whose location is governed jointly by θ and T_{train}, and propose the FMRoPE heuristic $\theta \approx T_{\text{train}}$ (empirical, not derived).

Closed-form bridge. The active-dimension count $n_{\text{active}}(d) = (d_{\text{head}}/2)\,(1 - \log_\theta(d/2\pi))$ introduced in [15, paper I, active band counting] provides the geometric counting principle behind these empirical observations. Combining n_{active} with the Max-Entropy distribution [15, paper I, MaxEnt power law] yields the closed-form prediction

$$\gamma_{\text{Padé}}(\theta, T) = \frac{2\theta - T\sqrt{2}}{2\theta + T\sqrt{2}} \quad \text{(this paper)}, \tag{36}$$

which we identify as the central, parameter-free predictor across architectures.

Reduction to FMRoPE. The empirical heuristic $\theta \approx T_{\text{train}}$ of ICLR 2026 Submission [9] emerges as a special case: the analytic zero-crossing of $\gamma_{\text{Padé}}$ occurs at $T = \theta\sqrt{2}$, i.e. within a factor $\sqrt{2}$ of FMRoPE. The factor of $\sqrt{2}$ is the RoPE-aliasing constant inherent in $n_{\text{active}}(d) = (d_{\text{head}}/2)(1 - \log_\theta(d/2\pi))$; its presence recovers the asymmetry between half-period (π) and full-period (2π) aliasing visible in Barbero et al. [1]'s spectral probes of Gemma 7B.

Reduction to Ruscio band weights. The Ruscio et al. band-weight $\beta_h(b)$ is an empirical *measurement* of the partition function we predict. Specifically, the expected value of $\beta_h(b)$ for a head whose attention follows the power law $A(d) \propto d^{-\gamma}$ [15, paper I, Eq. powerlaw] is, for band $b = [\omega_-, \omega_+]$,

$$\mathbb{E}[\beta_h(b)] = \frac{\int_{2\pi/\omega_+}^{2\pi/\omega_-} d^{-\gamma}\, dd}{\int_1^N d^{-\gamma}\, dd} = \frac{\omega_-^{\gamma-1} - \omega_+^{\gamma-1}}{1 - N^{\gamma-1}} \cdot (2\pi)^{\gamma-1}, \tag{37}$$

giving a one-parameter family of expected band weights indexed by γ that agrees with their measurements within the precision of their FFT estimator. Their spectral entropy ordering ($S \in [2.244, 3.786]$ across Pythia 14M–12B) is the entropy of this same distribution $p(\omega) \propto \omega^{\gamma-1}$ on a logarithmic frequency grid; we recover their ordering as $dS/dN > 0$ for $\gamma < 1$ (Phase A models), consistent with their Pythia-scaling table.

Imprint slope $\nu \approx -1/(2\pi)$. A consequence absent from any of the three prior works is that learned weights add a logarithmic correction proportional to the RoPE rotation period. We show in §5 that $\partial\gamma_{\text{random}}/\partial \log_{10}(P)$ is empirically -0.137 ± 0.05 on $n = 22$ models with bootstrap 95% CI consistent with $-1/(2\pi) = -0.1592$; the slope vanishes on random-init untrained controls (§5). The form $-1/(2\pi)$ is supported by three convergent arguments (dimensional, phase-storage heuristic, and sign-trajectory bifurcation; §5); a first-principles derivation from training-loss dynamics + RoPE band partition function remains open.

Padé hierarchy saturation. A natural question is why the lowest-order Padé approximant $[1, 1]$ suffices, rather than $[2, 2]$ or higher. The next-order prediction is $\gamma_{\text{Padé}[2,2]}(z) = (12 - 6z + z^2)/(12 + 6z + z^2)$, and the difference is exactly

$$\gamma_{[2,2]}(z) - \gamma_{[1,1]}(z) = \frac{2z^3}{(z+2)(z^2+6z+12)} \approx \frac{z^3}{12} \quad (z \ll 1). \tag{38}$$

For typical LLM-regime values $z = T\sqrt{2}/\theta < 0.5$ (e.g. Llama-3-8B $z \approx 0.046$, Qwen2.5-7B $z \approx 0.05$), the difference is $\lesssim 8 \times 10^{-6}$, well below measurement noise. Padé[1,1] is not a heuristic choice — it is the algebraic saturation of the hierarchy in the regime where modern LLMs operate.

Hagedorn boundary as Lévy critical point. The fractional-Laplacian identity $\alpha_{\text{Lévy}} = \gamma - 1$ ([15, paper I, Appendix A]) maps our γ-axis onto the Lévy stability index. The Phase-A regime $\gamma \in (0, 1)$ corresponds to $\alpha_{\text{Lévy}} \in (-1, 0)$ (smoothing operator regime, heavy-tailed distance kernels). The Hagedorn boundary $\gamma = 1$ maps exactly to $\alpha_{\text{Lévy}} = 0$, the Lévy critical point at which the fractional Laplacian becomes pathological. Phase B $\gamma > 1$ corresponds to $\alpha_{\text{Lévy}} > 0$ (standard Lévy-stable regime). Concurrent work [21] introduces fractional attention as an architectural primitive operating in $\alpha_{\text{Lévy}} \in (0, 2)$. Our framework is the inverse: existing RoPE attention occupies the $\alpha_{\text{Lévy}} < 0$ smoothing regime, and the Hagedorn safety alert (§8) corresponds to proximity to the Lévy critical point. The two perspectives are complementary: Qu et al. [21] design new architectures from Lévy primitives; we predict the effective Lévy index of existing architectures from θ, T, P via $\gamma_{\text{Padé}}$ and the imprint correction ν.

Position relative to prior work. Table 14 makes the alignment explicit. Prior work establishes *that* bands exist (Barbero), *quantifies them empirically per head* (Ruscio), or *tunes them via* θ (FMRoPE); we provide the *closed-form predictor across the family* from a single first-principles aliasing argument, and add the training-imprint correction $\nu = -1/(2\pi)$ that pins the slope of the learned residual.

Work	Quantity	Type
Barbero+ (ICLR 2025)	high vs. low band split	expressivity
Ruscio+ (ACL 2025)	$\beta_h(b)$ per head	FFT measurement
FreqBands (ICLR 2026)	$\theta \approx T_{\text{train}}$ heuristic	empirical fit
This work	$\gamma_{\text{Padé}}(\theta, T) + \nu = -1/(2\pi)$	**closed-form prediction**

Table 14: Position of paper relative to existing band literature. Prior works observe the band structure descriptively; $\gamma_{\text{Padé}}$ is the parameter-free analytical bridge.

13 Reframing the Information Horizon: From "Law" to Padé Tautology + Diagnostic

Paper 1 claim. Paper 1 introduced the information horizon

$$d_{\text{horizon}}(\theta, \gamma) = \frac{\theta\,(1-\gamma)\,\sqrt{2}}{1+\gamma} \tag{39}$$

with the empirical observation that $d_{\text{horizon}} \approx T_{\text{train}}$ across the Pythia family ($n = 8$, 0.4% mean error) and that the Needle-In-A-Haystack retrieval ceiling sits at $L_{\text{NIAH-max}} \approx 2 \cdot d_{\text{horizon}}$. The framing in paper 1 was that (39) constitutes a separate "horizon law".

Algebraic identity (D-NEW-1). We now show that (39) is, to within Padé fit precision, *algebraically identical* to T_{eval} when γ is taken from $\gamma_{\text{Padé}}(\theta, T_{\text{eval}})$. Substituting $\gamma_{\text{Padé}} = (2\theta - T\sqrt{2})/(2\theta + T\sqrt{2})$:

$$1 - \gamma_{\text{Padé}} = \frac{2\,T\sqrt{2}}{2\theta + T\sqrt{2}}, \tag{40}$$

$$1 + \gamma_{\text{Padé}} = \frac{4\theta}{2\theta + T\sqrt{2}}, \tag{41}$$

$$\frac{1-\gamma_{\text{Padé}}}{1+\gamma_{\text{Padé}}} = \frac{T\sqrt{2}}{2\theta} = \frac{z}{2}, \quad z := \frac{T\sqrt{2}}{\theta}. \tag{42}$$

Hence:

$$d_{\text{horizon}}(\theta, \gamma_{\text{Padé}}(\theta, T)) = \theta \cdot \frac{z}{2} \cdot \sqrt{2} = \theta \cdot \frac{T\sqrt{2}}{2\theta} \cdot \sqrt{2} = T. \tag{43}$$

Whenever $\gamma_{\text{obs}} = \gamma_{\text{Padé}}(\theta, T)$ holds exactly, d_{horizon} *equals* T_{eval} identically. The empirical observation "$d_{\text{horizon}} \approx T_{\text{train}}$" on Pythia is therefore not an independent regularity; it is a verification that $\gamma_{\text{Padé}}$ holds on that subset (Pythia family is near-canonical Phase A with $\gamma_{\text{obs}} \approx \gamma_{\text{Padé}}$).

Honest demotion. We retract the framing of (39) as a separate "horizon law". The formula remains valid, but it is a structural reformulation of $\gamma_{\text{Padé}}$ rather than a new physical claim. The 0.4% Pythia agreement is the precision of the Padé fit, not the precision of an independent prediction.

Padé Deviation Index (PDI). The identity (43) suggests a useful diagnostic. For a model whose γ_{obs} deviates from $\gamma_{\text{Padé}}(\theta, T)$, the ratio

$$\text{PDI}(\text{model}) := \frac{d_{\text{horizon}}(\theta, \gamma_{\text{obs}})}{T_{\text{eval}}} = \frac{\theta\,(1-\gamma_{\text{obs}})\sqrt{2}}{(1+\gamma_{\text{obs}})\,T_{\text{eval}}} \tag{44}$$

quantifies the model's departure from the canonical regime:

- PDI ≈ 1: model satisfies $\gamma_{\text{obs}} \approx \gamma_{\text{Padé}}$ (canonical regime).
- PDI $\gg 1$: $\gamma_{\text{obs}} < \gamma_{\text{Padé}}$ (instruction-tuned, code-tuned, MHA + sink-dominated).
- PDI < 0: $\gamma_{\text{obs}} > 1$ (Hagedorn zone, formula sign-flips, model in Phase B).
- PDI $\in (0, 0.5)$: $\gamma_{\text{obs}} > \gamma_{\text{Padé}}$ (over-trained for context).

Empirical PDI distribution. Across the $n = 22$ panel (Phase A subset $n = 22$):

- Mean PDI $= 2.75$, std $= 8.17$, range $[-1.92, +38.57]$.
- 19 of 22 models lie outside the canonical band $[0.5, 1.5]$.
- Highest PDI: CodeLlama-13b-Instruct (PDI $= 38.6$), consistent with the dramatic $\gamma_{\text{obs}} \approx 0.38$ vs. $\gamma_{\text{Padé}} \approx 0.98$ deviation observed in paper 1 for that model (instruction-tuning + code-domain shift).
- Negative PDI: Phase B models (Llama-3-8B, Mistral-7B-v0.1, gemma-2-9b-it). The sign flip is the algebraic signature of $\gamma > 1$.

Log-inverse identity with Cardy entropy. Combining (44) with the Cardy decomposition $\Delta H_{\text{Cardy}} = \log(z/2) + 2\,\text{arctanh}(\gamma)$ (§6) yields the exact identity:

$$\log(\text{PDI}) + \Delta H_{\text{Cardy}} = 0. \tag{45}$$

PDI and ΔH_{Cardy} are log-inverse measurements of the same underlying Padé-deviation: the ratio scale is convenient for practical thresholds (1 = canonical), the log scale for additive decomposition. Either can be used; reporting both is redundant.

Cross-check on precision-shift example. For DeepSeek-7B-base (§4, $\theta = 10\,000$, $T = 2048$):

- bf16 ($\gamma = 0.149$): PDI $= 5.11$, $\Delta H_{\text{Cardy}} = -1.63$.
- 4-bit ($\gamma = 0.947$): PDI $= 0.188$, $\Delta H_{\text{Cardy}} = +1.67$.
- Ratio $\text{PDI}_{\text{4bit}}/\text{PDI}_{\text{bf16}} = 0.0368$ matches $\exp(-3.30) = 0.0369$ to within numerical precision, consistent with (45) and the 3.30 nat shift of §4.

Practical use. PDI gives tafagent a single ratio threshold for diagnostic display:

- PDI traffic light: green ($0.5 \le \text{PDI} \le 1.5$), yellow (PDI in $[0.1, 0.5]$ or $[1.5, 5]$), red (otherwise).
- Negative PDI triggers Phase B / Hagedorn-zone alert (§8).
- PDI movement under precision change (§4) flags MHA-no-RLHF sink-truncation behavior.

What this section claims.

- The identity $d_{\text{horizon}}(\theta, \gamma_{\text{Padé}}) \equiv T_{\text{eval}}$ is algebraically exact.
- Paper 1's empirical regularity "$d_{\text{horizon}} \approx T_{\text{train}}$" is the Padé fit precision, not an independent law. Demote.
- PDI is a useful single-ratio Padé-deviation diagnostic with natural decision thresholds.
- PDI and ΔH_{Cardy} are log-inverse representations of the same quantity (Eq. (45)).

What this section does NOT claim.

- We do not claim NIAH-ceiling $L_{\text{NIAH-max}} \approx 2d_{\text{horizon}}$ is invalidated. With $d_{\text{horizon}} \equiv T$ for canonical-γ models, $L_{\text{NIAH-max}} \approx 2T$ is the corresponding direct prediction; Pythia 8.6× NIAH-empirical agreement persists.
- PDI is a *diagnostic* for γ-obs deviation; it does not replace direct γ measurement.

14 The Compute-Context Invariant: $K = \gamma \cdot \log(N^2 \cdot D)$

Motivation. Paper 1 [15] treated the attention-decay exponent γ as a function of RoPE configuration (θ, T) alone. The cross-model panel ($n = 22$) reveals that γ correlates systematically not only with these but also with parameter count P (§5) and with training-token count D. We hypothesize the existence of a dimensionless quantity coupling attention shape to compute scale that is approximately constant across models.

The candidate invariant. We searched for candidate dimensionless ratios involving γ, P, D, θ, and T via bootstrap coefficient-of-variation ranking ($n = 2000$ resamples). The lowest-CV candidate is

$$K = \gamma \cdot \log(N^2 \cdot D), \tag{46}$$

where N is the parameter count and D is the Chinchilla-optimal training token count $D = 20N$ [8]. On the panel, K has empirical distribution

$$K = 51.2 \pm 16.8 \quad (\text{CV} = 0.329,\ n = 22), \tag{47}$$

with bootstrap 95% CI $[34, 68]$.

