BEYOND MULTIPLICITY CORRECTION: SHARP RETRIEVAL THRESHOLDS FOR ATTENTION WITH CORRELATED DISTRACTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Redundant context affects attention in two distinct ways: multiplicity adds normalization mass, while score variation within correlated groups can produce extreme individual scores. We separate these effects in a planted retrieval model with exponentially many Gaussian distractors organized into supplied clusters whose number and size both grow exponentially with dimension. The model is a two-level generalized random energy model, whose classical free energy yields exact retrieval thresholds. Gaussian interpolation further shows that increasing positive correlation at fixed token count lowers the distractor free energy, unlike adding duplicates.

Our main result establishes a sharp limitation of count correction, which subtracts the log cluster size from pre-softmax scores: this correction removes multiplicity but retains within-cluster extremes. Even when the inverse temperature is chosen after observing all scores, count correction requires a signal at least $\sqrt{2\alpha_1}$, where α_1 is the exponential growth rate of the number of clusters. By contrast, averaging scores before exponentiation suppresses token-level fluctuations and lowers the optimal signal threshold to $\sqrt{2\rho\alpha_1}$, where ρ is the within-cluster correlation.

This gap persists after adding exponentially many singleton distractors, so group size alone does not identify the target, and extends, under explicit cumulant regularity assumptions, to fixed finite hierarchies of independent-increment scores. Finite-sample, stability, and participation results complement the asymptotic analysis. A controlled retrieval experiment with frozen sentence embeddings illustrates both sides of the distinction: averaging helps when the relevant item is a singleton, but can discard useful evidence when a relevant group contains multiple matching items.

1 INTRODUCTION

Repeated passages, related keys, and duplicated records can change attention even when the quality of an individual query–key score stays fixed. Two interventions are easily conflated: correlating a fixed collection of scores, and adding more copies of existing scores. Their effects need not have the same sign. A theory that attributes every redundancy failure to positive correlation risks recommending the wrong intervention.

We analyze one softmax retrieval step with a planted relevant item and a hierarchical Gaussian distractor ensemble. Our objective is to characterize the signal strength needed for its attention mass to approach one. We treat the hierarchy as a supplied partition, as could arise from document or record identifiers. Inferring semantic groups from learned keys is a separate problem. We also include exponentially many singleton distractors so that group sizes alone need not identify the target. The model permits exact comparisons of ordinary attention, subtraction of log cluster size, and averaging scores before exponentiation.

Our main result identifies what count correction leaves behind: dividing a cluster's weight by its size does not suppress its largest individual scores. Count correction therefore has a sharp signal threshold that cannot be lowered even by choosing the temperature after observing all scores (Theorem 3). Averaging the scores first achieves a strictly smaller threshold (Theorem 5). The hidden-singleton

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

result removes identification from group size alone, and the deeper and non-Gaussian results extend this same comparison. Finite-sample and stability bounds describe its scope. The free energy and concentration transitions themselves come from classical random energy model (REM) and generalized random energy model (GREM) calculations (Derrida, 1980; 1985; Bovier & Kurkova, 2004); we provide self-contained proofs. Finally, a frozen-encoder experiment illustrates how the two operations differ on natural-sentence scores, including cases where averaging relevant scores hurts retrieval; it is not an empirical validation of the Gaussian asymptotic model.

Relation to prior work. The Gibbs interpretation of attention and its connection to associative memory are established (Ramsauer et al., 2021). REM methods already yield precise softmax and initialization analyses (Dohmatob, 2025; Giorlandino & Goldt, 2025). Long-context scaling depends on the score geometry (Chen et al., 2025), and gap-counting unifies several concentration scales (Hayase & Karakida, 2026). Training-driven transitions (Lavie et al., 2026), correlated Hopfield equilibria (Schimmenti & Ciarchi, 2026), and clustered self-attention dynamics (Gao et al., 2026) concern different observables or dynamics. Our observable is the mass of an externally planted item after a single retrieval step. This distinction does not make the underlying GREM calculation new.

Likewise, correcting sampling density is an established idea, including in PointConv (Wu et al., 2019). ToMe explicitly accounts for multiplicity through a $+\log m$ adjustment after merging (Bolya et al., 2023); Double-P uses size-weighted centroids to approximate token attention (Ni et al., 2026). We study a different objective: equal prior mass for supplied groups. LUCID addresses correlated-key interference through preconditioning (Duvvuri et al., 2026). We do not introduce a general density correction or token-merging architecture. Our focus is their sharp retrieval limits in an explicit probabilistic model.

2 Model and Normalization

We model retrieval against a collection of distractor tokens that arrive in correlated clusters. We begin with equally sized distractor clusters; Proposition 20 later treats unequal cluster sizes through a fixed finite collection of distractor types. The ambient dimension is d, and we study the high-dimensional limit $d \to \infty$ through integers. A distractor is indexed by a pair (a, j): the index a identifies its cluster and j identifies the token within that cluster. There are M_d clusters and m_d tokens in every cluster, where

$$M_d = \lceil e^{\alpha_1 d} \rceil, \quad m_d = \lceil e^{\alpha_2 d} \rceil, \qquad \alpha = \alpha_1 + \alpha_2. \tag{1}$$

Thus α_1 is the exponential growth rate of the number of distinct clusters, α_2 is the exponential growth rate of the multiplicity within each cluster, and the total number of distractors is $N_d = M_d m_d \asymp e^{\alpha d}$.

To represent similarity inside a cluster, assign to each cluster a shared Gaussian factor G_a and to each token a token-specific Gaussian factor H_{aj}. All of these variables are mutually independent $N(0,1)$ random variables, and the distractor score is

$$X_{aj} = \sqrt{\rho d}\, G_a + \sqrt{(1-\rho)d}\, H_{aj}, \tag{2}$$

where $0 < \rho < 1$. The term G_a shifts every score in cluster a together, whereas H_{aj} distinguishes individual tokens within that cluster. Hence each X_{aj} is marginally $N(0, d)$, two different tokens in the same cluster have covariance ρd (and correlation ρ), and tokens in different clusters are independent. The endpoints give useful intuition: $\rho = 0$ would make all distractor scores independent, while $\rho = 1$ would make the tokens within a cluster exact score duplicates. Equation (2) is therefore a two-level Gaussian random-energy model for repeated or near-duplicate distractors.

Alongside these distractors, let X_* denote the score of the single relevant, or planted, item. At inverse temperature $\beta > 0$, ordinary softmax attention assigns that item the mass

$$\mathcal{Z}_d(\beta) = \sum_{a=1}^{M_d} \sum_{j=1}^{m_d} e^{\beta X_{aj}}, \quad L_d(\beta) = d^{-1} \log \mathcal{Z}_d(\beta), \qquad A_* = \frac{e^{\beta X_*}}{e^{\beta X_*} + \mathcal{Z}_d(\beta)}.$$

Here $\mathcal{Z}_d$ is the total unnormalized attention weight of the distractors and L_d is its log weight per dimension. The exact normalized log-odds against the planted item are

$$\frac{1}{\beta d} \log \frac{1 - A_*}{A_*} = \frac{L_d(\beta)}{\beta} - \frac{X_*}{d}. \tag{3}$$

Thus a negative limiting right-hand side gives retrieval, while a positive limit gives failure. We assume only $X_*/d \xrightarrow{\mathbb{P}} s \in \mathbb{R}$: the constant s is the planted item's asymptotic score advantage on the order-d scale. We do not require X_* to be independent of the distractors. Retrieval means $A_* \xrightarrow{\mathbb{P}} 1$, and all probability statements refer to the score ensemble.

The score model can be realized as a query–key dot product $X_{aj} = q^\top k_{aj}$. Here $q \in \mathbb{R}^d$ is a deterministic query vector with $\|q\|^2 = d$, and $k_{aj} \in \mathbb{R}^d$ is the key of token j in cluster a. Let $z_a \sim N(0, I_d)$ be the random vector shared by cluster a, and let $\epsilon_{aj} \sim N(0, I_d)$ be the token-specific random vector; all vectors z_a and ϵ_{aj} are mutually independent. Setting $k_{aj} = \sqrt{\rho}\, z_a + \sqrt{1-\rho}\, \epsilon_{aj}$ gives the score distribution in (2). A planted key k_* can be chosen so that $q^\top k_* = sd$. Under conventional scaled dot-product attention, the pre-softmax score is $t_d q^\top k/\sqrt{d}$ for a dimension-dependent sharpening factor t_d, so our β equals $t_d/\sqrt{d}$. Thus fixed β corresponds to $t_d = \beta\sqrt{d}$, and cannot be identified with fixed conventional scaling. A learned query depending on these keys need not satisfy the assumed law.

Define the one-level free energy for variance vd,

$$f(\beta; a, v) = \begin{cases} a + \beta^2 v/2, & \beta \le \sqrt{2a/v}, \\ \beta\sqrt{2av}, & \beta \ge \sqrt{2a/v}, \end{cases} \qquad f(\beta; a, 0) = a. \tag{4}$$

We use inverse temperature: larger β means sharper attention.

3 Exact thresholds and the sign of correlation

Theorem 1 (Classical two-level free energy, specialized)**.** *For fixed $\beta > 0$, define*

$$F(\beta) = \begin{cases} f(\beta; \alpha, 1), & \rho \le \alpha_1/\alpha, \\ f(\beta; \alpha_1, \rho) + f(\beta; \alpha_2, 1-\rho), & \rho > \alpha_1/\alpha. \end{cases} \tag{5}$$

Then the normalized distractor log-partition function $L_d(\beta) = d^{-1} \log \mathcal{Z}_d(\beta)$ converges to $F(\beta)$ in probability and in L^1. Equivalently, with $I_1(x) = x^2/(2\rho)$ and $I_2(y) = y^2/[2(1-\rho)]$,

$$F(\beta) = \max_{\substack{I_1(x) \le \alpha_1 \\ I_1(x) + I_2(y) \le \alpha}} \{\alpha - I_1(x) - I_2(y) + \beta(x+y)\}. \tag{6}$$

Consequently, if $X_*/d \xrightarrow{\mathbb{P}} s \in \mathbb{R}$, then $A_* \xrightarrow{\mathbb{P}} 1$ when $s > F(\beta)/\beta$ and $A_* \xrightarrow{\mathbb{P}} 0$ when $s < F(\beta)/\beta$.

In (6), x is the order-d deviation of the cluster-level parent shared by all tokens in a cluster, while y is the token-specific child deviation. The first constraint requires that such a rare parent exist among the M_d clusters; the second is the total entropy constraint requiring that a parent–child path with deviations (x, y) exist among all $M_d m_d$ tokens. Replacing the second constraint with $I_2(y) \le \alpha_2$ for every parameter choice gives an incorrect answer in the merged regime.

To describe the freezing transitions, define the three inverse-temperature scales

$$b_1 = \sqrt{2\alpha_1/\rho}, \qquad b_2 = \sqrt{2\alpha_2/(1-\rho)}, \qquad b = \sqrt{2\alpha}.$$

If $\rho \le \alpha_1/\alpha$, the parent constraint imposes no additional restriction on the optimizer and the free energy matches that of an i.i.d. REM with the same marginal variance, although the scores remain correlated. If $\rho > \alpha_1/\alpha$, equivalently $b_1 < b_2$, parent and child freezing occur separately, and the three branches of F are

$$F(\beta) = \begin{cases} \alpha + \beta^2/2, & \beta \le b_1, \\ \beta\sqrt{2\rho\alpha_1} + \alpha_2 + \beta^2(1-\rho)/2, & b_1 \le \beta \le b_2, \\ \beta\big(\sqrt{2\rho\alpha_1} + \sqrt{2(1-\rho)\alpha_2}\big), & \beta \ge b_2. \end{cases} \tag{7}$$

The surface $\beta s = F(\beta)$ is deliberately excluded from the retrieval dichotomy. The first-order limit alone does not determine its behavior.

Finite-dimensional refinement. Theorem 1 gives the limiting threshold as $d \to \infty$. Proposition 8 gives a separate finite-d refinement: it bounds the deviation of L_d from its mean and combines this with the deviation of X_*/d from s to give an explicit retrieval margin. This proposition is not needed to prove the later threshold results; it quantifies the finite-size scope of the asymptotic statement.

Theorem 2 (Correlation comparison)**.** *At any finite d, with fixed M_d clusters and m_d tokens per cluster, and for $0 < \rho < 1$, define the distractor-only token weights $p_{aj} = e^{\beta X_{aj}}/\mathcal{Z}_d$ and cluster weights $P_a = \sum_j p_{aj}$. Then*

$$\frac{\partial}{\partial \rho}\mathbb{E}L_d(\beta) = -\frac{\beta^2}{2}\mathbb{E}\left[\sum_a P_a^2 - \sum_{a,j} p_{aj}^2\right] \le 0.$$

In particular $F(\beta;\rho) \le F(\beta;0) = f(\beta;\alpha,1)$. For a common limiting signal s, i.i.d. success $\beta s > f(\beta;\alpha,1)$ implies success for every $\rho \in [0,1]$.

The finite-d derivative compares models with the same number of tokens and the same marginal $N(0,d)$ distribution for every score; only the correlation between tokens in the same cluster changes. It says that the expected log distractor weight decreases as ρ increases. It does not say that $\mathcal{Z}_d$ decreases for every realized score array, and an ordering of the mean alone does not order finite-sample retrieval probabilities. The asymptotic threshold comparison additionally uses concentration and convergence to the deterministic free energy F. Adding a literal duplicate is a different operation: it adds another positive term to $\mathcal{Z}_d$ and therefore increases the softmax denominator.

