Erratum to:
"Predicting How Transformers Attend"
§5.2 Theorem 5.2 — Heat capacity coefficient at $\gamma = 1$

Carles Marín
Independent Researcher
transformerkmarin@gmail.com

2026-05-02

Erratum summary

The original paper [1] states in §5.2 Theorem 5.2 that the heat capacity at the Hagedorn point obeys $C_V(\gamma{=}1, N) = (\log N)^2/4$. A self-audit using the Sócrates framework [3] (numerical verification + Sage exact rational arithmetic + SymPy symbolic integration) shows the correct asymptotic large-N coefficient is 1/12, not 1/4, i.e.,

$$C_V(\gamma{=}1, N) \xrightarrow{N\to\infty} \frac{(\log N)^2}{12}\,.$$

The error is a factor of three. It originates in the truncation of the Z-expansion at first order in $(1-\gamma)$. The qualitative conclusion ("C_V finite at Hagedorn for finite N, sharp in the $N \to \infty$ limit") is unaffected. The TAF Agent diagnostic tool [2] computes C_V via numerical derivative of the partition function, not via the analytic formula, and therefore reports the correct value automatically.

1 Original claim

§5.2 Theorem 5.2 of the original paper states:

> **Theorem 5.2 (Finiteness of C_V at Hagedorn).** For any finite N:
>
> $$C_V(\gamma{=}1, N) \;=\; \frac{(\log N)^2}{4}, \tag{1}$$
>
> which is finite for all $N < \infty$.

The original proof sketch reads:

> *Near* $\gamma = 1$: $Z(\gamma, N) \approx \log N \cdot (1 + (1-\gamma)\log N/2)$. *Then* $U = \log N/2 + (1-\gamma)(\log N)^2/4$. *Hence* $\mathrm{d}U/\mathrm{d}\gamma = -(\log N)^2/4$, *and* $C_V = -\gamma^2 \mathrm{d}U/\mathrm{d}\gamma = (\log N)^2/4$ *at* $\gamma{=}1$.

2 Where the proof goes wrong

The Taylor expansion of the partition function $Z(\gamma, N) = \sum_{d=1}^{N} d^{-\gamma}$ about $\gamma = 1$ is, to second order in $u \equiv 1 - \gamma$,

$$\begin{aligned} Z(\gamma, N) \;&=\; H_N \;+\; u\,(\log N)\,\frac{H_N}{2} \;+\; u^2\,(\log N)^2\,\frac{H_N}{6} \;+\; \mathcal{O}(u^3) \\ &\approx\; \log N \cdot \left[1 \;+\; \frac{(1-\gamma)\log N}{2} \;+\; \frac{(1-\gamma)^2(\log N)^2}{6} \;+\; \cdots\right], \end{aligned} \tag{2}$$

where $H_N = \sum_{d=1}^{N} 1/d \sim \log N + \gamma_E$ (Euler–Mascheroni). The original proof retained only the term linear in $(1-\gamma)$, dropping the quadratic term. Taking $\log Z$ and using $\log(1+x) = x - x^2/2 + \cdots$ gives

$$\log Z(\gamma, N) \;\approx\; \log(\log N) \;+\; \frac{(1-\gamma)\log N}{2} \;+\; \frac{(1-\gamma)^2(\log N)^2}{24} \;+\; \cdots, \tag{3}$$

where the second-order coefficient $1/24$ is the difference $1/6 - 1/8$ (the latter coming from the $-x^2/2$ in $\log(1+x)$).

Differentiating (3),

$$U(\gamma) \;=\; -\frac{\partial \log Z}{\partial \gamma} \;=\; \frac{\log N}{2} \;+\; (1-\gamma)\,\frac{(\log N)^2}{12} \;+\; \mathcal{O}\big((1-\gamma)^2\big)\,. \tag{4}$$

Hence

$$\left.\frac{\mathrm{d}U}{\mathrm{d}\gamma}\right|_{\gamma=1} \;=\; -\,\frac{(\log N)^2}{12}, \tag{5}$$

and

$$C_V(\gamma{=}1, N) \;=\; -\gamma^2 \left.\frac{\mathrm{d}U}{\mathrm{d}\gamma}\right|_{\gamma=1} \;=\; \frac{(\log N)^2}{12}\,, \tag{6}$$

which contradicts equation (1). The error is a factor of three.

3 Independent verification: cumulant identity

For the exponential family $p_\gamma(d) \propto d^{-\gamma}$ with sufficient statistic $\log d$ and natural parameter γ, the heat capacity equals the variance of the log-distance:

$$C_V(\gamma) \;=\; \gamma^2 \operatorname{Var}_{p_\gamma}(\log d)\,. \tag{7}$$

At $\gamma = 1$ the distribution is $p(d) = (1/d)/H_N$. Using the integral approximations $\int_1^N (\log x)/x\,\mathrm{d}x = (\log N)^2/2$ and $\int_1^N (\log x)^2/x\,\mathrm{d}x = (\log N)^3/3$, and $\int_1^N 1/x\,\mathrm{d}x = \log N$, one obtains

$$\mathbb{E}[\log d] \;\sim\; \frac{(\log N)^2/2}{\log N} \;=\; \frac{\log N}{2}, \tag{8}$$

$$\mathbb{E}[(\log d)^2] \;\sim\; \frac{(\log N)^3/3}{\log N} \;=\; \frac{(\log N)^2}{3}, \tag{9}$$

$$\operatorname{Var}(\log d) \;\sim\; \frac{(\log N)^2}{3} - \frac{(\log N)^2}{4} \;=\; \frac{(\log N)^2}{12}\,. \tag{10}$$

This independently gives (7) $\to (\log N)^2/12$ at $\gamma = 1$, confirming the corrected coefficient.