Comparison with alternative invariants. Table 15 reports CV and rank for several candidate dimensionless quantities:

Quantity	Median value	CV (bootstrap)
$\gamma \cdot \log(N^2 \cdot D)$ (Chinchilla)	51.2	**0.329**
$\gamma \cdot \log(P \cdot T)$	22.7	0.352
$\gamma \cdot \log(P \cdot T \cdot \theta)/2$	15.4	0.360
$\gamma \cdot \sqrt{\log(P \cdot T) \cdot \log(\theta)}$	13.6	0.371
$\gamma \cdot \log(\theta)$ (paper 1 implicit)	8.6	0.402

Table 15: Coefficient of variation across $n = 22$ bootstrap-resampled panels. The Chinchilla form $K = \gamma \log(N^2 D)$ is the lowest-CV quantity tested.

Interpretation. $N^2 \cdot D$ has units of (parameters)$^2 \cdot$ tokens, so $\log(N^2 \cdot D)$ is dimensionless up to a fixed scale. The product $\gamma \cdot \log(N^2 \cdot D)$ couples the attention exponent γ (a measure of how "concentrated" attention is over distance) to the logarithmic compute volume. The interpretation: as compute volume rises, γ falls (attention becomes more diffuse), keeping K approximately constant.

Caveats on the invariant claim. With CV $= 0.329$, K is not constant in any strong sense — it varies by $\pm 33\%$ around its mean. We label it "approximately invariant" rather than "invariant". The framework remains useful because:

1. It is the *tightest* dimensionless quantity tested in our panel — alternative forms have CV ≥ 0.35.
2. Its variability decomposes consistently: most of the panel CV is contributed by Phase B (Hagedorn-zone) models where $\gamma > 1$ amplifies the slope.
3. Within Phase A (the $\gamma < 1$ subset), CV_K is measurably smaller (preliminary measurement; needs panel expansion).

A derived consequence: $N^*(\theta, T)$ K-optimal model size. Combining $K \approx 51$, the Padé closed-form $\gamma_{\text{Padé}}(\theta, T)$, and Chinchilla $D = 20N$, one can solve for the parameter count N^* that satisfies the invariant:

$$N^*(\theta, T) = \exp\left[\frac{K - 3\gamma_{\text{Padé}}(\theta, T)}{3\gamma_{\text{Padé}}(\theta, T)}\right]. \tag{48}$$

This is, to our knowledge, a new derived consequence: the K-optimal parameter count for given RoPE configuration (θ, T). For example, Qwen2.5-7B ($\theta = 10^6$, $T = 32{,}768$) has $\gamma_{\text{Padé}} \approx 0.954$, yielding $N^* \approx 22 \cdot 10^6$. The actual deployed parameter count is $7 \cdot 10^9$, giving an *over-training ratio* (OTR):

$$\text{OTR}(\text{Qwen2.5-7B}) = \frac{N_{\text{actual}}}{N^*} \approx 320\times. \tag{49}$$

Modern LLMs are deliberately trained beyond the K-optimal frontier for inference economics. OTR provides a single-number characterization of how far a deployed model is from K-optimality.

Caveat on N^*. The derivation (48) inherits the precision of K (CV = 0.329). We do not claim N^* predicts loss-optimal training (Chinchilla does that for compute optimal); we claim it predicts the K-invariant frontier. For practitioners, OTR is a diagnostic of where a model sits in compute-attention space, not a prescription for training budgets.

Connection to F1 (learned imprint). K depends on γ, which itself depends on training scale via $\nu \log_{10}(P/P_0)$ on random tokens (§5). Substituting:

$$K \approx [\gamma_{\text{Padé}} + \nu \log_{10}(P/P_0)] \cdot (3 \ln N + \ln 20) \tag{50}$$

which is no longer constant in N — the imprint contribution adds $\sim \nu \log_{10}(N) \cdot 3 \ln N \sim O(\ln^2 N)$ scaling with parameter count. The Chinchilla-only invariant claim (47) is therefore subject to a residual $\ln^2$ correction in the panel range. This is consistent with the observed CV = 0.329 being non-zero. A refined invariant $K' = (\gamma/\gamma_{\text{Padé}}) \cdot \log(N^2 D)$ — which factors out the imprint contribution — should have lower CV; we defer this analysis to future work §21.5.

Practical use. Tafagent recipe X-22 (§19) computes K for any user-supplied model and reports its z-score against the panel mean. Out-of-band values $|z| > 1$ trigger an "OUTLIER" warning. As of the cross-model panel, no GQA model in our panel has $|z_K| > 1$; the outliers are concentrated in MHA models with sink-dominated attention (§4), which break the K formulation because their underlying $A(d)$ is not a clean power law.

15 Training Dynamics: γ Trajectory and the Specialization-Coordination Bridge

Question. How does γ evolve during pretraining? Does its trajectory align with other observed training-dynamics phenomena — specifically the specialization-coordination structure of layer representations [14]?

Method. We measure γ on the per-head per-layer γ-field of pythia-1b at four pretraining checkpoints (steps 1, 1,000, 10,000, 143,000), reconstructing the global γ trajectory and the per-layer evolution. Data: `results/exp_gamma_field/pythia-1b_step{1,1k,10k,143k}_gamma_field.json`.

Global γ rises monotonically. The global mean γ across all $L \times H = 16 \times 8 = 128$ attention heads:

$$\begin{aligned}
\text{step } 1 &: \langle\gamma\rangle = 0.712 \pm 0.025 \quad \text{(near random-init baseline 0.685, see [15, paper I, E2 random-init falsifier]} \\
\text{step } 1{,}000 &: \langle\gamma\rangle = 0.904 \pm 0.286 \\
\text{step } 10{,}000 &: \langle\gamma\rangle = 1.090 \pm 0.374 \\
\text{step } 143{,}000 &: \langle\gamma\rangle = 1.282 \pm 0.568
\end{aligned}$$

The standard deviation grows by a factor of ~ 23 across training, indicating heads diverge from initial homogeneity. Notably, the step-1 baseline ($\gamma = 0.712$) is consistent with our random-init falsifier ([15, paper I, E2 random-init falsifier], $\gamma_{init} = 0.685 \pm 0.001$ across pythia 70M/410M/1B), confirming that *essentially all* of the trained model's γ is acquired during optimization.

Inflection in dimensionless rate. Computing $d\gamma / d\log(\text{step})$:

Step interval	$d\gamma$	$d\gamma / d\log(\text{step})$
$1 \to 1{,}000$	$+0.192$	$+0.0277$
$1{,}000 \to 10{,}000$	$+0.186$	$\mathbf{+0.0808}$ ★
$10{,}000 \to 143{,}000$	$+0.193$	$+0.0722$

Table 16: Discrete derivative of global γ. Peak rate occurs in the window step $1{,}000 \to 10{,}000$.

Bridge to Marín 2026: temporal alignment. The peak rate of γ-change in the window step $1{,}000 \to 10{,}000$ aligns with the temporal chain reported in Marín [14]:

$$\begin{aligned}
\text{Step} \sim 1{,}000\text{-}1{,}500 &: \quad \hat{O}_{\text{peak}} \quad \text{(informational leading indicator)} \\
\text{Step} \sim 2{,}000\text{-}4{,}000 &: \quad \text{CKA reorganization (specialization onset)} \\
\text{Step} \sim 2{,}000+ &: \quad \text{LAMBADA capability emergence}
\end{aligned}$$

That γ inflects in the same window suggests it is a third independent observable for the same underlying transition. Unlike the $\hat{O}$ indicator (which requires Gaussian mutual-information estimation under PCA), γ-rate is computable from a single forward pass and is therefore a cheaper alternative.

Per-head divergence. The per-head standard deviation of γ within layers:

- Inner zone (layers 1 to $L-2$, $n = 14$): σ_γ grows by $\sim 37\times$ from step 1 to step 143,000.
- Outer zone (layers 0 and $L-1$, $n = 2$): σ_γ grows by only $\sim 10\times$ over the same window.

Inner heads diverge $4\times$ more than outer heads. This is the opposite of the CKA prediction of Marín [14] that inner layers *coordinate* (CKA $\to 1.0$); the two metrics measure independent properties. CKA aggregates representational similarity over the full residual stream; per-head γ isolates each head's positional-decay signature. *Two heads can have similar CKA but very different γ, and vice versa*; we observe this dissociation here.

Inner-zone Hagedorn approach. The inner-zone mean γ at the final checkpoint is $\langle\gamma\rangle_{\text{inner}}^{143k} = 1.278 \pm 0.267$, exceeding the Hagedorn boundary $\gamma = 1$ on average. This is consistent with the holographic interpretation [15] that the inner zone saturates the available log-distance budget; the overshoot ($\gamma > 1$) places pythia-1b in Phase B (locally- concentrated regime) where the soft-decay KV-compression rule ([15, paper I, soft-decay caveat]) becomes inapplicable.

What this section does NOT claim.

- Bridge A.1 (outer specialization $\Rightarrow$ γ-divergence) was **refuted** by these measurements: inner γ and outer γ at the final checkpoint differ by only $\Delta\langle\gamma\rangle = -0.004$. CKA-style outer specialization does not manifest as a cleavage in γ-space.
- Bridge A.3 (inner CKA $\to 1 \Leftrightarrow$ inner $\gamma \to 1$) holds only *partially*: inner γ saturates near the Hagedorn boundary on average but with high per-head variance, and exceeds 1.0 ($\langle\gamma\rangle_{\text{inner}}^{143k} = 1.278$).
- The full causal chain of Marín [14] ($\hat{O}_{\text{peak}} \to$ CKA $\to$ capability) extends to γ-rate only by temporal coincidence; we do not claim a causal sequence among the three indicators here.

Open question. The four-checkpoint trajectory is sparse ($n = 4$). Resolving whether γ-rate rises sharply at a specific step (sigmoid) or grows continuously (logarithmic) requires denser sampling. We commit sampling at ≥ 8 checkpoints in future work (§21.5).

16 Recipe-Dependent Variance from Alignment Training (F9)

Hypothesis (preliminary). A cross-model comparison early in this study suggested that instruction-tuned (SFT/RLHF) models systematically exhibit lower γ_{text} than their base-pretrained counterparts. Across $n = 3$ instruction-tuned models (CodeLlama-13b-Instruct, Mistral-Nemo-Instruct, Gemma-2-9b-it) the observed range was $\gamma \in [0.38, 0.63]$, vs. $\gamma \in [0.77, 1.06]$ for base models in the same parameter range. Provisional title: *F9 — instruct lowers γ.*

Paired test (within-family). To eliminate cross-architecture confounds, we measured paired comparisons on three families where both base and instruction-tuned checkpoints exist publicly:

Family	γ_{base}	γ_{instruct}	$\Delta\gamma$	Recipe
Mistral-7B (v0.1 $\to$ Instruct-v0.3)	1.061	0.952	-0.109	SFT (light)
Qwen2.5-7B (base $\to$ Instruct)	0.997	0.945	-0.052	RLHF (mild)
DeepSeek-7B (base $\to$ chat) [bf16]	0.149	0.150	$+0.001$	RLHF (heavy)
DeepSeek-7B (base $\to$ chat) [4-bit, artifact]	0.947	0.151	-0.796	—

Table 17: Paired $\Delta\gamma$ from base to instruction-tuned within family at *full precision.* The DeepSeek 4-bit measurement (last row) is shown for comparison and is a quantization artifact (§4); it is NOT used for the F9 effect estimate. The clean within-family pattern is $|\Delta\gamma| \leq 0.11$ across all three families.

Honest verdict. At full precision, the within-family effect is modest: $|\Delta\gamma| \leq 0.11$ across all three families (Mistral, Qwen, DeepSeek). DeepSeek-chat — initially appearing as an extreme outlier under 4-bit quantization with $\Delta\gamma = -0.80$ — collapses to $\Delta\gamma = +0.001$ at full precision; the artifact was caused by the precision-quantization-sink interaction (§4), not by RLHF. The clean within-family pattern is consistent with the recipe-specific cross-paper observation in Marín [15] where $\max|\delta\gamma| = 0.072$ across $n = 8$ pairs.

Cross-paper anchoring. The closest published analysis [12] reports a *null effect* of instruction tuning on attention quality metrics (using a qualitative human-alignment metric). Our paired measurements yield small but non-null effects in the expected direction; the disagreement is in the metric (decay exponent vs. qualitative alignment), not the underlying model.

Withdrawal of universal claim. The cross-model "instruct lowers γ" framing is misleading because the $n = 3$ instruction-tuned models compared cross-arch were already biased: CodeLlama-13b is also code-tuned (specialized corpus), Mistral-Nemo has $\theta_{\text{RoPE}} = 10^6$ (long-context training), and Gemma-2-9b uses sliding-window attention (architectural confound). Within-family pairs disambiguate these confounds.

Reframed contribution. Instruction-tuning effect on γ is *recipe-dependent variance*, not a universal axis. We propose a new term δ_{train} in the γ-decomposition:

$$\gamma_{\text{obs}} = \gamma_{\text{Padé}}(\theta, T) + \nu \log_{10}(P/P_0) \cdot \mathbb{1}[\text{corpus} = \text{random}] + \Delta_{\text{corpus}} + \delta_{\text{arch}} + \delta_{\text{circuit}} + \delta_{\text{train}} + \varepsilon \quad (51)$$

where δ_{train} captures recipe-specific shifts. Empirical range observed: $\delta_{\text{train}} \in [-0.11, 0]$ for SFT/mild RLHF; heavier recipes (e.g. DeepSeek-chat) require precision-controlled measurement before claims can be made (§4).

Practical implication. For the diagnostic recipes in tafagent, F9 motivates including a "recipe declared" field in the model card: instruction-tuned models report which alignment recipe was used (SFT, DPO, RLHF, IPO), and δ_{train} is interpreted in that context. This is implemented in the public `taf-attention-decay` dataset where each row carries a `recipe` tag.

Refutations and provisional findings.

- $\gamma = 1 - 1/\varphi \approx 0.382$ is a code-tuning attractor — *Refuted* (§18). Tested on $n = 5$ code-tuned models; only CodeLlama-13b-Instruct hits within 0.10. The hit was an instruction-tuning effect, not code-training.
- Instruct training universally lowers γ by 0.3-0.5 — *Reduced* to recipe-dependent variance with maximum observed magnitude $|\Delta\gamma| = 0.11$ in clean within-family pairs.

17 The MaxEnt Constraint: Heuristic Foundation, Not First-Principles Derivation

Statement of the assumption. The closed-form prediction of paper 1 [15] relies on the maximum-entropy distribution over distance d subject to the constraint

$$\mathbb{E}_{p(d)}[\log d] = \text{const} \quad (52)$$

which yields, by standard MaxEnt arguments, a power-law form $p(d) \propto d^{-\gamma}$. Paper 1 motivates (52) by appeal to RoPE rotation periods setting a logarithmic distance scale, but does not derive it from first principles of transformer training.