What actually condenses? We call

$$P_{\mathrm{cl}} := \sum_a P_a^2, \qquad P_{\mathrm{tok}} := \sum_{a,j} p_{aj}^2$$

the *cluster participation* and *token participation*, respectively. Throughout the paper, participation refers to these squared-mass sums. Equivalently, they are the conditional probabilities that two independent draws from the same distractor attention distribution choose the same cluster or the same token. If the attention mass is uniform over K clusters or tokens, the corresponding participation equals $1/K$, so its reciprocal is the associated effective number of attended objects. Their expectations have the following limits (proof in Appendix B):

$$(\mathbb{E}P_{\mathrm{cl}}, \mathbb{E}P_{\mathrm{tok}}) \longrightarrow \begin{cases} ((1-b_1/\beta)_+, (1-b_2/\beta)_+), & b_1 < b_2, \\ ((1-b/\beta)_+, (1-b/\beta)_+), & b_1 \ge b_2. \end{cases} \tag{8}$$

Here $x_+ = \max(x,0)$. In the intermediate phase $b_1 < \beta < b_2$, cluster participation is nonvanishing but maximum token mass tends to zero. Neither maximum cluster mass nor maximum token mass tends to one at any fixed finite β: their respective participation expectations remain strictly below one. These are distractor phases, distinct from the external planted-item transition.

4 Count correction has a finite optimal temperature

Ordinary softmax assigns one unnormalized weight $e^{\beta X_{aj}}$ to every token, so a cluster can receive more total mass simply because it contains more tokens. Count correction removes this multiplicity advantage using the supplied partition: a token with score X in a group of size m receives corrected weight $e^{\beta X}/m$, equivalently subtracting $\log m$ from its pre-softmax score βX.

We first consider the balanced model in which every distractor cluster has size m_d and the planted item forms a singleton group. Each distractor weight is therefore divided by m_d, while the planted weight is unchanged. Define the count-corrected planted mass and its fixed-temperature signal threshold by

$$A_*^{\mathrm{cnt}} = \frac{e^{\beta X_*}}{e^{\beta X_*} + \mathcal{Z}_d(\beta)/m_d}, \qquad h_{\mathrm{cnt}}(\beta) = \frac{F(\beta) - \alpha_2}{\beta}. \tag{9}$$

Applying the log-odds identity (3) with the corrected distractor denominator $\mathcal{Z}_d/m_d$ gives

$$\frac{1}{\beta d}\log\frac{1 - A_*^{\mathrm{cnt}}}{A_*^{\mathrm{cnt}}} \xrightarrow{\mathbb{P}} h_{\mathrm{cnt}}(\beta) - s.$$

Thus count-corrected retrieval succeeds when $s > h_{\mathrm{cnt}}(\beta)$ and fails when $s < h_{\mathrm{cnt}}(\beta)$. By Theorem 1, the ordinary-attention threshold is $F(\beta)/\beta$; since $d^{-1}\log m_d \to \alpha_2$, count correction lowers it by exactly α_2/β.

The correction does not change relative weights within or between the equally sized distractor clusters. If every group, including the relevant group, had the same size, the same additive shift would be applied to every pre-softmax score and the softmax weights would not change. In the present basic model the planted item is the unique singleton, so group size alone reveals its location; Corollary 6 removes this shortcut by adding exponentially many singleton distractors.

Theorem 3 (Optimal correction and adaptive-temperature lower bound)**.** *For $\alpha_1, \alpha_2 > 0$ and $0 < \rho < 1$, the fixed-temperature retrieval threshold $h_{\mathrm{cnt}}(\beta)$ is minimized uniquely at $\beta_{\mathrm{opt}} = \sqrt{2\alpha_1}$:*

$$\min_{\beta>0} h_{\mathrm{cnt}}(\beta) = \sqrt{2\alpha_1}. \tag{10}$$

Thus, no fixed β retrieves when $s < \sqrt{2\alpha_1}$, whereas the fixed choice $\beta = \sqrt{2\alpha_1}$ retrieves when $s > \sqrt{2\alpha_1}$. Moreover, even if β is selected after observing the realized distractor scores,

$$\inf_{\beta>0} \frac{\log \mathcal{Z}_d(\beta) - \log m_d}{\beta d} \xrightarrow{\mathbb{P}} \sqrt{2\alpha_1}. \tag{11}$$

If $s < \sqrt{2\alpha_1}$, then $\sup_{\beta>0} A_*^{\mathrm{cnt}}(\beta) \xrightarrow{\mathbb{P}} 0$, even when β is chosen after observing all scores. If $s > \sqrt{2\alpha_1}$, the fixed choice in (10) gives $A_*^{\mathrm{cnt}} \xrightarrow{\mathbb{P}} 1$.

Proof idea; complete proof in Appendix C. Before the first freezing point, $h_{\mathrm{cnt}}(\beta) = \alpha_1/\beta + \beta/2$, whose derivative $-\alpha_1/\beta^2 + 1/2$ changes from negative to positive at $\beta = \sqrt{2\alpha_1}$. This value occurs before the first freezing point in both regimes of (5), and direct differentiation shows that every later formula for h_{cnt} is increasing. For the data-dependent choice of β, convergence is uniform when β stays in a fixed bounded interval. Jensen's inequality rules out $\beta \to 0$, while the maximum distractor score rules out $\beta \to \infty$; hence a minimizing sequence cannot escape to either endpoint. □

Increasing β makes the softmax place more weight on the largest scores. For count-corrected weights $e^{\beta X}/m$, the required signal $h_{\mathrm{cnt}}(\beta)$ decreases until $\beta = \sqrt{2\alpha_1}$ and then increases; choosing a larger β therefore requires a stronger planted signal rather than improving retrieval. With M_d independent $N(0, d)$ distractor scores, the smallest s permitting retrieval after optimizing β is also $\sqrt{2\alpha_1}$, although the required s at a particular fixed β generally differs between the two models.

If one could instead use only the M_d noise-free cluster components $\sqrt{\rho d}G_a$, the smallest required signal would be $\sqrt{2\rho\alpha_1}$. Count correction does not attain this smaller value because it divides out cluster size but retains the token-specific noise $\sqrt{(1-\rho)d}H_{aj}$; the averaging method below suppresses that noise.

Proposition 4 (Upper bound on finite-sample failure probability)**.** *If $X_* = sd$ deterministically, then for any $0 < \varepsilon < 1$,*

$$\mathbb{P}\{A_*^{\mathrm{cnt}} \le 1 - \varepsilon\} \le \min\left\{1, \frac{1-\varepsilon}{\varepsilon} M_d \exp\left[d(\beta^2/2 - \beta s)\right]\right\}. \tag{12}$$

For $s > 0$, the exponent is minimized at $\beta = s$, giving $M_d e^{-ds^2/2}(1-\varepsilon)/\varepsilon$. The bound needs only marginal $N(0, d)$ distractor scores and does not assume their independence.

Let $\alpha_{1,d} = d^{-1}\log M_d$. The upper bound in (12) decreases exponentially with d when

$$\alpha_{1,d} + \frac{\beta^2}{2} - \beta s < 0. \tag{13}$$

This is possible exactly when $s^2 > 2\alpha_{1,d}$, and then (13) is equivalent to

$$s - \sqrt{s^2 - 2\alpha_{1,d}} < \beta < s + \sqrt{s^2 - 2\alpha_{1,d}}.$$

Thus this upper bound is useful only when β is neither too small nor too large. A small β leaves substantial total mass on the many distractors, while a large β amplifies rare high distractor scores in the moment bound. The proposition gives no failure claim outside this interval; only this particular upper bound ceases to guarantee success there.

5 AVERAGING REMOVES TOKEN-SPECIFIC SCORE NOISE

Count correction replaces each cluster's total weight $\sum_j e^{\beta X_{aj}}$ by the average of its exponentiated token scores, $m_d^{-1}\sum_j e^{\beta X_{aj}}$. This removes the factor caused by cluster size, but a single unusually large token-specific component H_{aj} can still dominate the average. Averaging scores first performs a different operation: for each cluster a, compute

$$\bar{X}_a = \frac{1}{m_d}\sum_{j=1}^{m_d} X_{aj},$$

and then assign one attention weight $e^{\beta \bar{X}_a}$ to that cluster. The planted item remains a singleton with score X_*. Define

$$\mathcal{Z}_d^{\mathrm{avg}} = \sum_a e^{\beta \bar{X}_a}, \quad A_*^{\mathrm{avg}} = \frac{e^{\beta X_*}}{e^{\beta X_*} + \mathcal{Z}_d^{\mathrm{avg}}}, \quad h_{\mathrm{avg}}(\beta) = \frac{f(\beta;\alpha_1,\rho)}{\beta}.$$

Here $\mathcal{Z}_d^{\mathrm{avg}}$ is the total distractor weight after replacing every cluster by its mean score, A_*^{avg} is the resulting planted-item mass, and $h_{\mathrm{avg}}(\beta)$ will be the required planted signal at fixed β.

Under the Gaussian score model in (2), averaging gives

$$\bar{X}_a = \sqrt{\rho d}\, G_a + \sqrt{(1-\rho)d}\,\frac{1}{m_d}\sum_{j=1}^{m_d} H_{aj}.$$

The second term is Gaussian with variance $(1-\rho)d/m_d$, so it becomes negligible on the order-d scale because m_d grows exponentially. Thus the averaged cluster scores asymptotically behave like the M_d independent shared components $\sqrt{\rho d}G_a$. This is the Gaussian model used for the asymptotic formulas in the theorem below; the finite-score comparison in its first sentence does not require Gaussian scores.

Theorem 5 (Count correction versus averaging)**.** *For every fixed collection of token scores, $\mathcal{Z}_d^{\mathrm{avg}} \le \mathcal{Z}_d/m_d$ and $A_*^{\mathrm{avg}} \ge A_*^{\mathrm{cnt}}$. Under the Gaussian score model in (2),*

$$d^{-1}\log \mathcal{Z}_d^{\mathrm{avg}} \xrightarrow{\mathbb{P}} f(\beta;\alpha_1,\rho), \qquad \frac{1}{\beta d}\log\frac{1-A_*^{\mathrm{avg}}}{A_*^{\mathrm{avg}}} \xrightarrow{\mathbb{P}} h_{\mathrm{avg}}(\beta) - s.$$

Thus averaging retrieves when $s > h_{\mathrm{avg}}(\beta)$ and fails when $s < h_{\mathrm{avg}}(\beta)$. Optimizing over β gives

$$\inf_{\beta>0} h_{\mathrm{avg}}(\beta) = \sqrt{2\rho\alpha_1} < \sqrt{2\alpha_1} = \inf_{\beta>0} h_{\mathrm{cnt}}(\beta). \tag{14}$$

Consequently, if $\sqrt{2\rho\alpha_1} < s < \sqrt{2\alpha_1}$, averaging retrieves with any fixed $\beta \ge b_1$, while count correction fails even if its β is selected after observing all distractor scores.

The first two inequalities follow from Jensen's inequality. They mean that, for the same fixed token scores, averaging before exponentiation gives a distractor denominator no larger than count correction and therefore gives the planted item at least as much attention mass. The complete proof is in Appendix D.

The asymptotic improvement is stronger: averaging suppresses the token-specific noise that count correction leaves inside $m_d^{-1}\sum_j e^{\beta X_{aj}}$. The interval of signals where averaging succeeds and count correction fails has width $\sqrt{2\alpha_1}(1-\sqrt{\rho})$. This result assumes that the groups are supplied correctly and that the planted item is a singleton; it does not claim that arbitrary clustering or merging improves attention.

Hidden-singleton variant. In the basic model, the planted item is the only singleton group, so its group size reveals its identity before any scores are examined. To remove this shortcut, add $L_d = \lceil e^{a_0 d}\rceil$ independent $N(0,d)$ distractor scores, each in its own singleton group, where $0 < a_0 < \alpha_1$. Place the planted item uniformly at random among these $L_d + 1$ singleton groups and hide its location. Define the two critical signal levels

$$c_{\mathrm{avg},0} := \max\{\sqrt{2\rho\alpha_1}, \sqrt{2a_0}\} < c_{\mathrm{cnt}} := \sqrt{2\alpha_1}. \tag{15}$$

Corollary 6 (Hidden-singleton separation)**.** *In this augmented model, count correction fails for every data-dependent choice of β when $s < c_{\mathrm{cnt}}$, while the fixed choice $\beta = c_{\mathrm{cnt}}$ retrieves when $s > c_{\mathrm{cnt}}$. Averaging retrieves whenever $s > c_{\mathrm{avg},0}$ using any fixed $\beta \geq \max\{b_1, \sqrt{2a_0}\}$. Hence every $s \in (c_{\mathrm{avg},0}, c_{\mathrm{cnt}})$ is retrieved by averaging and missed by count correction for every choice of β. If $a_0 \leq \rho\alpha_1$, then $c_{\mathrm{avg},0} = \sqrt{2\rho\alpha_1}$ and the original separation interval is unchanged. A method that observes only the partition, and not the scores, identifies the target with probability at most $1/(L_d + 1)$.*

Proof idea; complete proof in Appendix D. The augmented distractor denominator is the sum of the original clustered contribution and the new singleton contribution. After taking $d^{-1} \log$, this sum has the larger of their two limiting values. Adding distractors can only decrease the planted mass, so the adaptive-temperature failure result for count correction remains valid. For averaging, the clustered representatives require signal $\sqrt{2\rho\alpha_1}$ and the singleton distractors require signal $\sqrt{2a_0}$; their maximum is $c_{\mathrm{avg},0}$. The partition-only bound follows because the hidden target is uniform among the $L_d + 1$ singleton groups. □

The hidden-singleton construction removes identification from group size alone, but it still assumes that the target is a singleton rather than a member of a mixed relevant–irrelevant group.