4 Triple-system numerical verification

The error was first detected by the Sócrates audit framework (numerical Python), then confirmed by two independent computer-algebra systems:

A finite-N check at $N = 10^4$ gives $C_V = 7.89$ (exact discrete computation), within 12 % of the leading asymptotic $(\log N)^2/12 \approx 7.07$ — the residual is the $\mathcal{O}(\log N)$ sub-leading correction. The original coefficient $1/4$ predicts 21.2, discrepant by a factor 2.7 at this N and asymptoting to a factor 3.

Table 1: Three-way triangulation of the corrected coefficient.

System	Method	Result
Sócrates / NumPy	Direct $\mathrm{Var}_p(\log d)$ at γ=1, N=10^4	$C_V = 7.89$;
Sage docker	Exact rational arithmetic for $N \in \{10, 50, 100\}$; ratio/4 asymptotes at 1/3	ratio/12 →
SymPy	Symbolic evaluation of the integral identity in §3	$\mathrm{Var}(\log d)$ =

5 Impact

Qualitative conclusion of §5.2 unchanged. The original paper concludes that $\gamma = 1$ is a *crossover*, not a true phase transition, on the grounds that C_V is finite for finite N and sharp in the $N \to \infty$ limit. The corrected coefficient $(\log N)^2/12$ is also finite for finite N and divergent in the limit, so this conclusion is preserved.

Empirical Hagedorn peak unchanged. The original paper reported an empirical broad peak in C_V at $\gamma \approx 1.235$ with value 6.86 for $N = 2000$. With the corrected coefficient, the asymptotic prediction at $\gamma = 1$ is $(\log 2000)^2/12 \approx 4.81$ (rather than 14.4 originally), which is *closer to* the empirical peak value 6.86 at the slightly off-peak location $\gamma = 1.235$. The corrected formula thus removes the apparent factor-3 discrepancy between theory and the paper's own measurement.

Other §5.2 results. The Fisher-information / heat-capacity identity $I(\gamma) = C_V(\gamma)/\gamma^2$ is unaffected (it is derived from the cumulant relation that we have just re-confirmed). The information-geometric distance integral in §5.3 is also unaffected at the structural level; numerical values computed using the old coefficient should be rescaled by a factor of 1/3 near the Hagedorn boundary.

KV compression formulas (§7). The compression theorems in §7 do not depend on the §5.2 coefficient and are unaffected.

TAF Agent diagnostic tool. The companion diagnostic tool [2] computes C_V at runtime via the numerical central-difference of $U(\gamma)$, not via the analytic formula (1). It therefore returns the correct asymptotic value automatically and does not require any software update for this erratum. Users who have been reading C_V values from the agent have been receiving the corrected quantity all along; users who computed $C_V(\gamma{=}1, N)$ analytically from (1) should multiply their result by 1/3.

Other parts of the paper. The Padé formula for $\gamma_{\mathrm{Padé}}(\theta, T)$ in §3.3, the MaxEnt power-law derivation in §4, the area-law theorem in §5.5, the critical-layer formula in §6, the KV-compression theorems in §7, and the attention-grammar / RLHF / fingerprinting sections are not affected.

6 How the error was caught

The Sócrates framework was applied to the paper as part of a recursive self-audit. Six claims of varying complexity were submitted to algebraic, numerical, and (where applicable) formal verification: the Padé formula, its Taylor coefficients, the MaxEnt Lagrangian derivation, the C_V Hagedorn theorem (this erratum), the KV-compression theorem, and the D-SAGE-1 quadratic identity. The Padé formula, Taylor coefficients, MaxEnt derivation, and D-SAGE-1 identity passed verification cleanly. The KV-compression theorem passed within $\leq 7.5\%$ relative error in finite-N truth (i.e. the paper's "exact continuous formula" is in fact a continuum-integral

approximation to the discrete sum, exact as an asymptotic but not pointwise; the practical implementation in the agent now uses direct discrete computation). The C_V Hagedorn theorem failed: the analytic prediction $(\log N)^2/4$ exceeds the numerically computed value by a factor of three, and the discrepancy converges, not vanishes, as N grows. Hand-derivation reproduced the corrected coefficient $(\log N)^2/12$, which was then verified by Sage docker (exact rational arithmetic) and SymPy (symbolic integration). Three independent systems agree.

7 Acknowledgements

This erratum was generated through the Sócrates Audit Framework, an in-progress methodology for self-falsifying scientific claims. The framework's algebraic, numerical, and formal-verification stages were applied iteratively to its own author's published work. The fact that a real algebraic error survived prior internal review and was caught only by the framework constitutes the framework's first reported real-world validation. The author acknowledges Anthropic's Claude as a research instrument throughout the audit, the SageMath community for the Sage exact-rational computation, the Lean / Mathlib4 community for the formal verification infrastructure that was set up but ultimately not required for this erratum, and the SymPy community for the symbolic-integration backend.

References

[1] C. Marín. Predicting How Transformers Attend: Analytic Power-Law Theory, Phase Transitions, and Practical Compression Tools. *Zenodo preprint*, 2026. https://doi.org/10.5281/zenodo.19826343.

[2] C. Marín. TAF Agent — browser diagnostic tool for transformer attention. *HuggingFace Space*, 2026. https://huggingface.co/spaces/karlexmarin/taf-agent.

[3] C. Marín. The Sócrates Audit Framework: a methodology for self-falsifying empirical and mathematical claims. *In preparation*, 2026.

[4] T. Tao. Almost all orbits of the Collatz map attain almost bounded values. *arXiv:1909.03562*, 2019.