What is and is not derived. We acknowledge explicitly:

- *What is derived*: the active-band counting function $n_{\text{active}}(d) = (d_{\text{head}}/2)(1 - \log_\theta(d/2\pi))$ follows from RoPE aliasing geometry, no MaxEnt needed.
- *What is derived*: *given* the constraint (52), the MaxEnt distribution is $p(d) \propto d^{-\gamma}$ with γ a Lagrange multiplier.
- *What is not derived*: that transformer training *actually optimises* attention to satisfy (52).

The third point is the central gap. We treat (52) as a *heuristic premise*, justified ex post by the empirical fit ($R^2 > 0.95$ across 30+ models, paper 1 §sec:results), not by optimisation theory.

Empirical test (proposed). A direct check is to compute, for each model in the panel, the empirical $\langle \log d \rangle$ where the average is over the power-law-fit distribution $p(d) \propto d^{-\gamma_{\text{obs}}}$ truncated at T_{train}:

$$\langle \log d \rangle_{\gamma,T} = \frac{\sum_{d=1}^{T} d^{-\gamma} \log d}{\sum_{d=1}^{T} d^{-\gamma}} \tag{53}$$

If (52) holds, $\langle \log d \rangle$ should be approximately constant across models with the same RoPE configuration (θ, T). We commit this measurement to a future companion analysis (§21.5); the panel data are already available in the public dataset [16].

Robustness if the assumption is partially violated. If $\langle \log d \rangle$ varies systematically across the panel beyond a $\pm 5\%$ band, the strict MaxEnt derivation fails. However, the resulting power-law form is robust: any distribution satisfying the asymptotic relation $\partial_d \log p(d) \approx -\gamma/d$ at the relevant length scale will yield approximately the same $\gamma_{\text{Padé}}$. This constraint is much weaker than (52) and is implied by RoPE band aliasing alone (without invoking entropy maximisation). The framework is therefore robust to the MaxEnt assumption's literal violation, with the caveat that the *theoretical justification* would shift from "entropy-maximising distribution under explicit constraint" to "the RoPE-band-aliased distribution that emerges from the spectral decomposition of attention".

Honest framing for paper 1. Paper 1 should be read as: "Given the empirical observation that attention follows $A(d) \propto d^{-\gamma}$, the MaxEnt-with-log-distance-constraint construction provides a particularly clean closed-form for γ in terms of RoPE parameters. The constraint itself is not derived from training; it is consistent with training in that the prediction $\gamma_{\text{Padé}}$ matches measurements across 30+ models with median MAE 4.3%."

This is a weaker but defensible claim. The strong claim ("transformer training optimises $\mathbb{E}[\log d]$") is not warranted.

Implication for the framework. The framework remains operationally useful regardless of how the MaxEnt assumption is justified:

- $\gamma_{\text{Padé}}$ predicts the centroid of observed γ across the panel (paper 2 §15).
- Operational consequences (D_f window, α_{opt} NTK, d_{horizon}, etc.) follow from the power-law form and require only $A(d) \propto d^{-\gamma}$, not the MaxEnt derivation.
- Empirical regularities discovered in paper 2 (the imprint $\nu = -1/(2\pi)$, the precision-sensitivity sink behaviour, the bimodal distribution) hold regardless of the specific derivation pathway.

Path forward. We will measure the empirical $\langle \log d \rangle$ across the existing panel to determine whether the MaxEnt assumption holds within $\pm 5\%$. The result will be incorporated into the companion dataset [16] and any failure will be documented openly. Either outcome leaves the framework's operational claims intact; the difference is in the strength of the foundational story.

Crediting the critique. This caveat was strengthened by external critique pointing out that the constraint (52) is intuited from RoPE geometry rather than derived from optimisation. We thank the critic for the precision of the observation. We accept the framing correction and present here the refined version.

18 Withdrawn Findings: Honest Failure Catalog

Why this section exists. A research framework worth trusting reports its refutations as prominently as its confirmations. This section catalogs claims introduced in paper 1 [15] or in early drafts of paper 2 that we have since refuted, demoted, or scoped down based on subsequent evidence. Each entry includes the original claim, the experiment or analysis that refuted it, and the lesson learned.

18.1 F13 — $\gamma = 1 - 1/\varphi$ as a code-tuning attractor

Original observation. In session 29 (2026-04-28), four LLMs in our panel showed γ_{text} values close to famous mathematical constants:

- CodeLlama-13b-Instruct: $\gamma = 0.3823$, near $1 - 1/\varphi = 0.3820$ (golden conjugate, $|\text{err}| = 0.0003$)
- pythia-1.4b: $\gamma = 0.7051$, near $1/\sqrt{2} = 0.7071$
- Llama-2-7b: $\gamma = 0.2871$, near $1 - 1/\sqrt{2} = 0.2929$
- Mistral-Nemo-Instruct: $\gamma = 0.4284$, near $\log_{10}(e) = 0.4343$

The CodeLlama hit was particularly striking. We hypothesized that code-tuned models might converge to $\gamma = 1 - 1/\varphi$ as a discrete attractor in the γ-space.

Test (E5). We measured γ_{text} on five additional code-tuned models: `starcoder2-3b`, `starcoder2-7b`, `Qwen2.5-Coder-7B`, `deepseek-coder-6.7b-base`, plus the existing `CodeLlama-13b-Instruct-hf`.

Model	γ_{text}	Near $1 - 1/\varphi$?
starcoder2-3b (base)	0.994	no
starcoder2-7b (base)	0.774	no
Qwen2.5-Coder-7B (base)	0.945	no
DeepSeek-Coder-6.7B (base)	0.951	no
CodeLlama-13b-Instruct	**0.382**	**yes**

Table 18: Code-tuned cluster test. Only CodeLlama-13b-Instruct (the sole *instruction*-tuned model in the set) hits the candidate attractor.

Refutation. Four of five code-tuned base models cluster at $\gamma \in [0.77, 0.99]$, not at 0.382. The single hit was the only instruction-tuned model. The "code-tuning attractor" hypothesis is refuted; the CodeLlama finding is more parsimoniously explained by instruction-tuning effect (see §16).

Lesson. Four hits on famous constants in a continuous distribution of $n = 25$ panel measurements are within the regime of expected coincidences. A genuine attractor would manifest with $n \geq 8$ hits or a mechanistic derivation. We infer the original observation *post hoc* from the data; without independent test (E5), it would have entered the literature as a "discovery". This is the classic post-hoc-attractor failure mode and demonstrates the value of pre-registered falsification protocols.

18.2 $R_c^\star \approx 1.68$ as a clean GQA-ratio boundary

Original claim (paper 1 / paper 2 early draft). Across a small initial sample ($n = 4$), the GQA query-to-KV ratio $R_c = n_{\text{heads}}/n_{kv}$ appeared to predict post-induction-head capability emergence with a sharp boundary at $R_c^\star \approx 1.68$. We added it to the framework as the "ignition criterion" for in-context learning.

Refutation (cloud round 2+3 measurements, $n = 9$). Expanded panel measurements showed an overlap zone of $R_c \in [0.92, 17.08]$ with no sharp boundary. The killer counter-example: pythia-2.8b ($R_c = 0.47$, post-IH evidence) versus Llama-3-8B ($R_c = 0.92$, also post-IH). The original $R_c^\star$ was an artifact of the $n = 4$ sample distribution.

Demoted to. A diagnostic indicator with overlap zone, not a clean boundary. Robust only for $R_c \geq 2.0$, where post-IH behavior is consistently observed; below that threshold, the metric loses discriminative power. We cite this as an example of how boundary claims fail under sample expansion.

Lesson. Boundary claims with $n = 4$ are highly susceptible to sample artifacts. Subsequent expansion to $n = 9$ refuted cleanly. Pre-registration of the prediction would have caught this earlier.

18.3 Coordination zone fractional depth $1 - 2/N$ in γ-space

Original claim. Following the bridge to Marín [14], we initially proposed that the inner-zone coordination (CKA $\to$ 1.0) and outer-zone specialization (CKA decreasing) would manifest in γ-space as inner γ saturating near 1 (Hagedorn) and outer γ being distinct.

Test on pythia-1b 4 checkpoints. At step 143,000, the inner-zone mean $\langle\gamma\rangle_{\text{inner}} = 1.278 \pm 0.267$, outer-zone mean $\langle\gamma\rangle_{\text{outer}} = 1.283 \pm 0.066$. The difference is $\Delta = -0.004$ — essentially zero. CKA-based outer specialization does *not* manifest as γ-divergence (§15).

Refutation and reframing. The zero Δ refutes the "outer specialization in γ-space" claim. The inner-zone Hagedorn-saturation claim survives partially ($\langle\gamma\rangle_{\text{inner}}$ exceeds 1, near boundary) but with high per-head variance not predicted by the original framing. We retain only the temporal alignment between paper A's $\hat{O}$-peak and our γ-flow inflection, which is robust (§15).

Lesson. CKA and γ are independent metrics. Cross-paper bridges must be tested empirically before incorporating into a framework. The analogy "specialization $\Rightarrow$ divergence" is appealing but not metric-invariant.

18.4 Mittag-Leffler prefactor for $A(d)$

Original claim (sage round, paper 2 early draft). Symbolic analysis suggested that the empirical attention prefactor A_0 in $A(d) = A_0 \cdot d^{-\gamma}$ might match the Mittag-Leffler prefactor $1/\Gamma(1-\gamma)$, providing a connection to anomalous diffusion theory.

Test ($n = 39$). Across the panel, the empirical median ratio of measured A_0 to predicted $1/\Gamma(1-\gamma)$ was 0.23 with standard deviation 5.5 — no signal. The truncated power-law normalization $1/Z(\gamma, N)$ where $Z(\gamma, N) = \sum_{d=1}^{N} d^{-\gamma}$ instead matched with median ratio 1.30, std 0.56.

Refutation. Mittag-Leffler prefactor refuted. $A(d)$ follows truncated Zipf-like form, not anomalous-diffusion fractional Mittag-Leffler form.

Lesson. Symbolic-analysis matches between distributions can be coincidental. Empirical validation across $n = 39$ models was decisive. Avoid incorporating symbolic analogies into the framework without testing the prefactor (not just the exponent).

18.5 Universal R_c ignition criterion

Original (paper 1 §App α_S). A composite formula $\alpha_S = F_{R_c} \cdot F_{\text{Chin}}$ was proposed as a "semantic ignition constant" predicting in-context learning emergence from R_c and Chinchilla compute-optimality.

Provisional status. With $R_c^\star$ refuted (above), F_{R_c} component is also weak. Empirical accuracy of α_S on $n = 4$ models was 75%, but the known failure mode (the gate kills MHA models like pythia-2.8b that DO ignite) was not captured.

Demoted to. "preliminary research scaffold" with explicit caveat in paper 1's appendix; not a operational predictor for deployment.

Lesson. Composite formulas built from refutable components inherit the weakness of those components. Single-formula predictors are easier to refute and replace.

18.6 Holographic Schwarzschild metaphor for γ-flow

Original (early sage exploration). A speculative framing connected the γ-flow to the Schwarzschild metric and AdS/CFT correspondence, suggesting attention has an "event horizon" at $\gamma = 1$.

Refutation (EXP-METRIC-RoPE). Empirically, γ-flow under RoPE follows Euclidean fractional dimension $d_{\text{eff}} = 1/\gamma$, not hyperbolic / Schwarzschild geometry. The volume polynomial scaling tested differs from the hyperbolic prediction by orders of magnitude.

Lesson. Physics analogies must be tested against empirical geometry of the system, not adopted by aesthetic resonance. The fractional-Laplacian framing ([15, paper I, RG section]) survives; the Schwarzschild framing does not.

18.7 Soft-decay KV truncation as universal rule

Original (paper 1 §sec:kv_horizon_decay). The soft-decay weight $w(d) = (1-d/d_{\text{horizon}})^\gamma$ was proposed as a universal rule for KV-cache truncation, applicable across the panel.

Refutation (EXP-KV-decay, $n = 3$). Pythia-1b ($\gamma = 0.931$, $d_h = 505$, $d_h \ll T_{\text{train}}/2$) showed hard-truncation winning by perplexity. The soft-decay rule applies only when $d_h \gtrsim T_{\text{train}}/2$.

Demoted to. Régimen-bounded rule: applicable for $d_h \gtrsim T_{\text{train}}/2$, not universal. Documented in paper 1 NR-12 appendix.

Lesson. Universal rules need confirmation across the parameter range. Single- case validation is insufficient; the régime boundary must be identified explicitly.

18.8 Information-horizon "law" as Padé tautology

Original (paper 1). $d_{\text{horizon}}(\theta, \gamma) = \theta(1-\gamma)\sqrt{2}/(1+\gamma)$ was framed as a separate "horizon law" with the empirical $d_{\text{horizon}} \approx T_{\text{train}}$ on Pythia ($n = 8$, 0.4% err) presented as supporting evidence.

Demotion (Phase 7). Substituting $\gamma_{\text{Padé}}(\theta, T)$ into the formula yields $d_{\text{horizon}} \equiv T_{\text{eval}}$ identically (§13, Eq. (43)). The Pythia agreement therefore reflects the precision of the Padé fit on Pythia, not an independent regularity.

Reframed to. A diagnostic ratio PDI $= d_{\text{horizon,obs}}/T_{\text{eval}}$ (§13). When PDI $\neq 1$, the model deviates from the canonical Padé regime; the deviation is then the diagnostic signal.

Lesson. Empirical "laws" across a sub-panel can be tautological consequences of an existing identity. Algebraic substitution should be checked *before* reporting an empirical regularity as a separate claim.

18.9 $\kappa \cdot N_{\text{sem}} = \ln(2\pi)$ as a topological invariant

Original (paper 1 [15, paper I, topological invariant]). Empirical observation $\kappa \cdot (N - L_{\text{crit}}) = \ln(2\pi)$ on Pythia-1b ($n = 1$ data point, exact match) was framed as a candidate topological invariant of the crystallization process.

Demotion (Phase 7). Since κ is *defined* as $\kappa := \ln(2\pi)/N_{\text{sem}}$, and $N_{\text{sem}} := N - L_{\text{crit}}$, the relation $\kappa \cdot N_{\text{sem}} = \ln(2\pi)$ holds by definition for *any* model and any Galois Type. It is a tautology of the κ definition, not a topological invariant.

Reframed to. Reportable as a definitional consistency check, not as evidence for crystallization-invariant structure. The genuinely topological content of paper 1 lies in the gcd-based Galois Type classification ([15, paper I, Galois types]), not in this identity.

Lesson. A formula that "always works" across all candidate test cases because it follows from a definition is not an empirical claim. Audit "invariants" for tautology before publishing.