6 Practical meaning and scope

The theory assumes that the groups are supplied; it does not infer them from the keys. Subject to that assumption, the conclusions are stable to small errors, but the two corrections have different computational and semantic consequences.

Imperfect group statistics. Only relative count errors matter for count correction. If positive estimates satisfy

$$\max_{i \neq *} |\log(\widehat{m}_i/m_i) - \log(\widehat{m}_*/m_*)| \leq r_d, \tag{16}$$

then using the estimated counts changes $\log[(1 - A_*)/A_*]$, the log ratio of total distractor mass to planted mass, by at most r_d. In other words, if D is this log ratio under the true counts and $\widehat{D}$ is the same quantity under the estimated counts, then $|\widehat{D} - D| \leq r_d$. Likewise, if every averaged score has error at most u_d, the corresponding log ratio changes by at most $2\beta u_d$. Thus errors smaller than order d do not change the fixed-β conclusions away from the success–failure boundary. These bounds assume the groups are already available; they do not justify any particular clustering procedure. Proposition 13 gives the deterministic proof in Appendix D.

Cost and interpretation. Count correction still scores all N_d tokens, so its per-query cost remains $O(N_d d)$. Averaging can instead precompute M_d centroids in $O(N_d d)$ time and then score only those centroids in $O(M_d d)$ time per query. If all values have norm at most R, target mass A_* guarantees output error at most $2R(1 - A_*)$. This makes successful mass retrieval sufficient for value recovery, but not necessary; Proposition 14 gives the proof in Appendix D. It also explains an important limitation: averaging may erase the target when a supplied group mixes relevant and irrelevant items.

7 How broadly does the result apply?

The key result is a gap between the smallest planted signals required by the two methods: averaging succeeds above $\sqrt{2\rho\alpha_1}$, whereas count correction requires $\sqrt{2\alpha_1}$, as shown in (14).

A fixed finite score array. Before taking $d \to \infty$, consider one realized collection of distinct distractor scores $x_1, \ldots, x_{N_d}$ as fixed numbers. Their softmax weights at temperature β are $p_i(\beta) = e^{\beta x_i} / \sum_j e^{\beta x_j}$, and their entropy $S(p) = -\sum_i p_i \log p_i$ measures how broadly the distractor attention is spread. Proposition 15 shows that the best count-correction temperature is exactly the point where $S(p(\beta)) = \log m_d$, or equivalently where the effective number $e^{S(p)}$ of attended distractors is m_d. This optimization identity is known as entropic value-at-risk (EVaR); it is an existing result used here as an oracle benchmark, not a new contribution.

Deeper and non-Gaussian hierarchies. The two-level model has clusters and tokens, but the same comparison holds for any fixed number of Gaussian levels, such as source, document, passage, and token (Theorem 16). Gaussianity can also be relaxed when each level contributes centered, nondegenerate independent increments whose cumulant generating function is finite on $\mathbb{R}$ and has derivative tending to $\pm\infty$ at the two tails. For an increment Y_ℓ, the cumulant generating function is $\Lambda_\ell(t) = \log \mathbb{E}e^{tY_\ell}$; it records the exponential-tail behavior that determines the retrieval threshold. For a Gaussian increment it is simply a quadratic function of t, recovering the square-root formulas above. Theorem 18 proves the analogous count-versus-averaging gap for this non-Gaussian class. Full assumptions and proofs are in Appendices E–G; the results do not cover bounded-support saturation regimes, arbitrary learned score distributions, or unknown groups.

Boundary cases and heterogeneous groups. Appendix H records several extensions that delimit the basic comparison. Proposition 19 gives the optimal threshold when only a fraction $\gamma \log m_d$ of the multiplicity correction is subtracted. Proposition 20 allows a fixed finite collection of distractor types with unequal cluster sizes and also separates retrieval of a relevant group from retrieval of one member. Proposition 21 characterizes the inverse-count rule through exact replication invariance, while Corollary 22 shows that, in the two-transition case, no single temperature-independent effective context length can reproduce the full free-energy curve. The same appendix also records the endpoint and conventional-scaling checks used to delimit the asymptotic model.

8 Experiments and Limitations

The Gaussian simulations check the formulas within the stated model: finite-array temperature optimization approaches the predicted threshold, and cluster and token participation approach their respective limits. Separate constrained optimizations reproduce the free energy across 120 parameter settings and check the finite-array, deeper-hierarchy, and non-Gaussian extensions. Figure 1 and the full protocols are in Appendix I. These synthetic checks concern the theory itself. The following frozen-encoder experiment instead provides a controlled illustration of the two operations on learned sentence scores and of failure modes outside the theorem's assumptions; it is not a test of the Gaussian asymptotic law.

Frozen sentence retrieval. We use BANKING77, a collection of banking questions with 77 intent labels (Casanueva et al., 2020), and the frozen `all-MiniLM-L6-v2` sentence encoder (Sentence Transformers, 2021). Training sentences supply the keys; disjoint subsets of the official test split supply 308 validation and 924 evaluation queries. Scores are cosine similarities. Each context contains all 77 intents, initially with one relevant key and 38 singleton distractors. The other 38 intents contribute four or eight same-intent keys, or eight exact copies. Labels supply the groups; same-intent sentences need not be paraphrases. Three seeded context draws are shared by all methods. These are controlled contexts, not the standard intent-classification benchmark.

Ordinary attention and count correction weight individual keys. Averaging uses one mean score and one mean value per group, without renormalizing the mean key. Each value indicates its key's intent, so the relevant coordinate of the output is the total relevant value mass, equal to A_* when the target is a singleton. Each method independently selects a cosine-score multiplier from 81 values in $[0.25, 1024]$ by minimizing validation negative log relevant value mass. Evaluation queries are not used for this choice. Full settings, uncertainty estimates, and additional metrics are in Appendix J.

Multiplicity and score variation. Exact copies make count correction and averaging identical, as expected. With eight distinct keys per distractor group, their mean relevant masses are instead 0.342 and 0.440 after separate temperature selection (Table 1); the paired gain is 0.098 with 95% interval $[0.088, 0.108]$. Selecting one random representative per group gives 0.375, so retaining only one key does not explain the full gain. However, a label-aware oracle that chooses the best count-correction temperature separately for each evaluation query attains 0.511. Thus the learned-score experiment supports the distinction between the operations, but does not establish the theorem's all-temperature separation for these sentence scores.

When averaging loses relevant evidence. Giving the relevant intent four distinct keys reverses the comparison: count correction achieves 0.626 relevant mass and averaging 0.465. Variation

Table 1: Mean relevant value mass on 924 evaluation queries. Each method uses its own validation-selected temperature. The last column gives the paired averaging-minus-count difference with a 95% intent-block bootstrap interval. Groups are supplied using dataset labels.

Context	Ordinary	Count	Average	Difference [95% interval]
One key per intent	0.379	0.379	0.379	$+0.000$ [$+0.000$, $+0.000$]
Exact copies (8)	0.288	0.379	0.379	$+0.000$ [$+0.000$, $+0.000$]
Same intent (4)	0.267	0.345	0.406	$+0.062$ [$+0.056$, $+0.067$]
Same intent (8)	0.234	0.342	0.440	$+0.098$ [$+0.088$, $+0.108$]
Four relevant keys	0.614	0.626	0.465	-0.161 [-0.181, -0.142]
Mixed relevant group	0.234	0.243	0.005	-0.238 [-0.269, -0.208]

within a relevant group can contain useful matches to the query. Moving three distractor keys into the singleton's group is more damaging: averaging reduces relevant mass to 0.005, versus 0.243 for count correction. Here averaging also dilutes the relevant value to one quarter of the group's value. These cases are outside the singleton-target guarantee and show why group membership alone does not justify averaging.

What remains outside the theory. The model assumes exponentially many distractors, supplied groups, and a singleton target. Its exact phase diagram assumes Gaussian scores, while the non-Gaussian result still requires centered, nondegenerate independent increments with finite all-real exponential moments and steep cumulants. The temperature scaling also differs from conventional fixed scaled-dot-product attention. Our sentence experiment uses bounded cosine scores and label-supplied groups; it does not validate these asymptotic assumptions or a grouping algorithm. Appendix K reports a causal intervention on pre-softmax self-attention logits in a frozen DistilBERT model. It confirms that averaging suppresses within-group score extremes in actual attention, but its downstream effects are small, metric-dependent, and position-sensitive in a task on which the model performs near chance; it therefore does not support a general performance advantage.

Count correction removes multiplicity while retaining variation within a group. Averaging removes that variation as well. The theory shows why this yields a strict retrieval advantage against grouped distractors; the experiment shows why the same operation can remove useful evidence from a relevant group.

Reproducibility Statement

Every mathematical result stated in the main paper has a full proof in the appendix, with classical ingredients identified. The supplementary material contains the code and environment specifications for the direct optimizations, seeded Gaussian simulations, frozen-encoder experiment, and frozen-transformer diagnostic. Public model and data assets are revision-pinned and downloaded by the supplied scripts; fixed seeds and protocols regenerate the sentence splits, contexts, scores, calibration choices, and per-query or per-prompt outputs described in Appendices I–K. Generated results and model weights are not bundled. No model is trained. Numerical checks do not substitute for verification of the proofs.

AI use statement

In this work, the author used generative AI tools to assist with writing mathematical proofs, proposing and refining hypotheses, designing and evaluating the research methodology and numerical experiments, implementing the methods and verification code. The author also used generative AI tools to identify and check relevant literature and to improve the manuscript's clarity, structure, and formatting. The author takes full responsibility for the final content of the paper, including all text, claims, proofs, code, and artifacts produced with the assistance of generative AI.

REFERENCES

Amir Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155(3):1105–1123, 2012. doi: 10.1007/s10957-011-9968-2. URL https://link.springer.com/article/10.1007/s10957-011-9968-2.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JroZRaRw7Eu.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. doi: 10.1093/acprof:oso/9780199535255.001.0001.

Anton Bovier and Irina Kurkova. Derrida's generalised random energy models 1: Models with finitely many hierarchies. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 40 (4):439–480, 2004. doi: 10.1016/j.anihpb.2003.09.002. URL https://www.numdam.org/item/AIHPB_2004__40_4_439_0/.

Iñigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pp. 38–45, 2020. URL https://aclanthology.org/2020.nlp4convai-1.5/.

Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Critical attention scaling in long-context transformers, 2025. URL https://arxiv.org/abs/2510.05554.

Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Springer, 2 edition, 2010. doi: 10.1007/978-3-642-03311-7. URL https://link.springer.com/book/10.1007/978-3-642-03311-7. Corrected printing of the 1998 edition.

Bernard Derrida. Random-energy model: Limit of a family of disordered models. *Physical Review Letters*, 45(2):79–82, 1980. doi: 10.1103/PhysRevLett.45.79.

Bernard Derrida. A generalization of the random energy model which includes correlations between energies. *Journal de Physique Lettres*, 46(9):L401–L407, 1985. URL https://www.lps.ens.fr/~derrida/spinglass.html.

Elvis Dohmatob. Understanding softmax attention layers: Exact mean-field analysis on a toy problem. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.neurips.cc/paper_files/paper/2025/hash/2cf63608d264a3c65e68e3cff860c152-Abstract-Conference.html.

Sai Surya Duvvuri, Nirmal Patel, Nilesh Gupta, and Inderjit S. Dhillon. LUCID: Attention with preconditioned representations, 2026. URL https://arxiv.org/abs/2602.10410.

Qucheng Gao, Zuyi Yang, and Xiao Chen. Clustered attractor manifolds and dynamical condensation in self-attention, 2026. URL https://arxiv.org/abs/2608.08922.

Alessio Giorlandino and Sebastian Goldt. Two failure modes of deep transformers and how to avoid them: A unified theory of signal propagation at initialisation, 2025. URL https://arxiv.org/abs/2505.24333.

Tomohiro Hayase and Ryo Karakida. A unified framework for critical scaling of inverse temperature in self-attention, 2026. URL https://arxiv.org/abs/2605.12697.

N. K. Jana and B. V. Rao. Generalized random energy model II. *Journal of Statistical Physics*, 127: 841–850, 2007. doi: 10.1007/s10955-007-9288-y. URL https://link.springer.com/article/10.1007/s10955-007-9288-y.

Itay Lavie, Kirsten Fischer, Andrey Lekov, Frederic Van Maele, Zohar Ringel, and Moritz Helias. Phase transitions in attention: A bayesian theory of copy head emergence, 2026. URL https://arxiv.org/abs/2606.12058.