18.10 $K = 51.2$ as universal compute-context constant

Original (Sesion 29-30). The cross-model invariant $K := \gamma \cdot \log(N^2 \cdot D) \approx 51.2 \pm 16.8$ ($n = 22$, CV $= 0.329$) was reported as a universal compute-context coupling constant.

Demotion (Phase 7). The structural form $K = \gamma \cdot (3 \ln N + \ln 20)$ is general (combines the panel γ with Chinchilla's $D = 20N$). The specific value 51.2 is the panel *centroid*, dependent on the (γ, N) distribution of our test models. Different panels (different families, different size mixes) will produce different centroids while the structural form persists.

Reframed to. Structural form retained; numerical centroid labeled as panel-specific. Reproducibility requires reporting both the form and the panel composition.

Lesson. Empirical centroids of structural relations should not be claimed as universal constants without verification across independent panels.

18.11 $l_{1/2} \approx 6.62$ as a universal AbsPE invariant

Original (paper 1 §sec:absPE_layers). The 50% recovery layer $l_{1/2} = 1.5 + e^2 \cdot \ln(2) \approx 6.62$ was reported as a universal AbsPE invariant across the GPT-2 family.

Demotion (Phase 7). Solving $R(\ell) = 0.5$ from the universal AbsPE recovery curve $R(\ell) = 1 - \exp(-(\ell - 1.5)/e^2)$ (paper 1 Theorem T-P) gives exactly $l_{1/2} = 1.5 + e^2 \cdot \ln 2$. The numerical value is a tautology of $R(\ell)$ at $R = 0.5$; it carries no information beyond the recovery curve itself.

Reframed to. Definitional consequence of $R(\ell)$ at $R = 0.5$, not an independent invariant.

Lesson. "Universal constants" that are quantile-evaluations of an existing formula should be reported as such, not as independent observations.

18.12 $P(\text{hallu})$ formula at the Hagedorn boundary

Original (paper 1 §sec:P_hallu). The hallucination probability formula $P(\text{hallu}) = \max(0, 1 - (d_h/L)^{1-\gamma}) \cdot \sqrt{\chi}/(1 + \sqrt{\chi})$ applied across the panel.

Anomaly identified (Phase 7). At $\gamma = 1$ (Hagedorn boundary), $(d_h/L)^{1-\gamma} = (d_h/L)^0 = 1$, making the first factor $1 - 1 = 0$. The susceptibility factor $\sqrt{\chi}/(1 + \sqrt{\chi}) \to 1$ as $\chi \to \infty$. The product is $0 \cdot 1 = 0$.

But 36% of our panel sits at $\gamma \geq 1$ (§8) and these models *do* hallucinate. The formula gives $P = 0$ where empirical evidence shows non-zero risk.

Open problem. The formula needs regularization in the Hagedorn regime, e.g., replace $(d_h/L)^{1-\gamma}$ with $(d_h/L)^{\max(\varepsilon, 1-\gamma)}$ for small ε to floor the exponent. We mark this as a paper 2 open problem (§21.2); empirical calibration of ε requires Hagedorn-zone hallucination data.

Lesson. Closed-form falsifiable formulas can give pathological zeros at phase transitions; sanity-check at boundary cases as part of derivation.

18.13 Fisher-Rao distance dimensional issue

Original. The Fisher-Rao geodesic distance was written as $d_{\text{FR}}(\gamma_1, \gamma_2)^2 = \int_{\gamma_1}^{\gamma_2} I(\gamma)\, d\gamma$.

Correction (Phase 7). The Riemannian Fisher-Rao distance in 1D is $d_{\text{FR}}(\gamma_1, \gamma_2) = \int_{\gamma_1}^{\gamma_2} \sqrt{I(\gamma)}\, d\gamma$ (line element $\sqrt{I(\gamma)}\, d\gamma$). The squared form $\int I\, d\gamma$ is dimensionally a different quantity (KL-like cumulative information, not geodesic length).

Reframed to. Use canonical Riemannian form $d_{\text{FR}} = \int \sqrt{I}\, d\gamma$ throughout. The $\int I\, d\gamma$ form is retained but labeled as "cumulative Fisher information" to avoid ambiguity.

Lesson. Differential-geometric formulas have specific conventions; check units and dimensions before transcribing.

Withdrawn / demoted claim	Refuted / demoted by	Lesson
F13 $\gamma = 1 - 1/\varphi$ attractor	E5 ($n = 5$ code-tuned)	Post-hoc attractor on $n = 4$ susceptible to coincidence
$R_c^\star \approx 1.68$ boundary	cloud rounds 2+3 ($n = 9$)	Boundary claims need $n \geq 8$
$1 - 2/N$ in γ-space	pythia-1b 4-ckpt	CKA $\neq \gamma$; metrics independent
Mittag-Leffler prefactor	EXP-VALIDATE ($n = 39$)	Test prefactor, not just exponent
α_S semantic ignition	inherits $R_c^\star$ refutation	Composite formulas inherit weakness
Schwarzschild γ-metaphor	EXP-METRIC-RoPE	Test geometry, not adopt analogy
Soft-decay universal	EXP-KV-decay ($n = 3$)	Régime-bound rules need explicit boundaries
Phase 7–12 demotions (Sesion 31, 2026-04-30)		
$d_{\text{horizon}} \approx T_{\text{train}}$ as separate law	D-NEW-1 algebraic substitution	Check tautology before reporting empirical law
$\kappa \cdot N_{\text{sem}} = \ln(2\pi)$ as invariant	D-37 definition substitution	Audit "invariants" for tautology
$K = 51.2$ universal	D-DEEP-3 panel-centroid analysis	Centroids are panel-dependent
$l_{1/2} \approx 6.62$ universal	D-DEEP-5 quantile of $R(\ell)$	Quantiles are not separate constants
$P(\text{hallu})$ at $\gamma = 1$	D-DEEP-8 algebraic limit	Sanity-check formulas at phase boundaries
Fisher-Rao $d_{\text{FR}}^2 = \int I \, d\gamma$ form	D-DEEP-10 dimensional check	Differential geometry conventions matter
"0.3% match" for $\nu = -1/(2\pi)$	E1 robustness audit ($n = 22$)	Bootstrap sensitivity check before claiming precision

Table 19: Refutations and demotions catalog. Top half: paper 1 to paper 2 revisions. Bottom half: post-publication Rosetta-tablet algebraic audit (Sesion 31, 2026-04-30) added six tautology demotions and one dimensional fix.

18.14 Summary table

Net. Fourteen claims withdrawn, demoted, or scope-restricted across the paper 1 → paper 2 process plus the Rosetta-tablet algebraic audit. The framework is substantially leaner and more honest. Several "empirical regularities" from paper 1 (information horizon, κ invariant, $l_{1/2}$, $K = 51.2$ centroid) revealed themselves as tautological consequences of structural definitions once subjected to algebraic audit. The genuine empirical content concentrates around $\gamma_{\text{Padé}}$, ν imprint, ΔH_{Cardy} linear, the bimodal phase structure, and the precision-sensitivity boundary — claims that survive the audit unchanged.

19 Practical tools: TAF Agent

This section documents the operational tools that materialise the paper's findings. The TAF Agent (karlexmarin/taf-agent) is a browser-based diagnostic tool that runs the full TAF prediction chain for any transformer LLM, given the model's HuggingFace identifier or its `config.json`. All computation is local (Pyodide); no telemetry, no GPU required for the diagnostic itself.

19.1 Eight diagnostic recipes (v0.4)

The Agent exposes eight named recipes that combine paper II's findings into operational predictions:

X-1 — Custom training vs API. Computes break-even months for training a custom model versus paying for an API. Uses Chinchilla compute estimate ($6ND$) plus serve cost from recipe X-5. Output: `GO custom` / `API better` with break-even months.

X-2 — Long-context viability. Predicts whether a model will reliably serve a target evaluation context T_{eval}. Chains $\gamma_{Pad}(\theta, T_{\text{eval}}) \to$ architectural correction $\to d_{\text{horizon}} \to$ PDI $\to L_{\text{NIAH}}$ ceiling $\to$ hallucination rate $\to$ KV memory. Verdict: `YES / DEGRADED / NO` with mitigation if needed.

X-3 — Budget pre-flight. Given a USD budget, returns the largest feasible model under Chinchilla-optimal training [8]. Uses the $6ND$ FLOP estimate [11]: $N \cdot D \cdot 6 \cdot \text{cost-per-FLOP} \leq B$. Verdict: `GO / TINY-MODEL / MEMORY-LIMITED` with concrete (N, D).

X-5 — Hardware selection. Recommends GPU configuration for a target throughput (tokens/day). Uses peak FLOPs × MFU heuristic (0.4 default). Validated within ±20% against vLLM published numbers for common models.

X-19 — KV compression decision. Decides between soft decay, hard cutoff, and literature methods (SnapKV, PyramidKV) for KV cache compression at a target T_{eval}. Verdict: `USE SOFT DECAY` / `USE D_f CUTOFF` / `USE LITERATURE METHODS` / `USE HARD T_train`.

X-21 — Imprint Purity Diagnostic. Predicts γ_{random} on uniformly-random tokens via the learned-imprint formula $\gamma_{\text{random}} = \gamma_{Pad}(\theta, T) + \nu \log_{10}(P/14M)$ with $\nu = -1/(2\pi)$. Compares to measured γ_{random} to flag imprint anomalies (`CLEAN / OVER-IMPRINTED / UNDER-IMPRINTED`). See §5.

X-22 — Compute-Context Invariant. Computes $K = \gamma \log(N^2 \cdot D)$ with Chinchilla $D = 20N$ and checks against panel band 51.2 ± 16.8 (CV $= 0.329$, $n = 22$). Verdict: `IN-BAND / OUTLIER`. See §14.

X-23 — IH-Phase Detector. Returns the sign of $\Delta\gamma = \gamma_{\text{text}} - \gamma_{\text{random}}$ as a cheap induction-head probe. Cheaper than running ICL benchmarks. Cross-validated against the $N \geq 400$M emergence threshold of Olsson et al. [19].

19.2 The TAF Card

Profile mode runs all eight recipes simultaneously and produces a unified **TAF Card**: a single panel with verdict per dimension, key numbers, falsification status, and the v0.5 algebraic-consistency check (described below). The TAF Card is the canonical artifact for documenting whether a model is suitable for a deployment scenario. It is shareable via URL and exportable as JSON.

19.3 v0.5: Machine-verified consistency check

The v0.5 update (sesion 32, 2026-05-01) introduces a runtime sanity check that verifies the 12 D-SAGE algebraic identities of Appendix A, given a user-input γ value. The function `verify_algebraic_consistency(gamma)` returns a structured report:

```
{
  "input_gamma": 0.485,
```

```
  "phase": "A (gamma in (0,1))",
  "n_checks_total": 12,
  "n_checks_passed": 12,
  "all_consistent": true,
  "framework_verified_by": "Sage Groebner basis + Lean Mathlib4",
  "checks": {
    "D-SAGE-1": {claim: "2*eta^2 + eta*gamma_chi + 1 = 0", passes: true},
    "D-SAGE-2": {claim: "beta*chi = -1", passes: true},
    ...
  },
  "interpretation": "All 12/12 D-SAGE identities consistent (PASS)"
}
```

When checks fail. For any $\gamma \in (0, 1)$, all 12 checks pass identically (the identities are tautologies given the TAF exponent definitions). Failure of any check indicates one of:

- Numerical precision artifact (e.g., γ within 10^{-9} of phase boundary).
- Out-of-Phase-A input (function refuses to proceed if $\gamma \leq 0$ or $\gamma \geq 1$).
- Implementation bug in the framework code (the function would catch this).

The check is dimension-light: $O(1)$ floating-point operations. It runs in milliseconds and is exposed in the TAF Card via the **[Sci] Verify algebraic consistency** button. The function is also callable directly via Pyodide for power users.

Trust signal for users. The verify-consistency feature distinguishes the TAF Agent from other transformer-attention diagnostic tools (e.g., AI Scientist, mechanistic interpretability dashboards): it is the first such tool with formal machine-proof backing. The exact same algebraic facts run by Sage Groebner basis (or Lean Mathlib4 in offline mode) are checked at runtime in the browser. This closes the loop between the paper's theoretical claims and their operational consequences.

19.4 The diagnostic CLI (Diagnose mode)

For users who want to measure γ on a specific model directly (rather than relying on the panel-derived γ_{Pad} prediction), the **[Diag] Diagnose CLI** mode generates a Python command:

```
git clone https://github.com/karlesmarin/tafagent
cd tafagent
pip install torch transformers numpy
python -m tafagent.cli.diagnose <model-id> --T 4096 --device cuda
```

This produces a JSON output with measured γ_{text}, γ_{random}, $\Delta\gamma$, R^2 values, and a chain of intermediate quantities. The JSON can be re-uploaded to the TAF Agent via **Inspect mode** to feed it into the full TAF Card.

Run time: $\sim$ 5 minutes on CPU, $\sim$ 20–60 minutes on GPU for larger models. Public dataset [16] contains pre-computed measurements for 32 models.

19.5 Other modes

[VS] Compare Side-by-side TAF Card for 2–3 models on the same recipe. Useful for choosing between candidate models for a given deployment scenario.

[Inspect] Inspect config Paste raw `config.json`; tool parses architectural parameters and runs full Profile. For private models, in-development configs, or models not yet on HF Hub.

[Ask] Ask plain English Free-form question; an in-browser LLM (Qwen2.5-0.5B, $\sim$ 350 MB cached after first load) selects the appropriate recipe. Best for casual exploration; the deterministic TAF computation is always shown alongside.

[Recipe] Recipe + form Manual recipe selection with full parameter control. For users who need exact control over inputs.

[Phase] Phase diagram Scatter plot of 23 panel models on the $(\log\theta, \gamma)$ plane. Visualises the Hagedorn cluster, Phase B boundary, and Padé curve. Click a dot to load that model into the Recipe form.

19.6 Worked examples: TAF function outputs

To make the TAF Agent concrete, we show three worked examples reproducing real TAF computations. All three are obtainable in the browser tool with ~ 5 seconds each.

Example 1 — γ_{Pad} + architectural correction (Llama-3-8B at $T_{\mathbf{eval}} = 8K$).

```
Input:  theta=500000, T_train=8192, T_eval=8000,
        n_attention_heads=32, n_kv_heads=8, d_head=128,
        n_layers=32, n_params=8e9, has_SWA=False
TAF computation:
  z              = T_eval/theta             = 0.0160
  gamma_Pade     = (2*theta - T*sqrt2)
                   /(2*theta + T*sqrt2)     = 0.977
  gamma_text     = gamma_Pade - 0.012*n_kv  = 0.977 - 0.096
                                            = 0.881
  d_horizon      = theta*(1-g)*sqrt2/(1+g)  = 44766 tokens
  PDI            = d_horizon / T_eval       = 5.59
  L_NIAH ceiling = T_eval (PDI > 1)         = 8000 (full context useful)
  KV memory      = 2*L*T*n_kv*d_h*bytes(2)  = 1.07 GB (bf16)
Verdict X-2: YES (long-context viable, 95% NIAH retrieval expected)
```

Example 2 — `verify_algebraic_consistency` (Mistral-7B at $\gamma = 0.485$).

```
Input:  gamma = 0.485
TAF computation: 12 D-SAGE identities checked
  D-SAGE-1: 2*eta^2 + eta*g_chi + 1 = 0      [0.0e+00]  PASS
  D-SAGE-2: beta * chi               = -1    [0.0e+00]  PASS
  D-SAGE-4: alpha + chi              = 2     [0.0e+00]  PASS
  D-SAGE-5: alpha + g_chi  = 2*(2-g)         [0.0e+00]  PASS
  D-SAGE-6: beta * g_chi   = -2g^2+4g-3      [0.0e+00]  PASS
  Rushbrooke: 2*beta + g_chi - nu*1 = 0      [0.0e+00]  PASS (taut.)
  Josephson:  2 - alpha - nu*1      = 0      [0.0e+00]  PASS (taut.)
  Fisher residual = g(2g-3)/(1-g)             [-0.408]  IS NOT 0 (correct)
  eta_2gamma residual > 0 (Phase A)          [+1.193]  REFUTED PASS
  D-14 nu_imprint * 2*pi             = -1    [0.0e+00]  PASS
  D-SAGE-7: c_central * |nu| * 2pi   = 3     [0.0e+00]  PASS
  nu_beta_id: nu * beta              = -1    [0.0e+00]  PASS
```

```
Result: 12/12 D-SAGE identities consistent.
Framework algebraic structure: INTACT.
Verified by: Sage Groebner basis + Lean Mathlib4.
```

Example 3 — Compute-context invariant K (Mistral-7B).

```
Input:  gamma=0.485, N=7e9, D_chinchilla=20*N=1.4e11
TAF computation:
  K = gamma * log(N^2 * D)
    = 0.485 * log( (7e9)^2 * 1.4e11 )
    = 0.485 * log( 6.86e30 )
    = 0.485 * 70.84
    = 34.36

Panel band: 51.2 +/- 16.8 (CV = 0.329)
z-score:    (34.36 - 51.2) / 16.8 = -1.00
Verdict X-22: IN-BAND (1 sigma below mean - consistent with Chinchilla)
```

These three examples cover the framework's three claim types: empirical (Ex. 1), algebraic (Ex. 2), and diagnostic (Ex. 3). All are reproducible in the live browser tool at karlexmarin/taf-agent.

19.7 Reproducibility kit

The full reproducibility stack is open-source:

- Code: github.com/karlesmarin/tafagent
- Live demo: karlexmarin/taf-agent
- Dataset: karlexmarin/taf-attention-decay (32 models, 58 measurements, CC-BY-4.0)
- Sage verification: `analysis/sage_recursive_sweep_2026-04-30.sage`
- Lean Mathlib4 proofs: `lean_taf/taf/Taf/Identities.lean`

The TAF Agent is multi-language (English, Spanish, French, Chinese) with all 209 i18n keys complete. License: Apache-2.0 (code), CC-BY-4.0 (data).

Coverage at time of writing. v0.5 release (sesion 32, 2026-05-01). 8 recipes operational. 32 models profiled in dataset. 15 algebraic identities machine-verified. No funded team, no GPU dependency for the diagnostic itself. All running cost: $0.

20 Discussion

This paper has presented four constructive contributions to the TAF framework: a learned-imprint axis ($\nu = -1/(2\pi)$), an algebraic Cardy decomposition, a precision-sensitivity boundary with mechanism, and a bimodal phase structure with the Lévy stability link. It has also documented a quadratic algebraic identity (D-SAGE-1) that supersedes an earlier informal closure claim, and corrected a paper I claim ($\eta = 2\gamma$) by formal proof. The companion TAF Agent v0.5 tool brings these results into operational use.

This section synthesises the work as a whole and notes its limitations.

20.1 What survived rigorous audit

Across $\sim$ 132 Sócratic interrogations, $\sim$ 24 formula-level comparisons, eight chains of cross-formula derivations, dual-tool algebraic verification (Sage Groebner basis + Lean Mathlib4), and two independent literature scans of recent (2024-2026) transformer-attention and critical-exponents work, five claims survived as *robust phenomenological regularities* — sustained by multiple convergent paths of evidence on the current panel ($n = 22$–32 depending on availability). We deliberately avoid "universal" as a label here: universality requires reproduction beyond a single panel, beyond a single corpus, and ideally on synthetic-task validation experiments — none of which we provide. The five regularities below are the strongest candidates to become universal claims, contingent on future work:

1. **Learned-imprint slope** $\nu = -1/(2\pi)$: derived (not fitted) from RoPE rotation period, supported by three independent arguments (dimensional, phase-storage, sign-consistency), validated empirically across $n = 22$ models with 0.6% relative error on the panel mean. See §5.

2. **Architectural concentration law** $\gamma_{\text{text}} \approx \gamma_{Pad}(\theta, T) - 0.012\, n_{\text{kv}}$: correlational, $R^2 = 0.30$ cross-panel (14× improvement over γ_{Pad} alone). Not predictive per-model, but the dominant correction. See §9.

3. **Cardy decomposition** $\Delta H = \log(z/2) + 2\,\text{arctanh}(\gamma)$ with linear empirical slope ≈ 5 matching panel-averaged susceptibility $\langle\chi\rangle$. See §6.

4. **4-bit shift R^2-direction rule**: in MHA models without strong alignment, $R^2(\text{bf16}) < 0.9$ predicts 4-bit shifts γ *up*, $R^2 > 0.99$ predicts shifts *down*. Validated $n = 5/5$ sign-correct including the DeepSeek bf16 outlier reframe. See §4.

5. **D-SAGE-1 quadratic identity** $2\eta^2 + \eta\gamma_\chi + 1 = 0$: machine-verified by Sage Groebner basis and Lean Mathlib4. Algebraic constraint connecting anomalous dimension and susceptibility. See §10.

20.2 What was honestly demoted

A symmetric set of claims, framed in paper I or in early paper II drafts, did not survive the audit. We document the seven most consequential demotions in §18:

- $R_c^\star \approx 1.68$ as a sharp boundary (refuted; overlap $[0.92, 17.08]$ in panel).
- $\gamma = 1 - 1/\varphi$ as a code-tuning attractor (one-shot coincidence; statistical significance $p = 0.018$, marginal after Bonferroni correction).
- Mittag-Leffler prefactor $1/\Gamma(1-\gamma)$ (refuted with $n = 39$; truncated power law fits better).
- Universal soft-decay KV truncation rule (recipe-specific, not universal).
- "0.3% match" framing for ν (recalibrated to 13.7% relative error within panel CI; honest distinction).
- d_{horizon} "law" (algebraically tautological: $d_{\text{horizon}} \equiv T_{\text{eval}}$ when $\gamma = \gamma_{Pad}$).
- $F13$ universal $1 - 1/\varphi$ cluster (refuted across $n = 5$).

The "triple closure" of Rushbrooke, Josephson, and Fisher hyperscaling relations — a claim made internally during sesion 31 exploration — was likewise demoted. Sage Groebner basis revealed that Rushbrooke and Josephson are tautologies given the standard definitions of γ_χ and α_C, while Fisher is independent. The genuine algebraic content is the simpler quadratic identity (D-SAGE-1, §10).

20.3 Methodology contribution

We argue, perhaps more controversially than the empirical findings, that the *audit pipeline itself* is a contribution worth reporting. The following protocol detected our own errors and made the framework more honest:

1. **Cross-of-crosses semantic combination**: combine claims across abstraction levels (architecture → geometry → dynamics → phenomenology → interpretation). Reveals mechanisms invisible to single-level analysis.

2. **Triangulation requirement** (≥ 2 independent paths): single-path findings are quarantined until a second path confirms.

3. **Sage Groebner basis as algebraic decision procedure**: ideal-membership testing in PolynomialRing($\mathbb{Q}$) is strictly stronger than `sympy.simplify` heuristics, and it is what caught our $\eta = 2\gamma$ error. Free, dockerised, runs in seconds.

4. **Lean Mathlib4 as gold-standard formal proof**: machine-checked theorems in dependent type theory are publication-grade rigour. Our 15 TAF identities compile in 1973 jobs (Mathlib4 + 15) on a public Docker image.

5. **Post-finding lit-scan** (5-dimension, 7-source): every novel claim is checked against existing literature *before* attribution of novelty. Concurrent papers [18, 21] were found to use overlapping but distinct frameworks; the lit-scan made this explicit.

6. **Honest demotion catalogue**: failed claims are documented *by name*, not silently dropped. The catalogue (§18) is itself research output: it tells future readers what *not* to try.

The cumulative effect is that the framework's claims survive their own internal critique. We do not propose this as universal methodology, but we suggest that any quantitative framework benefits from a similar audit pipeline applied iteratively across drafts.

20.4 Limitations

The work has known limits, documented here for honesty:

Panel coverage. $n = 22$–32 models is insufficient for strong cross-architecture universal claims. The $R^2 = 0.30$ for the architectural concentration law leaves $\sim 70\%$ unexplained variance — most of which lies in specific models (Yi family at $n_{\text{kv}} = 4$, Qwen2.5-3B dual-regime, OLMo-7B $\Delta\gamma = 0.30$). These outliers are catalogued, not resolved.

Methodological dark zones. Several axes were not controlled in the panel and remain untested contributions to γ:

- *Tokenizer choice* (BPE vs Unigram vs SentencePiece) — could shift γ measurements; never tested.
- *Optimizer* (AdamW vs Lion vs SGD) — panel is AdamW-dominated.
- *Initialisation scheme* (μP vs standard) — only $n = 1$ measurement of γ_{init} at step zero.
- *LR schedule* (cosine vs WSD) — panel is cosine-dominated.
- *Cross-corpus comparability* — all measurements use a single corpus (Mongo); systematic cross-corpus variation untested.
- *Multi-modal models* (vision-LLMs) — out of scope.

- *State-space models* (Mamba, Hybrid) — different framework required; out of scope.

These constitute a research agenda for paper III or community work, not limitations of the present claims (which are scoped to AdamW + cosine + Mongo + standard tokenizer + RoPE attention).

Lévy fractional Laplacian framework. The mapping $\alpha_L vy = \gamma$ (with corrected $\eta = \gamma - 1$) to fractional Laplacian theory is suggestive, but our direct fractional Laplacian operator has not been measured on attention weights. The concurrent paper Qu et al. [21] formalises a related framework explicitly; cross-validation work is pending.

Bimodal phase claim. The Hartigan dip test on the panel of $n = 25$ gives $p = 0.72$, which does not reject unimodality at conventional significance. We have re-framed the bimodal claim accordingly: the cluster around $\gamma \to 1$ is dominantly an *industry design choice* (high-θ + low n_{kv} GQA models), not a phase-attraction mechanism. See §7 for the reframing.

Architectural law $k = 0.012$. The slope of the architectural concentration law was fit empirically; we have not derived it from first principles despite four hypotheses attempted. The constant remains a phenomenological parameter, valid for the panel but possibly drifting with future architectures.

20.5 Comparison with concurrent work

Three concurrent papers (October-November 2025) deserve explicit positioning:

Nakaishi et al. [18] — arXiv 2510.05554 "Critical attention scaling in long-context transformers" uses attention-score *rank-collapse* as the phase transition phenomenon. Their critical scaling $\beta_n \sim \log n$ is unrelated to our γ_{Pad} exponent (different formalism). *No theorem-level overlap*; the two frameworks could in principle be combined to give a two-axis (rank-collapse × decay-exponent) characterisation.

Tanaka et al. [26] — arXiv 2510.02458 "One-dimensional long-range Ising model" computes critical exponents in the long-range Ising regime $1/2 < \sigma < 1$. They derive $\eta(\sigma) = 2 - \sigma$. Mapping their convention to ours ($\sigma = \gamma - 1$) gives $\eta = 3 - \gamma$, not $\gamma - 1$. *Different convention* on the role of η — and our η is forced by the D-SAGE-1 quadratic, while their η comes from correlator scaling.

Qu et al. [21] — arXiv 2511.10208 "Fractional Neural Attention" (FNA) formalises attention as Lévy diffusion under fractional Laplacian. This is the closest concurrent work to our $\alpha_L vy = \gamma$ link. We have not yet completed full cross-validation; abstract-level alignment is encouraging. Future work will deconstruct whether their FNA dynamics yields our $\nu = -1/(2\pi)$ imprint slope as a derivative consequence.

20.6 Outlook

Near-term (paper III candidate). The dominant insight that may yet emerge from extending this work: TAF attention is best framed as *rational geometry of RoPE* (algebraic, non-Ising, $\beta < 0$ universality class) rather than as a direct analogue of statistical-mechanics phase transitions. A follow-up paper "Rational geometry of attention" could systematically reframe the exotic-physics language (Hagedorn, central charge, hyperscaling) into the simpler picture: rational *Pad* geometry + linear architectural correction + derived imprint slope. This would clarify what is *algebraically necessary* from what is *interpretive metaphor*.

Medium-term (community). The 12 D-SAGE machine-verified identities are now the formal core of the framework. The TAF Agent v0.5 `verify_algebraic_consistency` function exposes them at runtime. As the community contributes more panel data (cross-tokenizer, cross-optimizer), the architectural law's R^2 may improve; if it does not, the residual variance (currently $\sim 70\%$) is a clean target for new mechanism discovery.

Methodological propagation. We have demonstrated, at least in our own work, that Sage Groebner basis + Lean Mathlib4 dual-tool verification catches errors that `sympy.simplify` does not. We encourage other transformer-attention researchers to apply the same audit pipeline before publishing identities that depend on specific exponent definitions. Free, dockerised, fast.

Closing remark. This paper began as a panel-expansion campaign and an algebraic audit of paper I. It returned with four new constructive findings, seven demotions, one machine-verified central identity, and a corrected anomalous dimension. The framework is leaner and more honest. We thank the very tools — Sage and Lean — that found our own errors before reviewers would have.

21 Future Work

This section catalogs the open problems and experiments emerging from paper 2's analytical audit (Rosetta-tablet, Sesion 31, 2026-04-30). We distinguish between (i) experiments that would strengthen specific paper 2 claims, (ii) open theoretical problems that paper 2 explicitly leaves unresolved, and (iii) cross-paper integration directions.