Wentao Ni, Kangqi Zhang, Zhongming Yu, Oren Nelson, Mingu Lee, Hong Cai, Fatih Porikli, Jongryool Kim, Zhijian Liu, and Jishen Zhao. Double-P: Hierarchical top-p sparse attention for long-context LLMs, 2026. URL https://arxiv.org/abs/2602.05191.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2008.02217.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter, 2019. URL https://arxiv.org/abs/1910.01108.

Vincenzo Maria Schimmenti and Matteo Ciarchi. A bifurcation theory for the equilibria of modern hopfield networks, 2026. URL https://arxiv.org/abs/2609.07757.

Sentence Transformers. all-MiniLM-L6-v2: Model card, 2021. URL https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Frozen revision and export hashes recorded with the experiments.

Wenxuan Wu, Zhongang Qi, and Fuxin Li. PointConv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Wu_PointConv_Deep_Convolutional_Networks_on_3D_Point_Clouds_CVPR_2019_paper.html.

A FREE ENERGY AND GAUSSIAN COMPARISON: COMPLETE PROOFS

Throughout this appendix, $d \to \infty$, $\alpha_1, \alpha_2 > 0$, and $\alpha = \alpha_1 + \alpha_2$. We use $M_d = \lceil e^{\alpha_1 d} \rceil$ clusters and $m_d = \lceil e^{\alpha_2 d} \rceil$ tokens per cluster, as in (1); hence $N_d = M_d m_d$ and $d^{-1} \log M_d \to \alpha_1$, $d^{-1} \log m_d \to \alpha_2$, and $d^{-1} \log N_d \to \alpha$. The ceiling functions do not affect any limit. Unless stated otherwise, inverse temperatures are fixed and positive, and $\xrightarrow{\mathbb{P}}$ denotes convergence in probability.

A.1 COUNTING STATES SUBJECT TO THE CONSTRAINTS

Lemma 7 (Two-level counting principle)**.** *Let $a = 1, \dots, M_d$ index clusters and $j = 1, \dots, m_d$ index tokens within each cluster. Let $U_a \sim N(0, v_1 d)$ and $V_{aj} \sim N(0, v_2 d)$ be mutually independent, where $v_1, v_2 > 0$. For $X_{aj} = U_a + V_{aj}$, the limiting free energy is*

$$\max_{\substack{x^2/(2v_1) \le \alpha_1 \\ x^2/(2v_1) + y^2/(2v_2) \le \alpha}} \left\{ \alpha - \frac{x^2}{2v_1} - \frac{y^2}{2v_2} + \beta(x+y) \right\}. \tag{17}$$

Convergence holds in probability and in L^1.

Proof. Write $I_1(x) = x^2/(2v_1)$ and $I_2(y) = y^2/(2v_2)$. For a fixed interval B of positive length,

$$\frac{1}{d} \log \mathbb{P}(U_a/d \in B) \xrightarrow{d \to \infty} - \inf_{x \in B} I_1(x). \tag{18}$$

Since $U_a/d \sim N(0, v_1/d)$, its standard deviation is $\sqrt{v_1/d}$ and its density is $\sqrt{d/(2\pi v_1)} \exp[-dx^2/(2v_1)]$. Integrating this density gives the displayed limit; using interior subintervals establishes the lower bound. The identical statement holds for V_{aj}/d.

Upper bound. Choose R large enough that $R^2/(2v_1) > \alpha_1 + 1$ and $R^2/(2v_2) > \alpha + 1$. A union bound makes the probability that any U_a/d or V_{aj}/d leaves $[-R, R]$ tend to zero exponentially. Partition this compact interval into finitely many bins of width at most δ. For a pair of bins B, C, put $r_1 = \inf_B I_1$ and $r_2 = \inf_C I_2$, and define the finite index counts

$$K_B := \#\{a : U_a/d \in B\}, \qquad T_{BC} := \#\{(a,j) : U_a/d \in B,\ V_{aj}/d \in C\}.$$

Thus K_B counts clusters with $U_a/d \in B$, while T_{BC} counts cluster-token index pairs satisfying both bin conditions; it does not count the continuum of points in $B \times C$. By (18),

$$\mathbb{E}K_B = e^{d(\alpha_1 - r_1 + o(1))}, \qquad \mathbb{E}T_{BC} = e^{d(\alpha - r_1 - r_2 + o(1))}.$$

Fix $\eta > 0$. Markov's inequality, followed by a union bound over the finite bins, shows with probability tending to one that bins with $r_1 > \alpha_1 + \eta$ or $r_1 + r_2 > \alpha + \eta$ are empty, and all other bins satisfy

$$T_{BC} \le \exp\{d[\max(\alpha - r_1 - r_2, 0) + \eta]\}.$$

We next convert the count bound into a bound on the partition function $\mathcal{Z}_d = \sum_{a,j} e^{\beta X_{aj}}$. For each bin pair define its contribution

$$\mathcal{Z}_{BC} := \sum_{\substack{(a,j):\, U_a/d \in B \\ V_{aj}/d \in C}} e^{\beta X_{aj}}.$$

There are T_{BC} summands in $\mathcal{Z}_{BC}$. For every one of them, $X_{aj}/d = U_a/d + V_{aj}/d$ is at most $\sup_{(x,y) \in B \times C}(x+y)$. Hence

$$\mathcal{Z}_{BC} \le T_{BC} \exp\left\{ \beta d \sup_{(x,y) \in B \times C} (x+y) \right\}. \tag{19}$$

The bounds above control T_{BC}, and substituting them into (19) bounds each bin pair's contribution. Since the bin pairs partition all index pairs (a,j) on the event under consideration, $\mathcal{Z}_d = \sum_{B,C} \mathcal{Z}_{BC}$. For fixed δ, this is a sum over finitely many bin pairs, independently of d. Thus replacing the largest bin-pair bound by their sum changes $L_d = d^{-1} \log \mathcal{Z}_d$ by at most $(\log \text{number of bin pairs})/d = o(1)$.

First let $d \to \infty$, and then send $\eta \downarrow 0$. The relaxed constraints used above,

$$r_1 \le \alpha_1 + \eta, \qquad r_1 + r_2 \le \alpha + \eta,$$

therefore reduce to the stated constraints $r_1 \le \alpha_1$ and $r_1 + r_2 \le \alpha$. On this limiting set, $\alpha - r_1 - r_2 \ge 0$, so $\max(\alpha - r_1 - r_2, 0) = \alpha - r_1 - r_2$. All these sets lie in the compact square $[-R, R]^2$, and the objective is continuous, so their maxima converge to the maximum in (17). Finally send $\delta \downarrow 0$.

Lower bound. Fix (x, y) satisfying both constraints strictly. Choose small intervals B, C around these coordinates, and recall that $r_1 = \inf_{u \in B} I_1(u)$ and $r_2 = \inf_{v \in C} I_2(v)$. By choosing the intervals sufficiently small, the strict feasibility of (x, y) gives $r_1 < \alpha_1$ and $r_1 + r_2 < \alpha$.

Write $K_B = \sum_{a=1}^{M_d} \mathbf{1}\{U_a/d \in B\}$. It is binomial with mean $\mathbb{E}K_B = M_d \mathbb{P}(U_a/d \in B) = \exp\{d(\alpha_1 - r_1 + o(1))\}$, which grows exponentially because $\alpha_1 - r_1 > 0$. Its relative error $|K_B - \mathbb{E}K_B|/\mathbb{E}K_B$ tends to zero since $\mathrm{Var}(K_B)/(\mathbb{E}K_B)^2 \le 1/\mathbb{E}K_B \to 0$, Chebyshev's inequality gives $K_B/\mathbb{E}K_B \xrightarrow{\mathbb{P}} 1$; in particular $d^{-1} \log K_B \xrightarrow{\mathbb{P}} \alpha_1 - r_1$. Conditional on the cluster-level variables $(U_a)_a$, the $K_B m_d$ indicators for $V_{aj}/d \in C$ are independent Bernoulli variables. Their conditional total mean has exponential rate $\alpha - r_1 - r_2 > 0$. A binomial Chernoff bound therefore gives

$$\frac{1}{d} \log T_{BC} \xrightarrow{\mathbb{P}} \alpha - r_1 - r_2.$$

These second-level trials remain independent even when several tokens belong to the same selected cluster, because selection used only the U_a. No condition $r_2 < \alpha_2$ is required. Every index pair counted by T_{BC} satisfies $X_{aj}/d = U_a/d + V_{aj}/d \ge \inf_{(u,v) \in B \times C}(u + v)$. Its exponential weight is therefore at least $\exp\{\beta d \inf_{B \times C}(u + v)\}$, and hence

$$\mathcal{Z}_d \ge T_{BC} \exp\left\{ \beta d \inf_{(u,v) \in B \times C} (u + v) \right\}.$$

Combining this inequality with $d^{-1} \log T_{BC} \xrightarrow{\mathbb{P}} \alpha - r_1 - r_2$ gives, with probability tending to one,

$$\liminf_{d \to \infty} L_d \ge \alpha - r_1 - r_2 + \beta \inf_{(u,v) \in B \times C} (u + v).$$

Now shrink B and C to the chosen point (x, y). Continuity of I_1, I_2, and $u + v$ makes the right-hand side tend to $\alpha - I_1(x) - I_2(y) + \beta(x + y)$, proving the lower bound at every strictly feasible (x, y).

It remains to include points on the boundary of the feasible set. If (x, y) satisfies the non-strict constraints, then for every $0 < t < 1$ the point (tx, ty) satisfies them strictly: its two costs are $t^2 I_1(x) < \alpha_1$ and $t^2[I_1(x) + I_2(y)] < \alpha$, because $\alpha_1, \alpha > 0$. Apply the strict-point lower bound to (tx, ty) and let $t \uparrow 1$. The objective is continuous, so its value converges to the value at (x, y). Since the feasible set is compact and the objective is continuous, it has a maximizer; applying this argument to that maximizer proves the lower bound for the maximum.

L^1 convergence. We have already proved $L_d \xrightarrow{\mathbb{P}} F$. To strengthen this to L^1 convergence, we must show

$$\mathbb{E}|L_d - F| \longrightarrow 0. \tag{20}$$

It is enough to show that $(L_d)_d$ is uniformly integrable, meaning that large values make a uniformly negligible contribution to its expectation:

$$\lim_{K \to \infty} \sup_d \mathbb{E}\big[|L_d| \mathbf{1}\{|L_d| > K\}\big] = 0. \tag{21}$$

The elementary inequality

$$|L_d| \le d^{-1} \log N_d + \beta d^{-1} \max_{a,j} |X_{aj}|$$

reduces this to controlling the largest score. Each X_{aj} is $N(0, (v_1 + v_2)d)$, so a union bound gives

$$\mathbb{P}\left(\frac{\max_{a,j} |X_{aj}|}{d} > u \right) \le 2 N_d \exp\left[-\frac{d u^2}{2(v_1 + v_2)} \right].$$

Because $d^{-1}\log N_d \to \alpha$, the right-hand side decays exponentially in d for every sufficiently large fixed u. Integrating this tail bound shows

$$\sup_d \mathbb{E}\left[\left(d^{-1}\max_{a,j}|X_{aj}|\right)^2\right] < \infty.$$

Since $d^{-1}\log N_d$ is also uniformly bounded, the displayed bound on L_d gives $\sup_d \mathbb{E}L_d^2 < \infty$. This second-moment bound implies (21), because

$$\mathbb{E}\big[|L_d|\mathbf{1}\{|L_d| > K\}\big] \le K^{-1}\mathbb{E}L_d^2.$$

Finally let $Y_d = L_d - F$. The same uniform-integrability property holds for $(Y_d)_d$. For any $\varepsilon > 0$ and $K > \varepsilon$,

$$\mathbb{E}|Y_d| \le \varepsilon + K\mathbb{P}(|Y_d| > \varepsilon) + \mathbb{E}\big[|Y_d|\mathbf{1}\{|Y_d| > K\}\big].$$

First choose K so that the last term is uniformly small, then let $d \to \infty$ so that the middle term vanishes by $Y_d \xrightarrow{\mathbb{P}} 0$, and finally send $\varepsilon \downarrow 0$. This proves (20). □

A.2 Evaluation of the variational formula

Proof of Theorem 1. Apply Lemma 7 with $v_1 = \rho, v_2 = 1-\rho$. It remains to solve its strictly concave optimization problem.

The two constraints depend on x^2 and y^2, whereas $\beta(x+y)$ increases when either negative coordinate is replaced by its absolute value. Thus a maximizer can be chosen with $x, y \ge 0$. If $\rho \le \alpha_1/\alpha$, temporarily ignore the constraint $x^2/(2\rho) \le \alpha_1$. Define

$$g(x,y) := \alpha - \frac{x^2}{2\rho} - \frac{y^2}{2(1-\rho)} + \beta(x+y), \qquad \frac{x^2}{2\rho} + \frac{y^2}{2(1-\rho)} \le \alpha.$$

For every (x,y) satisfying the remaining constraint, Cauchy–Schwarz gives

$$x + y \le \left(\frac{x^2}{\rho} + \frac{y^2}{1-\rho}\right)^{1/2},$$

with equality exactly when $x/\rho = y/(1-\rho)$. Thus, for each fixed value of $x^2/(2\rho) + y^2/[2(1-\rho)]$, the largest possible value of g occurs at $x = \rho t$, $y = (1-\rho)t$. The remaining constraint becomes $0 \le t \le \sqrt{2\alpha}$, and on this line

$$g(\rho t, (1-\rho)t) = \alpha - \frac{t^2}{2} + \beta t.$$

The derivative of this one-variable function is $\beta - t$, so its maximum on $[0, \sqrt{2\alpha}]$ occurs at $t = \min(\beta, \sqrt{2\alpha})$. Hence

$$(x,y) = (\rho t, (1-\rho)t), \qquad t = \min(\beta, \sqrt{2\alpha}).$$

For this point, $x^2/(2\rho) = \rho t^2/2 \le \rho\alpha \le \alpha_1$, so it also satisfies the constraint that was temporarily omitted. Substitution gives $f(\beta;\alpha,1)$.

Put $b_1 = \sqrt{2\alpha_1/\rho}$ and $b_2 = \sqrt{2\alpha_2/(1-\rho)}$. If $\rho > \alpha_1/\alpha$, then $b_1 < b_2$. For $\beta \le b_1$, the point $(x,y) = (\beta\rho, \beta(1-\rho))$ satisfies

$$\frac{x^2}{2\rho} = \frac{\beta^2\rho}{2} \le \alpha_1, \qquad \frac{x^2}{2\rho} + \frac{y^2}{2(1-\rho)} = \frac{\beta^2}{2} \le \frac{\alpha_1}{\rho} < \alpha,$$

and the two equations $\partial g/\partial x = \partial g/\partial y = 0$ show that it is the maximizer. For $b_1 \le \beta \le b_2$, take

$$x = \sqrt{2\rho\alpha_1}, \qquad y = \beta(1-\rho).$$

Here $x^2/(2\rho) = \alpha_1$ and

$$\frac{x^2}{2\rho} + \frac{y^2}{2(1-\rho)} = \alpha_1 + \frac{\beta^2(1-\rho)}{2} \le \alpha_1 + \alpha_2 = \alpha.$$

With $\lambda_1 = \beta/b_1 - 1 \geq 0$ and $\lambda_2 = 0$, the derivative equations for the two constraints are

$$\beta - \frac{(1+\lambda_1+\lambda_2)x}{\rho} = 0, \qquad \beta - \frac{(1+\lambda_2)y}{1-\rho} = 0. \tag{22}$$

For $\beta \geq b_2$, choose

$$x = \sqrt{2\rho\alpha_1}, \qquad y = \sqrt{2(1-\rho)\alpha_2}.$$

This point satisfies $x^2/(2\rho) = \alpha_1$ and $x^2/(2\rho) + y^2/[2(1-\rho)] = \alpha$. The choices $\lambda_2 = \beta/b_2 - 1$ and $\lambda_1 = \beta/b_1 - \beta/b_2$ are nonnegative and satisfy (22). Because g is concave and both constraint functions are convex, these equations together with the displayed inequalities identify the global maximizer. Substitution gives (7), with the formulas agreeing at all boundaries.

Finally, write $D_d = \log \mathcal{Z}_d - \beta X_*$. Then $D_d/d \xrightarrow{\mathbb{P}} F - \beta s$ and $A_* = (1 + e^{D_d})^{-1}$. A strictly negative or positive limiting exponent proves the two retrieval conclusions without any independence requirement on X_*. □

A.3 Finite-size concentration and retrieval margins

The free-energy theorem gives a limiting statement. The following proposition is a separate finite-d refinement.

Proposition 8 (Gaussian concentration)**.** *For every $u > 0$,*

$$\mathbb{P}\{|L_d - \mathbb{E}L_d| \geq u\} \leq 2\exp[-du^2/(2\beta^2)]. \tag{23}$$

Let $e_d = |\mathbb{E}L_d - F| \to 0$ and $t_d(r) = \mathbb{P}(|X_*/d - s| > r)$. Except on an event of probability at most $2e^{-du^2/(2\beta^2)} + t_d(r)$,

$$\left| d^{-1} \log \frac{1 - A_*}{A_*} - (F - \beta s) \right| \leq e_d + u + \beta r. \tag{24}$$

Thus a positive margin $\beta s - F > e_d + u + \beta r$ implies $1 - A_* \leq \exp[-d(\beta s - F - e_d - u - \beta r)]$ on this event.

Proof. We use the standard Gaussian Lipschitz concentration inequality, see Boucheron et al. (2013): if W is a vector of independent $N(0,1)$ variables and $f(W)$ is c-Lipschitz in Euclidean distance, then

$$\mathbb{P}\{|f(W) - \mathbb{E}f(W)| \geq u\} \leq 2\exp\left[-\frac{u^2}{2c^2}\right].$$

Thus, to prove (23), it suffices to show that L_d is $(\beta/\sqrt{d})$-Lipschitz as a function of the independent standard Gaussian variables (G, H). The derivatives of $\log \mathcal{Z}_d$ are

$$\partial_{G_a} \log \mathcal{Z}_d = \beta\sqrt{\rho d}\, P_a, \qquad \partial_{H_{aj}} \log \mathcal{Z}_d = \beta\sqrt{(1-\rho)d}\, p_{aj}.$$

Therefore

$$\sum_a (\partial_{G_a} \log \mathcal{Z}_d)^2 + \sum_{a,j} (\partial_{H_{aj}} \log \mathcal{Z}_d)^2 = \beta^2 d \left[\rho \sum_a P_a^2 + (1-\rho) \sum_{a,j} p_{aj}^2\right] \leq \beta^2 d.$$

Hence, we obtain $\|\nabla \log \mathcal{Z}_d\|_2 \leq \beta\sqrt{d}$. Since $L_d = d^{-1} \log \mathcal{Z}_d$, its gradient has Euclidean norm at most $\beta/\sqrt{d}$, so L_d is $(\beta/\sqrt{d})$-Lipschitz.

On the concentration event, $|L_d - F| \leq |L_d - \mathbb{E}L_d| + |\mathbb{E}L_d - F| \leq u + e_d$. Intersect it with the event $|X_*/d - s| \leq r$ and use $d^{-1}\log[(1-A_*)/A_*] = L_d - \beta X_*/d$ to obtain (24). If $\Delta_d := \beta s - F - e_d - u - \beta r > 0$, then (24) gives

$$\frac{1}{d}\log\frac{1-A_*}{A_*} \leq -\Delta_d, \qquad 1 - A_* = A_*\frac{1-A_*}{A_*} \leq e^{-d\Delta_d},$$

which proves the final assertion of the proposition. No explicit rate is asserted for the deterministic bias e_d; this statement must not be read as a rate to the limiting threshold without that bias term. □

A.4 GAUSSIAN INTERPOLATION

Lemma 9 (Covariance differentiation)**.** *For a differentiable positive-definite covariance path $C(t)$ and a centered Gaussian vector $Y(t) \sim N(0, C(t))$, the log-sum-exp $g(y) = \log \sum_i e^{\beta y_i}$ satisfies*

$$\frac{d}{dt}\mathbb{E}g(Y(t)) = \frac{1}{2}\sum_{i,j} C'_{ij}(t)\mathbb{E}\partial_i\partial_j g(Y(t)). \tag{25}$$

Proof. Let ϕ_t be the density of $N(0, C(t))$. Its Fourier transform is

$$\widehat{\phi}_t(\xi) = \exp\left[-\frac{1}{2}\xi^\top C(t)\xi\right].$$

Differentiating in t gives

$$\frac{d}{dt}\widehat{\phi}_t(\xi) = -\frac{1}{2}\sum_{i,j} C'_{ij}(t)\xi_i\xi_j\widehat{\phi}_t(\xi) = \widehat{\left[\frac{1}{2}\sum_{i,j} C'_{ij}(t)\partial_i\partial_j\phi_t\right]}(\xi).$$

Therefore

$$\partial_t\phi_t(y) = \frac{1}{2}\sum_{i,j} C'_{ij}(t)\partial_i\partial_j\phi_t(y). \tag{26}$$

Writing $p_i(y) = e^{\beta y_i}/\sum_k e^{\beta y_k}$, the first two derivatives of g are

$$\partial_i g(y) = \beta p_i(y), \qquad \partial_i\partial_j g(y) = \beta^2\big[p_i(y)\mathbf{1}_{i=j} - p_i(y)p_j(y)\big]. \tag{27}$$

In particular, the second derivatives are bounded. Hence differentiating under the integral and using (26) yields

$$\begin{aligned}\frac{d}{dt}\mathbb{E}g(Y(t)) &= \int g(y)\,\partial_t\phi_t(y)\,dy \\ &= \frac{1}{2}\sum_{i,j} C'_{ij}(t)\int g(y)\,\partial_i\partial_j\phi_t(y)\,dy \\ &= \frac{1}{2}\sum_{i,j} C'_{ij}(t)\int \partial_i\partial_j g(y)\,\phi_t(y)\,dy,\end{aligned}$$

where the last equality is integration by parts twice. Since $g(y) \le \log n + \beta\max_i|y_i|$ and Gaussian densities and their derivatives have Gaussian tails, the differentiations and the boundary terms are justified. This is (25). Singular endpoints are obtained by continuity; only positive-definite $C(t)$ is needed in the displayed calculation. □

Proof of Theorem 2. Index the Gaussian score vector by $i = (a, j)$. Its covariance matrix satisfies

$$\frac{d}{d\rho}\operatorname{Cov}(X_{aj}, X_{bk}) = \begin{cases} d, & a = b,\ j \ne k, \\ 0, & \text{otherwise.}\end{cases}$$

Indeed, every score has variance d, so the derivative is 0 when $(a, j) = (b, k)$; two distinct scores have covariance ρd exactly when they belong to the same cluster. For $g(x) = \log\sum_{a,j} e^{\beta x_{aj}}$, (27) gives, for $j \ne k$,

$$\partial_{aj}\partial_{ak} g = -\beta^2 p_{aj}p_{ak}.$$

Applying (25) and dividing by d gives

$$\frac{d}{d\rho}\mathbb{E}L_d = \frac{1}{2d}\sum_a\sum_{j\ne k} d\,\mathbb{E}[-\beta^2 p_{aj}p_{ak}] = -\frac{\beta^2}{2}\mathbb{E}\sum_a\sum_{j\ne k} p_{aj}p_{ak}.$$

For each cluster,

$$\sum_{j\ne k} p_{aj}p_{ak} = \left(\sum_j p_{aj}\right)^2 - \sum_j p_{aj}^2.$$

Let $P_{\mathrm{cl}} = \sum_a P_a^2$ and $P_{\mathrm{tok}} = \sum_{a,j} p_{aj}^2$. Summing over a gives $\sum_a \sum_{j \neq k} p_{aj} p_{ak} = P_{\mathrm{cl}} - P_{\mathrm{tok}}$, and therefore

$$\frac{d}{d\rho} \mathbb{E} L_d = -\frac{\beta^2}{2} \mathbb{E}(P_{\mathrm{cl}} - P_{\mathrm{tok}}) \leq 0,$$

because $P_{\mathrm{cl}} - P_{\mathrm{tok}} = \sum_a \sum_{j \neq k} p_{aj} p_{ak} \geq 0$. Thus, for $0 < \rho_1 < \rho_2 < 1$,

$$\mathbb{E} L_d(\rho_2) - \mathbb{E} L_d(\rho_1) = -\frac{\beta^2}{2} \int_{\rho_1}^{\rho_2} \mathbb{E}(P_{\mathrm{cl}} - P_{\mathrm{tok}}) \, d\rho \leq 0.$$

Continuity extends this inequality to $\rho_1 = 0$ and $\rho_2 = 1$. Finally, L^1 convergence gives $\mathbb{E} L_d(\rho) \to F(\beta; \rho)$ for every fixed ρ, so $F(\beta; \rho) \leq F(\beta; 0) = f(\beta; \alpha, 1)$. If $\beta s > f(\beta; \alpha, 1)$, then $\beta s > F(\beta; \rho)$, and Theorem 1 gives $A_* \xrightarrow{\mathbb{P}} 1$. □

B PARTICIPATION WITHOUT AN INCORRECT WINNER-TAKE-ALL CLAIM

Proposition 10 (Expected participation)**.** *Equation (8) holds for every fixed $\beta > 0$ and $0 < \rho < 1$. A strictly positive limiting participation expectation does not assert a deterministic limiting participation value.*

Proof. Temporarily replace ρ and $1 - \rho$ by independent positive variances v_1, v_2, and write

$$X_{aj}^{(v)} = \sqrt{v_1 d} \, G_a + \sqrt{v_2 d} \, H_{aj}, \qquad L_d(v_1, v_2) = \frac{1}{d} \log \sum_{a,j} e^{\beta X_{aj}^{(v)}}.$$

The covariance is

$$\mathrm{Cov}(X_{aj}^{(v)}, X_{bk}^{(v)}) = d \left(v_1 \mathbf{1}_{a=b} + v_2 \mathbf{1}_{(a,j)=(b,k)} \right).$$

Consequently,

$$\partial_{v_1} \mathrm{Cov}(X_{aj}^{(v)}, X_{bk}^{(v)}) = d \mathbf{1}_{a=b}, \qquad \partial_{v_2} \mathrm{Cov}(X_{aj}^{(v)}, X_{bk}^{(v)}) = d \mathbf{1}_{(a,j)=(b,k)}. \tag{28}$$

Let

$$g(x) = \log \sum_{a,j} e^{\beta x_{aj}}.$$

By (27),

$$\partial_{aj} \partial_{bk} g(x) = \beta^2 \left(p_{aj} \mathbf{1}_{(a,j)=(b,k)} - p_{aj} p_{bk} \right),$$

where ∂_{aj} means differentiation with respect to the score coordinate x_{aj}. Applying Lemma 9, using $L_d = d^{-1} g$, and substituting the first identity in (28) gives

$$\begin{aligned} \partial_{v_1} \mathbb{E} L_d &= \frac{1}{2d} \sum_{a,j} \sum_{b,k} d \mathbf{1}_{a=b} \, \mathbb{E}[\partial_{aj} \partial_{bk} g] = \frac{\beta^2}{2} \mathbb{E} \left[\sum_{a,j} p_{aj} - \sum_a \sum_{j,k} p_{aj} p_{ak} \right] \\ &= \frac{\beta^2}{2} \mathbb{E} \left[1 - \sum_a \left(\sum_j p_{aj} \right)^2 \right] = \frac{\beta^2}{2} (1 - \mathbb{E} P_{\mathrm{cl}}). \end{aligned} \tag{29}$$

Likewise, the second identity in (28) gives

$$\begin{aligned} \partial_{v_2} \mathbb{E} L_d &= \frac{1}{2d} \sum_{a,j} \sum_{b,k} d \mathbf{1}_{(a,j)=(b,k)} \, \mathbb{E}[\partial_{aj} \partial_{bk} g] = \frac{\beta^2}{2} \mathbb{E} \left[\sum_{a,j} p_{aj} - \sum_{a,j} p_{aj}^2 \right] \\ &= \frac{\beta^2}{2} (1 - \mathbb{E} P_{\mathrm{tok}}). \end{aligned} \tag{30}$$

Equations (29) and (30) are exact for every finite d.