21.1 Strengthening experiments (paper 2 follow-up)

E3 — Panel expansion to $n \geq 30$. Several paper 2 claims sit at the boundary of statistical resolution on the current $n = 22$ panel:

- ν imprint slope: bootstrap CI $[-0.260, -0.008]$ wide; pythia-31m drives $\sim 30\%$ of the signal (§5).
- Hartigan dip test on bimodal γ_{text}: requires $n \geq 30$ for reliable rejection of unimodality (§7).
- C-8 Cardy linear regression: $r = 0.69$, $p = 0.003$ on $n = 16$; expansion would tighten slope estimate (§6).

Lit-scan (2026-04-30) confirmed no published cross-LLM γ panel of $n \geq 27$ exists, so these measurements must be acquired in-house. Five candidate models for inclusion: yi-6B, gemma-2-2b, phi-3-medium-4k, qwen2.5-1.5B, internlm3-8b. Estimated cost: 6–8h GPU.

ν trajectory across Pythia checkpoints. The strongest test of the imprint mechanism (§5) is to fit ν separately at each of the 35 documented Pythia-70M training checkpoints [2] and verify the predicted asymptote $\nu \to -1/(2\pi)$ as step $\to \infty$. A monotonic approach to the constant would constitute substantial empirical support for the phase-storage interpretation; deviation would invite a richer functional form. Estimated cost: 4h GPU (35 checkpoints, fixed protocol).

4-bit precision-sensitivity verification ($n \geq 3$ MHA no-RLHF). The mechanism in §4 (4-bit truncates long-tail attention $\Rightarrow$ reveals $\gamma_{\text{Padé}}$ for MHA-no-RLHF) currently rests on $n = 1$ data point (DeepSeek-7B-base). We predict that any MHA model without strong alignment training, measured at 4-bit, will give $\gamma_{\text{4bit}} \approx \gamma_{\text{Padé}}(\theta, T)$. Candidates: pythia-1B (predicted $\gamma_{\text{Padé}} \approx 0.747$), pythia-2.8B, Llama-2-7B-base, GPT-NeoX-20B, OPT-6.7B. The shift magnitude prediction $\Delta\gamma_{\text{4bit}} \propto (1 - \gamma_{\text{bf16}})$ becomes testable at $n \geq 3$.

F9 paired base/instruct measurements ($n \geq 6$). The recipe-dependent RLHF effect on γ (§16) currently has $n = 3$ paired measurements (Mistral, Qwen, DeepSeek). Adding three pairs (Llama-2 base/chat, gemma-7B base/instruct, phi-3 base/medical) would shift this finding from "borderline" to "confirmed recipe-dependent".

C-8 controlled-z extension to other architectural families. The Pythia-only subset confirms the Cardy slope $b = 2/(1 - \langle\gamma\rangle^2)$ at $r = 0.97$ (§6, Eq. (16)). The same test on the GPT-2 family (AbsPE, $T = 1024$ effective) and the Qwen family (GQA) would establish whether the prediction holds across architectures. Existing paper 1 data covers GPT-2; the analysis is mostly re-aggregation.

$c_{\text{central}} = 3$ direct measurement. The CFT-based derivation $c = 3$ from the $\gamma \to 0$ limit of S_γ ([15, paper I, Appendix A]) assumes the Cardy formula applies. Direct measurement of attention entropy $S(L) = -\sum_d p(d) \log p(d)$ vs. $\log L$ on small-γ models (Llama-2-7b $\gamma \approx 0.29$, DeepSeek-base $\gamma \approx 0.15$, gpt-2-medium $\gamma \approx 0.78$), fitting $S = (c/3) \log L$, recovers c empirically. Lit-scan (2026-04-30) confirms log-scaling of attention entropy is established [18], but the fit-to-c angle is novel.

Hartigan dip test (CPU only, existing data). Formal statistical test of γ_{text} bimodality (§7) on the existing $n = 25$ panel. Implementation: `diptest` package, 30 minutes CPU.

21.2 Open theoretical problems

$\nu = -1/(2\pi)$ first-principles derivation. The three convergent arguments supporting this constant (§5: dimensional, phase-storage heuristic, sign-trajectory bifurcation) do not yet constitute a formal proof. A complete derivation would involve:

1. Setting up a Lagrangian for the trained-attention distribution with constraints (i) normalization, (ii) MaxEnt $\mathbb{E}[\log d] = \text{const}$, (iii) phase coherence $\int p(d) \cos(d\,\omega_0)\, dd = C_{\text{phase}}(P)$.
2. Showing $C_{\text{phase}}(P) = (1/(2\pi))\, \log_{10}(P)$ from training-loss landscape analysis (open).
3. Linearization at the panel γ-region to extract $\nu = -1/(2\pi)$ exactly.

Step (1) is straightforward; the Phase-coherence integral has closed-form via Mellin transform. Step (3) is mechanical algebra once (2) is established. Step (2) — the rate of phase-coherence accumulation per parameter from gradient-descent dynamics — is the genuine open problem.

P_{critical} closed-form derivation. The IH formation threshold ([15, paper I, IH section]) at $P_{\text{critical}} \approx 400\,\text{M}$ on Pythia (D-DEEP-12) is empirical. A closed-form derivation from the bifurcating γ_{text} vs. γ_{random} trajectories during training would predict $P_{\text{critical}}(\theta, T_{\text{train}}, d_{\text{head}}, \cdots)$ across architectures. Current linear-trajectory models give $P_{\text{critical}} \approx 1.6$ (way too small), suggesting the trajectory is logistic (saturates) rather than linear. Required: fit $\gamma(\text{step})$ for γ_{text} and γ_{random} separately; identify functional form; intersect.

$P(\text{hallu})$ regularization at $\gamma \geq 1$. The Hagedorn-boundary anomaly (§18 D-DEEP-8): the formula gives $P(\text{hallu}) = 0$ at $\gamma = 1$ but Hagedorn-zone LLMs hallucinate empirically. Two candidate fixes:

1. Floor the exponent: $(d_h/L)^{\max(\varepsilon, 1-\gamma)}$ with $\varepsilon \in (0, 0.1)$ calibrated against Hagedorn-zone hallu data.
2. Rebuild the formula with the susceptibility factor entering multiplicatively in a different way that preserves the boundary.

Choice between (1) and (2) requires Hagedorn-zone hallucination benchmarks — not currently available in standard NIAH protocols.

Cross-derivation of empirical ν via direct γ regression. The clean test of ν uses the direct regression $\gamma_{\text{random}} \sim \log_{10}(P)$, not the ΔH_{Cardy} regression. The latter is confounded by per-model $z = T\sqrt{2}/\theta$ variation (§13). On the existing $n = 21$ random panel, direct regression gives slope -0.081 ($p = 0.10$), weaker than the Cardy regression's $+0.16$ but with the correct sign. Panel expansion should clarify whether the direct slope converges to $-1/(2\pi)$.

MoE imprint relation. On OLMoE (1.3B active / 7B total), the imprint formula $\gamma_{\text{random}} = \gamma_{\text{Padé}} + \nu \log_{10}(P)$ fails by ~ 0.6 for either P_{active} or P_{total} (§5). The effective parameter count for MoE imprint is unclear. Required: $n \geq 3$ MoE models tested at γ_{random}, fit empirical $P_{\text{eff}}^{\text{MoE}}$, characterize functional form.

21.3 Cross-paper integration directions

Bridge to fractional-attention architectures. Qu et al. [21] introduce Fractional Neural Attention as an architectural primitive operating in the standard Lévy regime $\alpha_{\text{Lévy}} \in (0, 2)$. Our framework predicts that existing RoPE attention occupies $\alpha_{\text{Lévy}} \in (-1, 0)$ (smoothing regime) approaching the Lévy critical point at the Hagedorn boundary (§8). A direct comparison between FNA's empirical performance and our prediction at $\gamma = 1 + \alpha_{\text{FNA}}$ would test the bridge: does designed-fractional attention align with our γ-target? Cost: $\sim$ 4h to set up FNA reference implementation and measure γ, plus comparison.

Bridge to specialization-coordination dynamics. Three of the five paper A bridges (14) remain underexplored. Bridge #2 ($\hat{O}$-peak vs. γ-flow inflection) requires re-running the γ measurement protocol on the 35 Pythia checkpoints underlying paper A's CKA analysis. Bridge #3 (inner-zone CKA $\to$ 1 as Hagedorn saturation) survives only partially in our pythia-1b 4-checkpoint test (§15); cross-validation on a wider checkpoint set would resolve.

Bridge to universal information chain. Marín [17] (paper B) frames language as one of several domains exhibiting Universal Information Structure (UIS): spectral, Bayesian, MDL-compressed, Gödelian. Paper 2's $\nu = -1/(2\pi)$ imprint and the $c_{\text{central}} = 3$ result ([15, paper I, Appendix A]) instantiate clauses (i) spectral and (iii) MDL specifically. A cross-domain test would seek analogous constants in non-language UIS instantiations (e.g., spectral analog of ν in audio, in image, in biological sequence data); a non-trivial match would support the universality claim.

21.4 Tooling roadmap

TAF Forge — inverse-design tool. The closed-form $\theta_{\text{design}}(\gamma_{\text{target}}, T)$ inversion (§13) plus the $N^*(\theta, T)$ K-optimal predictor (§14) plus the PDI diagnostic suffice for an inverse-design tool: given target γ, target T, and budget P, output the recommended architecture $(\theta, d_{\text{head}}, N_{\text{layers}}, N_{\text{params}})$ and report PDI / Hagedorn-zone proximity. Estimated dev: $\sim 16 - 22$h ($\sim 5 - 7$ days). Companion to the existing TAF Agent.

TAF dataset registry expansion. The current registry (`karlexmarin/taf-attention-decay`, $n = 58$ measurements / 32 models) will be extended with the panel-expansion measurements above. Future: per-checkpoint γ trajectories for Pythia-70M, GPT-NeoX, OLMo to enable training-dynamics meta-analysis without re-running.

Workshop note: Closed-form γ-decay from active RoPE band counting. Optional 6–8pp companion paper for ICLR Tiny Papers or NeurIPS workshop. Focuses tightly on the $\gamma_{\text{Padé}}$ = (active band count via aliasing geometry) chain, citing the paper 2 ν imprint and Cardy linear results as evidence. Strengthens the analytic claim independently of paper 2's broader honest-revisions framing.

21.5 Open puzzles

A small set of empirical findings remains unexplained by the framework as it stands. They are documented here as open puzzles, not as limitations of paper II's claims (which are scoped to the panel and methodology described).

Yi family persistent anomaly. Yi-6B and Yi-9B (architecture: $n_{\text{kv}} = 4$, $\theta = 5\text{M}$, $T_{\text{train}} = 4096$) measure $\gamma_{\text{text}} \approx 0.46$ and $\gamma_{\text{text}} \approx 0.55$ respectively. The architectural concentration law (§9) predicts $\gamma_{Pad}(\theta, T) - 0.012 \cdot 4 \approx 0.99 - 0.048 = 0.94$. The deviation $\Delta\gamma \approx -0.48$ is the largest single residual in the panel and is not captured by either Padé baseline or any of the six axes in the decomposition.

Sesion 32 update (random corpus): Yi-9B random-tokens measurement yields $\gamma_{\text{random}} = 1.061$ (CI $[0.916, 1.175]$, $R^2 = 0.988$, $n = 30$). On the random corpus the Yi anomaly *disappears*: $\gamma_{\text{random}} \approx 1.06$ matches the Padé saturated regime ($\gamma_{Pad} \to 1$ as $z \to 0$). This isolates the anomaly to text corpus only, supporting the *tokeniser-driven* hypothesis below over the LayerNorm-topology and extreme-θ alternatives.

Hypotheses worth pursuing:

- *Tokeniser-driven*: Yi tokeniser is Chinese-heavy; cross-corpus γ measurements never run.
- *LayerNorm topology*: Yi family uses non-standard post-LayerNorm; effect on γ untested.
- *Extreme θ outlier*: $\theta = 5\text{M}$ with $T_{\text{train}} = 4096$ gives $z = T/\theta = 8 \times 10^{-4}$, deeper into the Padé saturated regime than any other panel model. The architectural law's slope may not extrapolate to this regime.

Resolution would require a controlled experiment with Yi-architecture trained on standard (non-Chinese-heavy) corpus, or alternatively, re-measuring Yi on multiple corpora to disentangle tokeniser from architecture.

Methodological dark zones. Several training-pipeline axes were not controlled in the panel and remain unmeasured contributors to γ:

- *Tokeniser choice* (BPE, Unigram, SentencePiece, custom)
- *Optimizer* (AdamW dominant; Lion/SGD untested)
- *LR schedule* (cosine dominant; WSD/constant untested)
- *Batch size* (correlated with N in panel)
- *Initialisation scheme* (μP vs standard, $n = 1$ measurement at step zero)
- *Continued pre-training* ($n = 3$–4 examples; not statistical)
- *Gradient clipping* (clipping at 1.0 dominant; unmeasured at 0.1 or 10)
- *Mixed precision* (bf16 dominant; fp16 outliers documented but not panel-tested)
- *Sequence packing* (modern training uses; document-mask correctness assumed but not audited)

• *Quantisation-aware training* (rare in panel; could change 4-bit shift R^2-direction rule)

Each of these is an axis where TAF claims have not been validated. Future panel expansion should target controlled comparison pairs (same architecture, varying axis) to isolate per-axis γ-contribution.

Famous-constant clustering anomaly. Four panel models sit within 0.01 of named irrational constants: CodeLlama-13B at $\gamma \approx 1 - 1/\varphi$ (golden-conjugate), Pythia-1.4B at $\gamma \approx 1/\sqrt{2}$, Llama-2-7B at $\gamma \approx 1 - 1/\sqrt{2}$ (the AM-GM anchor of Theorem 10.2), and Mistral-Nemo at $\gamma \approx \log_{10}(e)$. With four candidate constants and 22 panel models at tolerance 0.01, the expected number of matches by chance is ~ 0.88; observed 4 gives $p \approx 0.018$ (Poisson), marginal and not Bonferroni-significant. We do not commit to any mechanism explaining this clustering. If real, the AM-GM anchor ($\gamma_{\min} = 1 - 1/\sqrt{2}$ for Llama-2-7B) is the only one with algebraic standing in the framework. Resolution requires panel expansion to test whether the apparent clustering survives.