It remains to justify passing these derivatives to the limit $d \to \infty$. Parameterize

$$v_i = t_i^2, \qquad t_i > 0,$$

and define

$$\Phi_d(t_1, t_2) := \mathbb{E}\left[\frac{1}{d}\log\sum_{a,j}\exp\{\beta\sqrt{d}(t_1 G_a + t_2 H_{aj})\}\right].$$

For every realized Gaussian array, the expression inside the expectation is convex, and hence Φ_d is convex. Lemma 7 and its L^1 conclusion give pointwise convergence

$$\Phi_d(t_1, t_2) \longrightarrow \Phi(t_1, t_2) := F(\beta; t_1^2, t_2^2).$$

Fix one coordinate, hold the other coordinate constant, and write the resulting one-variable functions as $\Phi_d(t)$ and $\Phi(t)$. For every $h > 0$ small enough, convexity gives

$$\frac{\Phi_d(t) - \Phi_d(t-h)}{h} \le \Phi_d'(t) \le \frac{\Phi_d(t+h) - \Phi_d(t)}{h}.$$

Taking $d \to \infty$ in the two secant slopes yields

$$\frac{\Phi(t) - \Phi(t-h)}{h} \le \liminf_{d\to\infty} \Phi_d'(t) \le \limsup_{d\to\infty} \Phi_d'(t) \le \frac{\Phi(t+h) - \Phi(t)}{h}.$$

If Φ is differentiable at t, sending $h \downarrow 0$ makes both outer terms converge to $\Phi'(t)$. Therefore

$$\Phi_d'(t) \longrightarrow \Phi'(t). \tag{31}$$

Applying this argument separately to t_1 and t_2, and using

$$\partial_{v_i} = \frac{1}{2t_i}\partial_{t_i}, \qquad v_i = t_i^2,$$

shows that the derivatives in (29) and (30) converge to the corresponding derivatives of the limiting free energy.

The general-variance limiting free energy is

$$F(\beta; v_1, v_2) = \begin{cases} f(\beta; \alpha_1, v_1) + f(\beta; \alpha_2, v_2), & \alpha_1/v_1 < \alpha_2/v_2, \\ f(\beta; \alpha, v_1 + v_2), & \alpha_1/v_1 \ge \alpha_2/v_2. \end{cases}$$

For the one-level free energy,

$$\partial_v f(\beta; a, v) = \begin{cases} \beta^2/2, & \beta \le \sqrt{2a/v}, \\ \beta\sqrt{a/(2v)}, & \beta \ge \sqrt{2a/v}. \end{cases} \tag{32}$$

The two expressions agree when $\beta = \sqrt{2a/v}$. The derivatives of the two branches of $F(\beta; v_1, v_2)$ also agree when $\alpha_1/v_1 = \alpha_2/v_2$, so the differentiability required in (31) holds.

From (29) and (30),

$$\lim_{d\to\infty} \mathbb{E}P_{\text{cl}} = 1 - \frac{2}{\beta^2}\partial_{v_1} F(\beta; v_1, v_2), \qquad \lim_{d\to\infty} \mathbb{E}P_{\text{tok}} = 1 - \frac{2}{\beta^2}\partial_{v_2} F(\beta; v_1, v_2). \tag{33}$$

Set $v_1 = \rho$ and $v_2 = 1 - \rho$. If $b_1 < b_2$, then the first branch of the general-variance formula applies, and (32) gives

$$\lim_{d\to\infty} \mathbb{E}P_{\text{cl}} = \left(1 - \frac{b_1}{\beta}\right)_+, \qquad \lim_{d\to\infty} \mathbb{E}P_{\text{tok}} = \left(1 - \frac{b_2}{\beta}\right)_+.$$

If $b_1 \ge b_2$, then

$$F(\beta; v_1, v_2) = f(\beta; \alpha, v_1 + v_2).$$

At $v_1 + v_2 = 1$, both partial derivatives in (33) equal the variance derivative in (32) with $a = \alpha$ and $v = 1$. Hence

$$\lim_{d\to\infty} \mathbb{E}P_{\text{cl}} = \lim_{d\to\infty} \mathbb{E}P_{\text{tok}} = \left(1 - \frac{b}{\beta}\right)_+.$$

These are exactly the two cases in (8). □

For a probability vector w, $(\max_i w_i)^2 \le \sum_i w_i^2 \le \max_i w_i$. Thus vanishing expected token participation implies $\max_{a,j} p_{aj} \xrightarrow{\mathbb{P}} 0$ by Markov's inequality. If the maximum mass tended to one in probability, its participation would tend to one in probability and, being bounded, in expectation. Formula (8) rules that out for both tokens and clusters at finite β. The positive cluster-participation limit in the intermediate regime implies a nonvanishing probability of a macroscopic cluster weight, not that one cluster absorbs all mass with probability tending to one. The richer random cascade description is classical (Bovier & Kurkova, 2004) and is not needed for these statements.

C COUNT CORRECTION AND OPTIMAL TEMPERATURE: COMPLETE PROOFS

The count-corrected log odds are exactly

$$\log \frac{1 - A_*^{\mathrm{cnt}}}{A_*^{\mathrm{cnt}}} = \log \mathcal{Z}_d - \log m_d - \beta X_*. \tag{34}$$

Their normalized limit is $F - \alpha_2 - \beta s$, giving the fixed-temperature threshold in (9).

C.1 THE UNIQUE LIMITING OPTIMUM

Lemma 11 (Shape of the corrected threshold). *For $0 < \rho < 1$, $h_{\mathrm{cnt}}(\beta) := (F(\beta) - \alpha_2)/\beta$ is strictly decreasing on $(0, \sqrt{2\alpha_1})$ and strictly increasing on $(\sqrt{2\alpha_1}, \infty)$.*

Proof. Set

$$c := \sqrt{2\alpha_1}, \qquad b := \sqrt{2\alpha}, \qquad b_1 := \sqrt{\frac{2\alpha_1}{\rho}}, \qquad b_2 := \sqrt{\frac{2\alpha_2}{1-\rho}}.$$

Suppose first that $\rho \le \alpha_1/\alpha$. Then $F(\beta) = f(\beta; \alpha, 1)$ by (5), and (4) gives

$$h_{\mathrm{cnt}}(\beta) = \begin{cases} \dfrac{\alpha_1}{\beta} + \dfrac{\beta}{2}, & 0 < \beta \le b, \\ b - \dfrac{\alpha_2}{\beta}, & \beta \ge b. \end{cases} \tag{35}$$

The derivatives of the two expressions in (35) are

$$h'_{\mathrm{cnt}}(\beta) = \begin{cases} -\dfrac{\alpha_1}{\beta^2} + \dfrac{1}{2}, & 0 < \beta < b, \\ \dfrac{\alpha_2}{\beta^2}, & \beta > b. \end{cases} \tag{36}$$

For the first derivative,

$$-\frac{\alpha_1}{\beta^2} + \frac{1}{2} \begin{cases} < 0, & 0 < \beta < c, \\ = 0, & \beta = c, \\ > 0, & \beta > c. \end{cases}$$

Moreover, $c = \sqrt{2\alpha_1} < \sqrt{2(\alpha_1 + \alpha_2)} = b$, and hence, (36) is negative on $(0, c)$ and positive on (c, ∞). The two formulas in (35) agree at $\beta = b$, since

$$\frac{\alpha_1}{b} + \frac{b}{2} = b - \frac{\alpha_2}{b},$$

which follows from $b^2 = 2(\alpha_1 + \alpha_2)$.

Now suppose that $\rho > \alpha_1/\alpha$. This inequality is equivalent to $b_1 < b_2$. Substituting the three formulas for F from (7) into h_{cnt} gives

$$h_{\mathrm{cnt}}(\beta) = \begin{cases} \dfrac{\alpha_1}{\beta} + \dfrac{\beta}{2}, & 0 < \beta \le b_1, \\ \sqrt{2\rho\alpha_1} + \dfrac{\beta(1-\rho)}{2}, & b_1 \le \beta \le b_2, \\ \sqrt{2\rho\alpha_1} + \sqrt{2(1-\rho)\alpha_2} - \dfrac{\alpha_2}{\beta}, & \beta \ge b_2. \end{cases} \tag{37}$$

Therefore

$$h'_{\mathrm{cnt}}(\beta) = \begin{cases} -\dfrac{\alpha_1}{\beta^2} + \dfrac{1}{2}, & 0 < \beta < b_1, \\ \dfrac{1-\rho}{2}, & b_1 < \beta < b_2, \\ \dfrac{\alpha_2}{\beta^2}, & \beta > b_2. \end{cases} \tag{38}$$

Because $0 < \rho < 1$,

$$c = \sqrt{2\alpha_1} < \sqrt{\frac{2\alpha_1}{\rho}} = b_1.$$

Thus the first expression in (38) is negative on $(0, c)$ and positive on (c, b_1), while the other two expressions are strictly positive. The formulas in (37) agree at $\beta = b_1$ and $\beta = b_2$; explicitly,

$$\frac{\alpha_1}{b_1} + \frac{b_1}{2} = \sqrt{2\rho\alpha_1} + \frac{b_1(1-\rho)}{2},$$
$$\sqrt{2\rho\alpha_1} + \frac{b_2(1-\rho)}{2} = \sqrt{2\rho\alpha_1} + \sqrt{2(1-\rho)\alpha_2} - \frac{\alpha_2}{b_2},$$

using $b_1^2 = 2\alpha_1/\rho$ and $b_2^2 = 2\alpha_2/(1-\rho)$.

In both parameter cases, $h'_{\mathrm{cnt}}(\beta) < 0$ for $0 < \beta < c$ and $h'_{\mathrm{cnt}}(\beta) > 0$ for $\beta > c$. Hence the unique minimizer is

$$\beta = c = \sqrt{2\alpha_1}.$$

Finally,

$$h_{\mathrm{cnt}}(c) = \frac{\alpha_1}{c} + \frac{c}{2} = c,$$

because $c^2 = 2\alpha_1$. This proves (10). □

C.2 Uniform control over all temperatures

Lemma 12 (Maximum scores and compact uniform convergence)**.** *Let $V_d = d^{-1}\max_{a,j} X_{aj}$ and $B_d = (N_d d)^{-1}\sum_{a,j} X_{aj}$. Then $B_d \xrightarrow{\mathbb{P}} 0$, and $V_d \xrightarrow{\mathbb{P}} v$ where*

$$v = \begin{cases} \sqrt{2\alpha}, & \rho \le \alpha_1/\alpha, \\ \sqrt{2\rho\alpha_1} + \sqrt{2(1-\rho)\alpha_2}, & \rho > \alpha_1/\alpha. \end{cases} \tag{39}$$

For each $0 < t < T < \infty$, $\sup_{\beta\in[t,T]} |L_d(\beta) - F(\beta)| \xrightarrow{\mathbb{P}} 0$.

Proof. Using (2) and $N_d = M_d m_d$, write

$$B_d = \frac{1}{N_d d}\sum_{a=1}^{M_d}\sum_{j=1}^{m_d}\left(\sqrt{\rho d}\, G_a + \sqrt{(1-\rho)d}\, H_{aj}\right)$$
$$= \frac{\sqrt{\rho d}}{M_d d}\sum_{a=1}^{M_d} G_a + \frac{\sqrt{(1-\rho)d}}{N_d d}\sum_{a=1}^{M_d}\sum_{j=1}^{m_d} H_{aj}. \tag{40}$$

The two sums in (40) are independent and centered. Therefore $\mathbb{E}B_d = 0$ and

$$\mathrm{Var}(B_d) = \frac{\rho d}{M_d^2 d^2} M_d + \frac{(1-\rho)d}{N_d^2 d^2} N_d = \frac{\rho}{M_d d} + \frac{1-\rho}{N_d d}.$$

Since $M_d \to \infty$ and $N_d \to \infty$, Chebyshev's inequality gives, for every $\varepsilon > 0$,

$$\mathbb{P}(|B_d| > \varepsilon) \le \frac{1}{\varepsilon^2}\left(\frac{\rho}{M_d d} + \frac{1-\rho}{N_d d}\right) \longrightarrow 0.$$

Hence $B_d \xrightarrow{\mathbb{P}} 0$.

We next prove the limit of V_d. For every fixed $\beta > 0$, $e^{\beta d V_d} \le \mathcal{Z}_d(\beta) \le N_d e^{\beta d V_d}$. Taking logarithms and dividing by βd gives

$$\frac{L_d(\beta)}{\beta} - \frac{\log N_d}{\beta d} \le V_d \le \frac{L_d(\beta)}{\beta}. \tag{41}$$

For each fixed β, Theorem 1 gives

$$\frac{L_d(\beta)}{\beta} \xrightarrow{\mathbb{P}} \frac{F(\beta)}{\beta}, \qquad \frac{\log N_d}{\beta d} \longrightarrow \frac{\alpha}{\beta}. \tag{42}$$

Moreover, combining (4) and (5), and dividing by β gives

$$\lim_{\beta\to\infty}\frac{F(\beta)}{\beta}=\begin{cases}\sqrt{2\alpha}, & \rho\le\alpha_1/\alpha,\\ \sqrt{2\rho\alpha_1}+\sqrt{2(1-\rho)\alpha_2}, & \rho>\alpha_1/\alpha,\end{cases}=v.$$

Fix $\varepsilon>0$. Choose a finite number $\beta_0>0$ sufficiently large that

$$\left|\frac{F(\beta_0)}{\beta_0}-v\right|<\varepsilon,\qquad\frac{\alpha+1}{\beta_0}<\varepsilon.$$

Since $d^{-1}\log N_d\to\alpha$, for all sufficiently large d, we have $d^{-1}\log N_d\le\alpha+1$. On the event $|L_d(\beta_0)-F(\beta_0)|\le\beta_0\varepsilon$, the upper bound in (41) gives

$$V_d\le\frac{F(\beta_0)}{\beta_0}+\varepsilon\le v+2\varepsilon,$$

and the lower bound gives

$$V_d\ge\frac{F(\beta_0)}{\beta_0}-\varepsilon-\frac{\alpha+1}{\beta_0}\ge v-3\varepsilon.$$

The probability of this event tends to one by (42). Hence

$$\mathbb{P}(|V_d-v|>3\varepsilon)\longrightarrow 0,$$

and since $\varepsilon>0$ was arbitrary, $V_d\xrightarrow{\mathbb{P}}v$.

Finally, fix $0<t<T<\infty$. Define

$$Q_d:=\frac{1}{d}\max_{a,j}|X_{aj}|.$$

Since

$$\frac{\partial L_d(\beta)}{\partial\beta}=\frac{1}{d}\sum_{a,j}p_{aj}(\beta)X_{aj},$$

we have

$$\left|\frac{\partial L_d(\beta)}{\partial\beta}\right|\le\frac{1}{d}\sum_{a,j}p_{aj}(\beta)|X_{aj}|\le Q_d.$$

Therefore, for every $\beta,\gamma>0$, $|L_d(\beta)-L_d(\gamma)|\le Q_d|\beta-\gamma|$.

Since every X_{aj} is marginally $N(0,d)$, for every $R>0$, a union bound gives

$$\mathbb{P}(Q_d>R)\le 2N_d\exp\left(-\frac{dR^2}{2}\right).\tag{43}$$

Choose R so that $R^2/2>\alpha+1$. Since $d^{-1}\log N_d\to\alpha$, the right-hand side of (43) tends to zero. Hence $\mathbb{P}(Q_d\le R)\longrightarrow 1$.

Let $\varepsilon>0$. Because F is continuous on the compact interval $[t,T]$, choose $\delta>0$ such that

$$|\beta-\gamma|\le\delta\quad\Longrightarrow\quad|F(\beta)-F(\gamma)|\le\frac{\varepsilon}{3},$$

and decrease δ further so that $R\delta\le\varepsilon/3$. Choose finitely many grid points

$$t=\beta_0<\beta_1<\cdots<\beta_K=T$$

such that every $\beta\in[t,T]$ is within distance δ of at least one β_k. At each fixed grid point, $L_d(\beta_k)\xrightarrow{\mathbb{P}}F(\beta_k)$. Because the grid contains only finitely many points, the union bound gives

$$\mathbb{P}\left(\max_{0\le k\le K}|L_d(\beta_k)-F(\beta_k)|>\frac{\varepsilon}{3}\right)\longrightarrow 0.$$

On the event $Q_d \le R$ and $\max_{0\le k\le K} |L_d(\beta_k) - F(\beta_k)| \le \varepsilon/3$, choose for each $\beta \in [t, T]$ a grid point β_k satisfying $|\beta - \beta_k| \le \delta$. Then

$$\begin{aligned} |L_d(\beta) - F(\beta)| &\le |L_d(\beta) - L_d(\beta_k)| + |L_d(\beta_k) - F(\beta_k)| + |F(\beta_k) - F(\beta)| \\ &\le R\delta + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} \le \varepsilon. \end{aligned}$$

Therefore

$$\mathbb{P}\left(\sup_{\beta\in[t,T]} |L_d(\beta) - F(\beta)| > \varepsilon \right) \longrightarrow 0.$$

This proves the compact uniform convergence. □

Proof of Theorem 3. Put $c := \sqrt{2\alpha_1}$ and

$$H_d(\beta) = \frac{\log \mathcal{Z}_d(\beta) - \log m_d}{\beta d}. \tag{44}$$

Lemma 11 gives

$$h_{\mathrm{cnt}}(\beta) \ge c \quad \text{for every } \beta > 0, \qquad h_{\mathrm{cnt}}(\beta) = c \iff \beta = c. \tag{45}$$

For every fixed $\beta > 0$, (34) and $L_d(\beta) \xrightarrow{\mathbb{P}} F(\beta)$ give

$$\frac{1}{d} \log \frac{1 - A_*^{\mathrm{cnt}}(\beta)}{A_*^{\mathrm{cnt}}(\beta)} = \beta \left[\frac{L_d(\beta) - d^{-1} \log m_d}{\beta} - \frac{X_*}{d} \right] \xrightarrow{\mathbb{P}} \beta \big[h_{\mathrm{cnt}}(\beta) - s \big]. \tag{46}$$

If $s < c$, then the limit in (46) is strictly positive for every fixed β, and therefore $A_*^{\mathrm{cnt}}(\beta) \xrightarrow{\mathbb{P}} 0$. If $s > c$, choose $\beta = c$. Then the limit is $c(c - s) < 0$, and hence $A_*^{\mathrm{cnt}}(c) \xrightarrow{\mathbb{P}} 1$. This proves the fixed-temperature assertions.

We next prove (11). Jensen's inequality gives

$$\mathcal{Z}_d(\beta) \ge N_d \exp\left\{ \frac{\beta}{N_d} \sum_{a,j} X_{aj} \right\} = N_d e^{\beta d B_d}. \tag{47}$$

where $B_d = (N_d d)^{-1} \sum_{a,j} X_{aj}$. Also, the largest summand in $\mathcal{Z}_d(\beta)$ gives

$$\mathcal{Z}_d(\beta) \ge e^{\beta d V_d}. \tag{48}$$

Substituting (47) and (48) into (44) yields, for every $\beta > 0$,

$$H_d(\beta) \ge \frac{\log M_d}{\beta d} + B_d, \qquad H_d(\beta) \ge V_d - \frac{\log m_d}{\beta d}. \tag{49}$$

The limit v in (39) is strictly larger than c. We first aim to prove the lower bound

$$\inf_{\beta>0} H_d(\beta) \ge c - \varepsilon \qquad \text{for } 0 < \varepsilon < \min\left\{1, \frac{v - c}{4}\right\}. \tag{50}$$

For such an ε, choose fixed numbers $0 < t < T < \infty$ such that

$$\frac{\alpha_1}{2t} > c + 1, \qquad v - \frac{\alpha_2 + 1}{T} > c + 2\varepsilon. \tag{51}$$

By Lemma 12 and the definitions of M_d and m_d, with probability tending to one, we have $B_d \ge -\varepsilon$, $V_d \ge v - \varepsilon$, $d^{-1} \log M_d \ge \frac{\alpha_1}{2}$ and $d^{-1} \log m_d \le \alpha_2 + 1$. On this event, (49) gives

$$H_d(\beta) \ge \frac{\alpha_1}{2t} - \varepsilon \qquad \text{for } 0 < \beta < t, \tag{52}$$

and

$$H_d(\beta) \ge v - \varepsilon - \frac{\alpha_2 + 1}{T} \qquad \text{for } \beta > T. \tag{53}$$

Both lower bounds in (52) and (53) exceed $c - \varepsilon$. On the fixed compact interval $[t, T]$,

$$\sup_{\beta \in [t,T]} |H_d(\beta) - h_{\mathrm{cnt}}(\beta)| \xrightarrow{\mathbb{P}} 0, \tag{54}$$

because $H_d(\beta) = L_d(\beta)/\beta - d^{-1}\log(m_d)/\beta$ and Lemma 12 gives uniform convergence of L_d there. Combining (54) with (45) shows

$$\inf_{\beta \in [t,T]} H_d(\beta) \geq c - \varepsilon \tag{55}$$

with probability tending to one. Equations (52), (53), and (55) imply (50) with probability tending to one. Conversely,

$$\inf_{\beta > 0} H_d(\beta) \leq H_d(c) \xrightarrow{\mathbb{P}} h_{\mathrm{cnt}}(c) = c. \tag{56}$$

Since $\varepsilon > 0$ is arbitrary, (50) and (56) prove (11).

It remains to show uniform failure when $s < c$. Set $\delta := (c - s)/4 > 0$ and choose

$$\varepsilon := \min\left\{\frac{\delta}{2}, \frac{1}{2}, \frac{v - c}{8}\right\}.$$

Then $0 < \varepsilon < \min\{1, (v - c)/4\}$ and $\varepsilon < \delta$. The lower-bound argument above, with this choice of ε, shows that, with probability tending to one,

$$H_d(\beta) \geq c - \varepsilon \quad \text{for every } \beta > 0. \tag{57}$$

On the same event, we may also require $X_*/d \leq s + \delta$. For every $\beta \geq t$, (34) and (57) give

$$\begin{aligned} \log \frac{1 - A_*^{\mathrm{cnt}}(\beta)}{A_*^{\mathrm{cnt}}(\beta)} &= \beta d \left[H_d(\beta) - \frac{X_*}{d}\right] \\ &\geq \beta d(c - s - \delta - \varepsilon) \\ &\geq td(c - s - 2\delta) = \frac{td(c - s)}{2}. \end{aligned} \tag{58}$$

For $0 < \beta < t$, use (47) directly in (34):

$$d^{-1} \log \frac{1 - A_*^{\mathrm{cnt}}}{A_*^{\mathrm{cnt}}} \geq d^{-1} \log M_d + \beta(B_d - X_*/d). \tag{59}$$

Decrease t, if necessary, while preserving (51), so that

$$t(|s| + 2) < \frac{\alpha_1}{4}. \tag{60}$$

With probability tending to one, we have $|B_d| \leq 1$, $|X_*/d| \leq |s| + 1$, and $d^{-1}\log M_d \geq \alpha_1/2$. On this event, (59) and (60) imply, for every $0 < \beta < t$,

$$d^{-1} \log \frac{1 - A_*^{\mathrm{cnt}}(\beta)}{A_*^{\mathrm{cnt}}(\beta)} \geq \frac{\alpha_1}{2} - t(|s| + 2) > \frac{\alpha_1}{4}. \tag{61}$$

Together, (58) and (61) show that

$$\inf_{\beta > 0} \log \frac{1 - A_*^{\mathrm{cnt}}(\beta)}{A_*^{\mathrm{cnt}}(\beta)} \xrightarrow{\mathbb{P}} +\infty.$$

This implies $\sup_{\beta > 0} A_*^{\mathrm{cnt}}(\beta) \xrightarrow{\mathbb{P}} 0$. □

Proof of Proposition 4. The event $A_*^{\mathrm{cnt}} \leq 1 - \varepsilon$ is equivalent to $\mathcal{Z}_d/m_d \geq [\varepsilon/(1-\varepsilon)]e^{\beta s d}$. For every fixed (a, j), $X_{aj} \sim N(0, d)$, and hence $\mathbb{E}e^{\beta X_{aj}} = \exp(\beta^2 d/2)$. Using linearity of expectation,

$$\mathbb{E}\left[\frac{\mathcal{Z}_d}{m_d}\right] = \frac{1}{m_d} \sum_{a=1}^{M_d} \sum_{j=1}^{m_d} \mathbb{E}e^{\beta X_{aj}} = M_d \exp(\beta^2 d/2).$$

Markov's inequality proves (12).