Formal verification (Lean Mathlib4 + Sage Groebner). The 15 algebraic identities of TAF critical exponents (Appendix A.4) have been independently verified by two complementary tools: (i) **Sage Groebner basis** as an algebraic decision procedure in PolynomialRing($\mathbb{Q}$) with the defining ideal of TAF exponents, and (ii) **Lean Mathlib4** as machine-checked proof in dependent type theory (build success: 1973/1973 jobs, all `ring` / `field_simp` tactics pass). Both tools detected that the original "triple closure" claim was incorrect (Rushbrooke + Josephson are tautologies; Fisher is independent), and replaced it with the new identity **D-SAGE-1**: $2\eta^2 + \eta \cdot \gamma_\chi + 1 = 0$, which algebraically refutes paper 1's $\eta = 2\gamma$ claim and pins $\eta = \gamma - 1$ as the correct anomalous dimension. This is arguably the first transformer-attention framework with formal machine-proof backing. Lean code: `lean_taf/taf/Taf/Identities.lean`; Sage script: `analysis/sage_recursive_sweep_2026-04-30.sage`. A future 3-month effort could extend the formal verification to the analytical claims (AM-GM bound $\gamma_\chi \geq 2\sqrt{2}$, Cardy slope $= 2\,\mathrm{arctanh}(\gamma)$, and $\log(\mathrm{PDI}) + \Delta H_{\mathrm{Cardy}} = 0$) for an independent paper III *"Formally verified TAF identities"* suitable for ITP / CPP / JAR venues. The TAF Agent v0.5 (`verify_algebraic_consistency`) implements all 12 D-SAGE checks as runtime sanity tests on user-input γ values, flagging bf16 outliers / quantization artifacts when any check fails.

22 Conclusion

Paper II completes a phenomenological framework for transformer attention decay along five constructive axes (precision-direction rule, learned-imprint axis, Cardy entropy decomposition, bimodal phase structure, and machine-verified algebraic backbone) and records seven honest revisions of earlier claims. The strongest empirical claim is the precision-direction rule: 5/5 sign-correct on bf16/4-bit paired measurements of full multi-head attention, mechanistically grounded in NF4 truncation of long-tail attention mass, and immediately actionable for practitioners deploying 4-bit inference. The strongest theoretical claim is the quadratic identity **D-SAGE-1**: $2\eta^2 + \eta\,\gamma_\chi + 1 = 0$, verified twice independently (Sage Groebner basis and Lean Mathlib4). Beyond these, the architectural concentration relation $\gamma_{\text{text}} \approx \gamma_{Pad} - 0.012\,n_{kv}$ is documented as an in-sample correlation that does not generalise out-of-sample, and the imprint slope $\nu \approx -1/(2\pi)$ as a phenomenological prediction with wide CI and an honest within-model null.

The methodological contribution may be the more transferable one: a five-stage audit pipeline (Sócratic interrogation, Sage Groebner basis, empirical validation, literature scan, triangulation), with optional Lean Mathlib4 cross-check, applied to all 15 algebraic identities of the framework. The pipeline is software-implementable, domain-agnostic, and produces a written record of failures (§18) instead of silent dropping.

We do not claim a fundamental theory of transformers. The $n = 22$–32 panel is moderate, single-corpus, AdamW + cosine-LR-dominated, and $\sim 70\%$ of the cross-panel variance in γ_{text}

remains unexplained. Mixture-of-experts, multi-modal models, state-space models, and cross-corpus comparability are explicitly out of scope. What this paper offers is a leaner, more honest version of paper I: stronger where the audit confirms, demoted where it does not, and machine-verifiable where the algebra permits.

The dataset (`karlexmarin/taf-attention-decay`, 79 records), diagnostic tool (`karlexmarin/taf-agent`), and Lean proof code are released alongside this paper for independent verification and reproduction.

A Formal verification of TAF algebraic identities

Setup (Phase A: $\gamma \in (0, 1)$)

The TAF critical exponents (with hyperscaling consistency at $d = 1$) are:

$$\nu_c = \frac{1}{1-\gamma} \qquad \beta_c = \gamma - 1 \qquad \eta_c = \gamma - 1$$
$$\alpha_C = 2 - \frac{1}{1-\gamma} \qquad \chi = \frac{1}{1-\gamma} \qquad \gamma_\chi = \frac{1}{1-\gamma} + 2(1-\gamma)$$

The 15 algebraic identities below have been independently verified by two complementary tools: (i) **Sage Groebner basis** as an algebraic decision procedure in PolynomialRing($\mathbb{Q}$), and (ii) **Lean Mathlib4** as machine-checked proof in dependent type theory. Lean build succeeds: 1973/1973 jobs.

Master table

#	Theorem	Statement	Step-by-step proof
1	**D-SAGE-1 ★★**	$2\eta^2 + \eta \cdot \gamma_\chi + 1 = 0$	$\eta = \gamma - 1 \Rightarrow 2\eta^2 = 2(\gamma-1)^2$; $\eta\gamma_\chi = (\gamma - 1)/(1-\gamma) + 2(\gamma-1)(1-\gamma) = -1 - 2(\gamma-1)^2$; sum: $2(\gamma-1)^2 - 1 - 2(\gamma-1)^2 + 1 = 0$
2	**D-SAGE-2**	$\beta \cdot \chi = -1$	$\beta\chi = (\gamma-1)/(1-\gamma) = -1$ since $\gamma - 1 = -(1-\gamma)$
3	**D-SAGE-3**	$\nu \cdot \beta = -1$	$\nu = \chi$ in Phase A; equivalent to D-SAGE-2
4	**D-SAGE-4**	$\alpha_C + \chi = 2$	$[2 - 1/(1-\gamma)] + 1/(1-\gamma) = 2$ (cancellation)
5	**D-SAGE-5**	$\alpha_C + \gamma_\chi = 2(2-\gamma)$	$[2 - 1/(1-\gamma)] + [1/(1-\gamma) + 2(1-\gamma)] = 2 + 2(1-\gamma) = 4 - 2\gamma$
6	**D-SAGE-6**	$\beta \cdot \gamma_\chi = -2\gamma^2 + 4\gamma - 3$	$\beta\gamma_\chi = -1+2(\gamma-1)(1-\gamma) = -1+(4\gamma-2\gamma^2-2)$
7	**Rushbrooke** (tautology)	$2\beta + \gamma_\chi - \nu = 0$	$2(\gamma-1) + 1/(1-\gamma) + 2(1-\gamma) - 1/(1-\gamma) = 2\gamma - 2 + 2 - 2\gamma = 0$
8	**Josephson** (tautology)	$2 - \alpha_C - \nu = 0$	$2-[2-1/(1-\gamma)]-1/(1-\gamma) = 0$ (cancellation)
9	**Fisher residual**	$\gamma_\chi - (2-\eta)\nu = \gamma(2\gamma-3)/(1-\gamma)$	$(2-\eta)\nu = (3-\gamma)/(1-\gamma)$; $\gamma_\chi - (2-\eta)\nu = (\gamma-2)/(1-\gamma) + 2(1-\gamma) = [(\gamma-2) + 2(1-\gamma)^2]/(1-\gamma) = \gamma(2\gamma-3)/(1-\gamma)$
10	**fisher_zero_iff ★**	Fisher $= 0 \iff \gamma \in \{0, 3/2\}$	From #9: numerator $\gamma(2\gamma-3) = 0$. Phase A excludes both
11	**eta_2gamma_res**	$2(2\gamma)^2 + 2\gamma\gamma_\chi + 1 = (-4\gamma^3 + 5\gamma + 1)/(1-\gamma)$	$8\gamma^2 + 2\gamma/(1-\gamma) + 4\gamma(1-\gamma) + 1$; $\times(1-\gamma)$: $(2\gamma+1)^2(1-\gamma) + 2\gamma = -4\gamma^3 + 5\gamma + 1$
12	**eta_2gamma_fail**	residual $> 0\ \forall\gamma \in (0,1)$	$\gamma^3 < \gamma$ on $(0,1) \Rightarrow -4\gamma^3 + 5\gamma + 1 > \gamma + 1 > 1 > 0$. Refutes paper 1's $\eta = 2\gamma$
13	**D-14**	$\nu_{\text{imprint}} \cdot 2\pi = -1$	$(-1/(2\pi)) \cdot 2\pi = -1$ (paper 1 base, dimensional)
14	**D-SAGE-7 ★**	$c \cdot \lvert\nu_{\text{imprint}}\rvert \cdot 2\pi = 3$	$3 \cdot (1/(2\pi)) \cdot 2\pi = 3$. Dimensional closure
15	**nu_beta_id**	$\nu \cdot \beta = -1$	$\nu\beta = (1/(1-\gamma))(\gamma-1) = -1$

Verification methodology

Sage Groebner basis (algebraic decision). Defining ideal in PolynomialRing($\mathbb{Q}, [g, \nu, \beta, \alpha, \eta, \gamma_\chi]$) with the 5 generators $\{\nu(1-g) - 1,\ \beta - g + 1,\ \eta - g + 1,\ \alpha + \nu - 2,\ \gamma_\chi - \nu - 2(1-g)\}$. The Groebner basis includes the polynomial $\eta^2 + \frac{1}{2}\eta\gamma_\chi + \frac{1}{2}$, equivalent to **D-SAGE-1**. Each identity reduces to 0 in the ideal $\iff$ identity holds. Fisher reduces to $-4\eta - \gamma_\chi - 1$, NOT in the ideal $\Rightarrow$ Fisher is independent (not implied by Rushbrooke + Josephson).

Lean Mathlib4 (dependent type theory). Each theorem is proven via tactic combination `unfold; field_simp; ring`. The `ring` tactic is a decision procedure for commutative rings: any polynomial equality can be proven automatically. Lean kernel verifies the resulting proof term in the Calculus of Inductive Constructions.

Reproducibility.

```
# Sage verification (~30s)
```

```
docker run --rm -v "$(pwd)/analysis:/work" sagemath/sagemath:latest \
    sage /work/sage_recursive_sweep_2026-04-30.sage

# Lean verification (~10min first build, cached after)
docker run --rm -v "$(pwd)/lean_taf:/work" \
    leanprovercommunity/lean:latest \
    -c "cd /work/taf && lake build"
```

Build success: 1973/1973 jobs (Mathlib4 + 15 TAF theorems), exit code 0. Sage results: `analysis/sage_recursive_sweep_results.json`. Lean code: `lean_taf/taf/Taf/Identities.lean`.

Lit-scan novelty assessment

Status	Count	Theorems
NOVEL TAF-specific	12	D-SAGE-1, 2, 3, 4, 5, 6, 7 + fisher_residual + fisher_zero_iff + eta_2gamm
Classical reframed	2	rushbrooke + josephson (1963 classical, framed as tautologies)
Paper 1 base	1	D-14 ($\nu \cdot 2\pi = -1$, from Marin paper 1)

Searched: Wikipedia critical exponents, Scholarpedia scaling laws, Cambridge phase notes, MIT statistical mechanics notes, arXiv 2024-2026 critical exponents + transformer attention. NONE of the 12 TAF-specific identities appear in classical critical phenomena literature. Reason: TAF uses non-standard exponent system ($\beta = \gamma - 1 < 0$ in Phase A regime). Concurrent paper Nakaishi et al. [18] (arXiv 2510.05554) uses different formalism (rank-collapse, not γ-decay) — no overlap. Concurrent Tanaka et al. [26] (arXiv 2510.02458) on 1D long-range Ising derives $\eta(\sigma) = 2 - \sigma$ in different convention (maps to our $\eta = 3 - \gamma$, not $\gamma - 1$) — no overlap.

Significance. This appendix documents the first machine-verified catalog of TAF-specific algebraic identities. The dual-tool verification (Sage Groebner basis + Lean Mathlib4) detected and corrected an earlier informal claim of "triple closure" (Rushbrooke + Josephson + Fisher all closing simultaneously, which was *wrong*: only D-SAGE-1 quadratic is the genuine algebraic constraint, with Fisher being independent). The TAF Agent v0.5 web tool implements all 12 D-SAGE checks as runtime sanity tests via the `verify_algebraic_consistency` function.

B Formula × formula comparison catalogue

This appendix documents 24 pairwise comparisons of TAF formulas performed during the Sócratic audit (Sesion 32 V11/V12). Each comparison classifies the relation between two formulas as **equivalent**, **contradictory**, **independent**, **nested** (one is special case of other), **coincidence in limits**, or **related via calculus**. The catalogue serves as a structural map of the framework: it makes explicit which formulas are tied together algebraically, which are independent, and which contradict (revealing errors).

#	Pair	Status	Reasoning
1	$\eta = \gamma - 1$ vs $\eta = 2\gamma$ (paper 1)	Contradictory	D-SAGE-1 satisfied by $\gamma - 1$, not 2γ (residual > 0 Phase A).
2	$\gamma_\chi = \frac{1}{1-\gamma} + 2(1-\gamma)$ vs $c = 3$ (γ=0)	Coincidence at limit	$\gamma_\chi(0) = 3 = c_{\text{central}}$; c is special value of γ_χ.
3	d_{horizon} vs T_{train}	Independent	d_{horizon} from geometry; T_{train} from training. PDI ratio relates them as diagnostic.
4	$\nu_{\text{cross-panel}} = -1/(2\pi)$ vs $\nu_{\text{within}} = -0.080$	Same mechanism, different scale	Reconciled by Chinchilla $D = 20N$ rescaling (factor $\sqrt{20}/\sqrt{2}$).
5	$\gamma_{Pad}[1,1]$ vs $\gamma_{Pad}[2,2]$	Coincidence at $z \ll 1$	Saturated for LLM regime ($\lvert err \rvert = 9.39\%$ vs 9.33%).
6	Architectural law vs *Lvy* α	Independent levels	Arch is L0 phenomenology; *Lvy* is L6 interpretation.
7	ΔH_{Cardy} vs χ	Related via calculus	$d\Delta H/d\gamma = 2\chi/(1+\gamma)$. Antiderivative relationship.
8	D-SAGE-1 vs Fisher relation	Nested (D-SAGE-1 $\supset$ Fisher zeros)	D-SAGE-1 is identity $\forall\gamma$; Fisher is the specific point where curve crosses zero.
9	AM-GM bound $2\sqrt{2}$ vs $c = 3$	Bound vs anchor	$3 > 2\sqrt{2}$; γ=0 is above floor; $\gamma_{\min}$=$1-1/\sqrt{2}$ saturates floor.
10	$z^* = (\sqrt{17}-3)/2$ vs $1-1/\varphi$	Independent algebraic	Different minimal polynomials; both $\in \overline{\mathbb{Q}}$, distinct.
11	γ_{text} correlational vs γ_{Pad} predictive	Nested (text $\supset$ Padé baseline)	Text is fit; Padé is theory-baseline. R^2: 0.30 vs 0.021.
12	NIAH ceiling vs P_{hallu}	Complementary, distinct facets	Both depend on d_{horizon}. NIAH = success rate; P_{hallu} = failure rate.
13	Fisher info $I = C_V/\gamma^2$ vs χ	Independent geom roles	Fisher diverges at $\gamma \to 0$ AND $\gamma \to 1$; χ only at γ=1.
14	D-SAGE-1 vs RG flow fixed points $\{0, 1\}$	Compatible	At γ^*=0: D-SAGE-1 is $2{\cdot}1+(-1){\cdot}3+1 = 0$ ✓. At γ^*=1: limit consistent.
15	K invariant vs Chinchilla $D = 20N$	K assumes Chinchilla	K-band 51.2 ± 16.8 valid only in Chinchilla-trained regime.
16	Padé $z^3/12$ vs $\nu_{\text{imprint}} = -1/(2\pi)$	Independent constants	$1/12 \neq 1/(2\pi)$; no algebraic link evident.
17	DeepSeek bf16 (γ=0.149) vs DeepSeek 4-bit (γ=0.95)	Same model, opposing precisions	R^2-direction rule confirmed: bf16 R^2=0.49 $\to$ 4-bit shifts UP.
18	Llama-2-7B $1-1/\sqrt{2}$ vs CodeLlama-13B $1-1/\varphi$	Same form, distinct constants	Both "$1-1/x$" but x irrationals algebraically distinct.
19	$\nu = -1/(2\pi)$ derived vs empirical fit -0.158	Coincidence within CI	0.6% relative error validates 3-argument derivation.
20	Famous-constant cluster vs random null	Marginal ($p = 0.018$)	4 of 88 chances at $\lvert\Delta\rvert \leq 0.01$. Not Bonferroni-significant.
21	D-DEEP-15 $\log(\text{PDI}) + \Delta H = 0$ vs PDI definition	Algebraic identity (Phase 7 form)	Both forms reconcile at $\Delta H = \log(z/\sqrt{2}) + 2\,\text{arctanh}(\gamma)$.
22	Architectural law vs 6-axis decomposition	Nested (6-axis collapses to 2-axis)	Only n_{kv} significant beyond Padé; R^2 identical.
23	$\eta = \gamma - 1$ (D-SAGE-1) vs $\alpha_{Lvy} = \gamma - 1$ (concurrent paper)	Coincidence (overdetermination)	Two independent paths converge to $\gamma - 1$; robust evidence.
24	AM-GM bound vs Hagedorn $\gamma \to 1$ divergence	Floor-ceiling pair	$\gamma_\chi \in [2\sqrt{2}, \infty)$ on Phase A.

Distribution of statuses.

Status	Count
Equivalent / coincidence at limits	6
Contradictory	1
Independent	5
Nested (one $\supset$ other)	4
Same mechanism distinct scale	2
Related via calculus	1
Marginal/statistical	2
Compatible / complementary	3
Total	**24**

Implications. Five comparison patterns deserve highlighting:

1. **Over-determination of $\eta = \gamma - 1$** (#23): two *independent* derivation paths converge — both D-SAGE-1 algebraically and the Lévy framework [21] arrive at the same value. This is robust evidence beyond either path alone.

2. **Phase A floor-ceiling structure** (#24): $\gamma_\chi \in [2\sqrt{2}, \infty)$ across the full Phase A regime. The susceptibility is bounded below (AM-GM, Theorem 10.2) and unbounded above (Hagedorn divergence).

3. **D-SAGE-1 supersedes Fisher** (#8): the quadratic identity contains all valid algebraic content of the Fisher hyperscaling relation; the simpler form replaces three.

4. **Famous-constant cluster is statistically marginal** (#20): explicit $p = 0.018$ value before Bonferroni — does not survive multiple testing correction. Catalogued as curiosity, not law.

5. **6-axis decomposition collapses to 2-axis** (#22): nature prefers simplicity. θ via Padé and n_{kv} suffice; other axes do not contribute meaningfully to R^2.

The full catalogue is documented at greater length in `analysis/SOCRATIC_FORMULA_INTERROGATION_2026-04` in the project repository, including detailed reasoning per comparison.

C Pre-release peer feedback and editorial response

This appendix documents pre-release peer-style feedback that we received and the editorial actions we took. We include this because (i) it is itself an instance of the audit working as intended, (ii) it shows what *not* to claim, and (iii) it preserves the constructive critique for future readers.

Strengths confirmed

The pre-release reviewer confirmed four points:

- Architectural concentration law $\gamma_{\text{text}} \approx \gamma_{Pad} - 0.012\, n_{\text{kv}}$ has *practical engineering value*: GQA pushes γ toward 1; MHA toward broader distributions. Industry empiricism aligns.
- Precision sensitivity (4-bit vs bf16) is *undervalued but critical*: cuantisation can artefactually create power laws, especially in MHA with attention sinks.
- Honest demotions section signals "proceso científico real, not marketing".
- "γ as state variable" framing is *conceptually powerful*: reduces the complexity of attention behaviour to a single interpretable scalar.

Weaknesses identified — and our editorial responses

$\nu = -1/(2\pi)$ ***numerologically attractive vs. genuinely derived.*** The bootstrap CI on the empirical ν is wide ($[-0.260, -0.008]$); pythia-31m drives much of the slope. The reviewer flagged this as "constante bonita que encaja aproximadamente" rather than "ley física derivada". **Our response:**

- We tone down the framing. §5 now explicitly states "We do not claim a complete first-principles derivation, but three independent arguments converge on this value."
- We launched (sesion 32, 2026-05-01) a Pythia-70M trajectory experiment across 9 checkpoints (step1000 to step143000) to test whether ν converges *monotonically* toward $-1/(2\pi)$ during training. **Honest null result**: within-model $d\gamma_{\text{random}}/d\log(\text{step})$ does *not* monotonically approach $-1/(2\pi)$. The within-model trajectory shows γ_{random} increasing from ~ 1.0 to ~ 1.1 over training and saturating; only the very last checkpoint pair (step100000→step143000) yields a local slope of -0.152 near the predicted value, indistinguishable from noise at $n = 2$. The cross-panel slope $\nu = -1/(2\pi)$ is therefore a *cross-architecture / cross-scale phenomenon*, not a within-training trajectory toward an asymptote.
- This null result *strengthens* the phenomenological framing: $\nu = -1/(2\pi)$ is an empirical regularity across the parameter-count axis, not an attractor reachable by training a single model. The result aligns with the reviewer's concern that the cross-panel ν is more "empirical organisation" than "derived law".
- We mark $\nu = -1/(2\pi)$ as a phenomenological prediction under continued empirical scrutiny — the within-model null is one honest data point against over-interpretation.

MaxEnt is "heurístico, no derivado de entrenamiento". The reviewer correctly identifies this as a foundational gap. We had already documented this in §17 but the implications were under-emphasised. **Our response:**

- Strengthen language: MaxEnt is a *useful heuristic* for organising observed regularities, not a derivation from gradient-descent dynamics.
- Future work (§21) includes connecting imprint dynamics to actual training trajectories (loss, gradient norms) — the path that would replace MaxEnt with a derivation.

Low R^2 values: ~ 0.27 for imprint, ~ 0.30 for n_{kv}. Reviewer: "más dirección correcta que teoría cerrada". Correct. **Our response:**

- Abstract now explicitly states "$R^2 = 0.30$ — the dominant correction beyond Padé baseline, but $\sim 70\%$ of variance remains unexplained".
- Yi family persistent anomaly catalogued as the single biggest unresolved residual in §21.5.

MoE breaks the framework. The reviewer correctly notes that imprint and parameter counting ambiguate for Mixture-of-Experts (MoE) models. **Our response:**

- Explicitly out of scope for paper II. MoE handling proposed as paper III direction: P_{eff} = entropy-weighted active parameters.
- Mixtral measurement in dataset is currently $\gamma_{\text{global}} = 0.92$ without per-expert breakdown — admitted as limitation.

"Mezcla peligrosa" of physics language and weak regression. Reviewer flagged the risk of overinterpretation: terms like "universality class" and "Cardy law real" suggest more than the $n = 22$–32 panel can support. **Our response:**

- Tone-down audit performed: "anti-Ising universality class" → "anti-Ising-*like* regime"; "Cardy law" → "Cardy-*form* decomposition"; "Hagedorn essential singularity" framed as "Hagedorn-analogue boundary".
- Abstract explicitly labels the framework as "phenomenological — a quantitative organisation of empirical regularities, not a derivation from first principles".
- Five surviving regularities described as "robust phenomenological regularities, candidates for universality contingent on future work" rather than "universal candidates" simpliciter.

Six paper III directions accepted from the critique

The reviewer's concrete suggestions are accepted as paper III roadmap:

1. **Distribution-level modelling** (not just γ exponent): mixture of power-law + attention sink + cutoff; tempered Lévy stable distributions. Would explain 4-bit chaos mechanistically.
2. **Training-dynamics connection**: measure $\gamma(t)$, gradient norms, attention entropy during training. Find $d\gamma/dt = f(\text{loss}, \text{grad}, \text{scale})$. *Would close the MaxEnt gap.*
3. n_{kv} **as information-theoretic**: replace heuristic with mutual information across heads / KV bank redundancy. Hypothesis: γ proportional to attention-space compression.
4. **MoE rigorous handling**: P_{eff} = entropy-weighted active parameters. Brings Mixtral, DeepSeek-MoE into the framework.
5. **Synthetic-task validation**: small transformers, controlled tasks. Separates causality from correlation (current panel is real-LLM noisy).
6. **Permanent linguistic discipline**: maintain analogues ("-like", "-form") as descriptive tools, never as load-bearing universality claims.

Reflection

The critique was the strongest single editorial signal we received on paper II. Its incorporation is the difference between a paper that overclaims and a paper that calibrates. We thank the reviewer indirectly: tools and people that catch errors before publication are how the framework remains honest.

References

[1] Federico Barbero et al. Round and round we go! what makes rotary positional encodings useful? In *ICLR*, 2025.

[2] Stella Biderman et al. Pythia: A suite for analyzing large language models across training and scaling. In *COLM*, 2024.

[3] J. L. Cardy. Operator content of two-dimensional conformally invariant theories. *Nuclear Physics B*, 270:186–204, 1986. doi: 10.1016/0550-3213(86)90552-3. Original Cardy formula derived from modular invariance of CFT on torus.

[4] M. E. Fisher. The theory of equilibrium critical phenomena. *Reports on Progress in Physics*, 30(2):615–730, 1967. doi: 10.1088/0034-4885/30/2/306. Original Fisher relation $\gamma = (2-\eta)\nu$ + comprehensive scaling laws review.

[5] Alireza Ghafarollahi and Markus J. Buehler. SciAgents: Automating scientific discovery through multi-agent intelligent graph reasoning, 2024.

[6] R. B. Griffiths. Thermodynamic inequality near the critical point for ferromagnets and fluids. *Physical Review Letters*, 14(16):623–624, 1965. doi: 10.1103/PhysRevLett.14.623. Griffiths inequality $\alpha + \beta(\delta + 1) \geq 2$.

[7] J. A. Hartigan and P. M. Hartigan. The dip test of unimodality. volume 13, pages 70–84, 1985.

[8] Jordan Hoffmann et al. Training compute-optimal large language models. 2022. Chinchilla scaling law, $D = 20N$ token-parameter ratio.

[9] ICLR 2026 Submission. Frequency bands in RoPE: Base frequency and context length shape the interpolation–extrapolation trade-off, 2025. URL https://openreview.net/forum?id=PR1PPxvG9Q. OpenReview ID PR1PPxvG9Q.

[10] B. D. Josephson. Inequality for the specific heat: Derivation. *Proceedings of the Physical Society*, 92(2):269–275, 1967. Original Josephson hyperscaling relation $2 - \alpha = \nu d$.

[11] J. Kaplan et al. Scaling laws for neural language models. 2020. Original $6ND$ FLOP estimate for transformer training compute.

[12] X. Liu et al. Scaling laws of RoPE-based extrapolation. 2023. Establishes $d_{\text{extra}}/d = \log_\theta(T_{\text{train}}/2\pi)$ for RoPE.

[13] Chris Lu et al. The AI Scientist: Towards fully automated open-ended scientific discovery, 2024.

[14] Carles Marín. Universal specialization-coordination structure in language model training. 2026. doi: 10.5281/zenodo.19199838. URL https://zenodo.org/records/19199838. Paper A. Pythia checkpoint specialization-coordination dynamics.

[15] Carles Marín. Predicting how transformers attend: Analytic power-law theory, phase transitions, and practical compression tools. 2026. doi: 10.5281/zenodo.19826343. URL https://zenodo.org/records/19826343. Companion paper I; this work is paper II.

[16] Carles Marín. TAF Attention-Decay Dataset, 2026. URL https://huggingface.co/datasets/karlexmarin/taf-attention-decay. $n = 58$ measurements / 32 models / 12 families. CC-BY-4.0.

[17] Carles Marín. The universal information chain. 2026. doi: 10.5281/zenodo.19224167. URL https://zenodo.org/records/19224167. Paper B. Cross-domain bridges via CAAS architecture.

[18] K. Nakaishi et al. Critical attention scaling in long-context transformers. 2025.

[19] Catherine Olsson et al. In-context learning and induction heads. Anthropic Transformer Circuits Thread, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/. Induction-head emergence threshold $N \geq$ 400M; alias for olsson2022induction.

[20] A. Pelissetto and E. Vicari. Scaling laws. Scholarpedia, 2025. URL http://www.scholarpedia.org/article/Scaling_laws. Standard reference on scaling laws and hyperscaling relations.

[21] Cheng Kevin Qu, Andrew Ly, and Pulin Gong. Fractional neural attention for efficient multiscale sequence processing, 2025. URL https://arxiv.org/abs/2511.10208. Concurrent work: introduces fractional Laplacian as architectural primitive; complementary to our analytic γ_{Pad} prediction (this paper, Sec. spectral_reframe).

[22] F. Ruscio et al. Beyond position: Wavelet-like properties spontaneously emerge in transformer attention. In *ACL*, 2025.

[23] G. S. Rushbrooke. On the thermodynamics of the critical region for the Ising problem. *Journal of Chemical Physics*, 39(3):842–843, 1963. doi: 10.1063/1.1734329. Original Rushbrooke inequality $\alpha + 2\beta + \gamma \geq 2$.

[24] Theodore R. Sumers et al. Cognitive architectures for language agents (CoALA), 2023.

[25] X. Sun et al. When attention sink emerges in language models: An empirical view. In *ICLR*, 2025.

[26] H. Tanaka et al. One-dimensional long-range Ising model: two (almost) equivalent approximations, 2025. URL https://arxiv.org/abs/2510.02458. 1D long-range Ising; derives $\eta(\sigma) = 2 - \sigma$ in different convention.

[27] B. Widom. Equation of state in the neighborhood of the critical point. *Journal of Chemical Physics*, 43(11):3898–3905, 1965. doi: 10.1063/1.1696618. Widom scaling hypothesis; relation $\gamma = \beta(\delta - 1)$.

[28] Wikipedia contributors. Critical exponent, 2025. URL https://en.wikipedia.org/wiki/Critical_exponent. Reference list of standard critical exponents and their hyperscaling relations.

[29] G. Xiao et al. Efficient streaming language models with attention sinks, 2023.