Completing the square $\beta^2/2 - \beta s = (\beta - s)^2/2 - s^2/2$ gives the minimizing positive temperature when $s > 0$. This minimizes the upper bound for a given s, whereas $\beta = \sqrt{2\alpha_1}$ minimizes the asymptotic signal threshold; these are different optimization problems. □

D AVERAGING, ESTIMATION ERROR, AND VALUE RECOVERY

Proof of Theorem 5. For every cluster,

$$e^{\beta \bar{X}_a} = \exp\left\{\frac{1}{m_d}\sum_{j=1}^{m_d}\beta X_{aj}\right\} \le \frac{1}{m_d}\sum_{j=1}^{m_d} e^{\beta X_{aj}}.$$

by Jensen's inequality. Summing over a gives $\mathcal{Z}_d^{\mathrm{avg}} = \sum_{a=1}^{M_d} e^{\beta \bar{X}_a} \le \mathcal{Z}_d/m_d$. Therefore, we have

$$A_*^{\mathrm{avg}} = \frac{e^{\beta X_*}}{e^{\beta X_*} + \mathcal{Z}_d^{\mathrm{avg}}} \ge \frac{e^{\beta X_*}}{e^{\beta X_*} + \mathcal{Z}_d/m_d} = A_*^{\mathrm{cnt}}.$$

For the Gaussian model,

$$\bar{X}_a = \sqrt{\rho d}\, G_a + R_a, \qquad R_a := \sqrt{(1-\rho)d}\,\frac{1}{m_d}\sum_{j=1}^{m_d} H_{aj}.$$

Since $m_d^{-1}\sum_j H_{aj} \sim N(0, m_d^{-1})$, $R_a \sim N\left(0, \frac{(1-\rho)d}{m_d}\right)$. The R_a are independent across clusters and independent of $(G_a)_a$. Hence, by the union bound, for every $u > 0$,

$$\mathbb{P}\left\{\max_a |R_a| > ud\right\} \le 2M_d \exp\left[-\frac{u^2 d m_d}{2(1-\rho)}\right] \longrightarrow 0.$$

Put $Y_a := \sqrt{\rho d}\, G_a$. For every realized array,

$$e^{-\beta \max_a |R_a|}\sum_a e^{\beta Y_a} \le \sum_a e^{\beta(Y_a + R_a)} \le e^{\beta \max_a |R_a|}\sum_a e^{\beta Y_a}.$$

Taking logarithms and dividing by d gives

$$\left|\frac{1}{d}\log \mathcal{Z}_d^{\mathrm{avg}} - \frac{1}{d}\log\sum_{a=1}^{M_d} e^{\beta\sqrt{\rho d}G_a}\right| \le \frac{\beta}{d}\max_a |R_a| \xrightarrow{\mathbb{P}} 0.$$

The remaining sum has $M_d = \exp(\alpha_1 d + o(d))$ independent $N(0, \rho d)$ scores. The one-level version of Lemma 7 therefore gives

$$\frac{1}{d}\log\sum_{a=1}^{M_d} e^{\beta\sqrt{\rho d}G_a} \xrightarrow{\mathbb{P}} f(\beta; \alpha_1, \rho),$$

and thus $d^{-1}\log \mathcal{Z}_d^{\mathrm{avg}} \xrightarrow{\mathbb{P}} f(\beta; \alpha_1, \rho)$. Since

$$\log\frac{1 - A_*^{\mathrm{avg}}}{A_*^{\mathrm{avg}}} = \log \mathcal{Z}_d^{\mathrm{avg}} - \beta X_*,$$

we have

$$\frac{1}{\beta d}\log\frac{1 - A_*^{\mathrm{avg}}}{A_*^{\mathrm{avg}}} \xrightarrow{\mathbb{P}} \frac{f(\beta; \alpha_1, \rho)}{\beta} - s = h_{\mathrm{avg}}(\beta) - s.$$

Thus $A_*^{\mathrm{avg}} \xrightarrow{\mathbb{P}} 1$ when $s > h_{\mathrm{avg}}(\beta)$ and $A_*^{\mathrm{avg}} \xrightarrow{\mathbb{P}} 0$ when $s < h_{\mathrm{avg}}(\beta)$. Put $b_1 := \sqrt{2\alpha_1/\rho}$. The one-level free energy $f(\beta; \alpha_1, \rho)$ gives

$$h_{\mathrm{avg}}(\beta) = \begin{cases} \dfrac{\alpha_1}{\beta} + \dfrac{\rho\beta}{2}, & 0 < \beta \le b_1, \\ \sqrt{2\rho\alpha_1}, & \beta \ge b_1. \end{cases}$$

The derivative of the first expression is $-\alpha_1/\beta^2 + \rho/2 < 0$ for $0 < \beta < b_1$, so it is decreasing and its value at b_1 is $\sqrt{2\rho\alpha_1}$. Hence $\inf_{\beta>0} h_{\mathrm{avg}}(\beta) = \sqrt{2\rho\alpha_1}$ with equality for every $\beta \ge b_1$.

Since $0 < \rho < 1$, $\sqrt{2\rho\alpha_1} < \sqrt{2\alpha_1}$. For every s strictly between these values, $A_*^{\mathrm{avg}} \xrightarrow{\mathbb{P}} 1$ for every fixed $\beta \ge b_1$, while Theorem 3 gives $\sup_{\beta>0} A_*^{\mathrm{cnt}}(\beta) \xrightarrow{\mathbb{P}} 0$ even when the count-correction temperature is selected after observing the distractor scores. □

Proof of Corollary 6. Let $Y_1, \dots, Y_{L_d}$ be the added singleton distractor scores and put $Z_{0,d}(\beta) := \sum_{\ell=1}^{L_d} e^{\beta Y_\ell}$. Since $L_d = \exp(a_0 d + o(d))$ and the Y_ℓ are independent $N(0,d)$ variables, the one-level free-energy formula gives $d^{-1} \log Z_{0,d}(\beta) \xrightarrow{\mathbb{P}} f(\beta; a_0, 1)$. Count correction leaves each added singleton unchanged because its group size is one. Thus the total distractor weights in the augmented model are

$$D_{\mathrm{cnt},0}(\beta) := \frac{\mathcal{Z}_d(\beta)}{m_d} + Z_{0,d}(\beta), \qquad D_{\mathrm{avg},0}(\beta) := \mathcal{Z}_d^{\mathrm{avg}}(\beta) + Z_{0,d}(\beta).$$

Write

$$A_{*,0}^{\mathrm{cnt}}(\beta) := \frac{e^{\beta X_*}}{e^{\beta X_*} + D_{\mathrm{cnt},0}(\beta)}, \qquad A_{*,0}^{\mathrm{avg}}(\beta) := \frac{e^{\beta X_*}}{e^{\beta X_*} + D_{\mathrm{avg},0}(\beta)}.$$

Their exact log odds are

$$\log \frac{1 - A_{*,0}^{\mathrm{cnt}}(\beta)}{A_{*,0}^{\mathrm{cnt}}(\beta)} = \log D_{\mathrm{cnt},0}(\beta) - \beta X_*, \qquad \log \frac{1 - A_{*,0}^{\mathrm{avg}}(\beta)}{A_{*,0}^{\mathrm{avg}}(\beta)} = \log D_{\mathrm{avg},0}(\beta) - \beta X_*. \quad (62)$$

For positive U, V, observe that

$$\max\{\log U, \log V\} \le \log(U+V) \le \max\{\log U, \log V\} + \log 2.$$

The term $d^{-1} \log 2$ vanishes. Applying this bound to the two sums defining $D_{\mathrm{cnt},0}$ and $D_{\mathrm{avg},0}$, and using Theorems 3 and 5, gives

$$\frac{1}{\beta d} \log \frac{1 - A_{*,0}^{\mathrm{cnt}}(\beta)}{A_{*,0}^{\mathrm{cnt}}(\beta)} \xrightarrow{\mathbb{P}} h_{\mathrm{cnt},0}(\beta) - s, \qquad \frac{1}{\beta d} \log \frac{1 - A_{*,0}^{\mathrm{avg}}(\beta)}{A_{*,0}^{\mathrm{avg}}(\beta)} \xrightarrow{\mathbb{P}} h_{\mathrm{avg},0}(\beta) - s, \quad (63)$$

where $h_{\mathrm{cnt},0}(\beta) = \max\{h_{\mathrm{cnt}}(\beta), f(\beta; a_0, 1)/\beta\}$ and $h_{\mathrm{avg},0}(\beta) = \max\{h_{\mathrm{avg}}(\beta), f(\beta; a_0, 1)/\beta\}$.

Put $c := \sqrt{2\alpha_1}$. Theorem 3 gives $h_{\mathrm{cnt}}(\beta) \ge c$ for every $\beta > 0$ and $h_{\mathrm{cnt}}(c) = c$. This implies $h_{\mathrm{cnt},0}(\beta) \ge h_{\mathrm{cnt}}(\beta) \ge c$ for every $\beta > 0$. At $\beta = c$, since $a_0 < \alpha_1$, we also have $c > \sqrt{2a_0}$, and the definition (4) gives

$$\frac{f(c; a_0, 1)}{c} = \sqrt{2a_0} < c.$$

Therefore, $h_{\mathrm{cnt},0}(\beta) \ge c$ for every $\beta > 0$ and $h_{\mathrm{cnt},0}(c) = c$.

If $s > c$, the first limit in (63) at $\beta = c$ is negative, so $A_{*,0}^{\mathrm{cnt}}(c) \xrightarrow{\mathbb{P}} 1$. If $s < c$, then samplewise

$$A_{*,0}^{\mathrm{cnt}}(\beta) = \frac{e^{\beta X_*}}{e^{\beta X_*} + \mathcal{Z}_d(\beta)/m_d + Z_{0,d}(\beta)} \le \frac{e^{\beta X_*}}{e^{\beta X_*} + \mathcal{Z}_d(\beta)/m_d} = A_*^{\mathrm{cnt}}(\beta).$$

Hence

$$\sup_{\beta > 0} A_{*,0}^{\mathrm{cnt}}(\beta) \le \sup_{\beta > 0} A_*^{\mathrm{cnt}}(\beta) \xrightarrow{\mathbb{P}} 0$$

by Theorem 3, which proves failure for every data-dependent temperature choice.

For averaging, Theorem 5 gives $h_{\mathrm{avg}}(\beta) = \sqrt{2\rho\alpha_1}$ for $\beta \ge b_1$, and (4) gives

$$\frac{f(\beta; a_0, 1)}{\beta} = \sqrt{2a_0} \qquad \text{for } \beta \ge \sqrt{2a_0}.$$

Thus, for every fixed $\beta \ge \max\{b_1, \sqrt{2a_0}\}$, we have

$$h_{\mathrm{avg},0}(\beta) = \max\{\sqrt{2\rho\alpha_1}, \sqrt{2a_0}\} = c_{\mathrm{avg},0}.$$

If $s > c_{\mathrm{avg},0}$, the second limit in (63) is negative, so $A_{*,0}^{\mathrm{avg}}(\beta) \xrightarrow{\mathbb{P}} 1$. Moreover,

$$\sqrt{2\rho\alpha_1} < \sqrt{2\alpha_1} = c, \qquad \sqrt{2a_0} < \sqrt{2\alpha_1} = c,$$

because $0 < \rho < 1$ and $a_0 < \alpha_1$. Hence $c_{\mathrm{avg},0} < c$, proving (15) and the stated separation interval. If $a_0 \le \rho\alpha_1$, then $\sqrt{2a_0} \le \sqrt{2\rho\alpha_1}$ and therefore $c_{\mathrm{avg},0} = \sqrt{2\rho\alpha_1}$, so the original interval is unchanged.

There are $L_d + 1$ singleton groups, and the planted item is placed uniformly among them. A method that sees only the group memberships and group sizes cannot distinguish these singleton groups. If it selects singleton ℓ with probability q_ℓ, then its success probability is

$$\sum_{\ell=1}^{L_d+1} q_\ell \frac{1}{L_d+1} = \frac{1}{L_d+1}.$$

This bound does not apply to methods that also use the scores. □

Proposition 13 (Deterministic relative-log-error bound)**.** *Let D and $\widehat{D}$ be the log odds against the planted item obtained from count correction using the true counts m_i and estimated counts $\widehat{m}_i$. Under (16), $|\widehat{D} - D| \le r_d$. If every score used by averaging, including the planted score, is perturbed by at most u_d, then its log odds change by at most $2\beta u_d$.*

Proof. For count correction,

$$e^D = \sum_{i \neq *} e^{\beta(X_i - X_*)} \frac{m_*}{m_i}.$$

Put $e_i := \log(\widehat{m}_i/m_i)$. Replacing m_i by $\widehat{m}_i$ multiplies the ith summand by $e^{e_* - e_i}$, which lies in $[e^{-r_d}, e^{r_d}]$ by (16). Hence $e^{-r_d} e^D \le e^{\widehat{D}} \le e^{r_d} e^D$, and taking logarithms gives $|\widehat{D} - D| \le r_d$.

For averaging, perturbing both scores in $X_i - X_*$ changes this difference by at most $2u_d$. Each odds summand is therefore multiplied by a factor in $[e^{-2\beta u_d}, e^{2\beta u_d}]$, giving the second bound in the same way. Thus either conclusion is unchanged at fixed β and away from the limiting boundary when the corresponding error is $o(d)$. □

Proposition 14 (Value error)**.** *Let $\widehat{v} := \sum_i A_i v_i$, where $A_i \ge 0$, $\sum_i A_i = 1$, and $\|v_i\| \le R$. Then*

$$\|\widehat{v} - v_*\| \le 2R(1 - A_*).$$

The same bound holds after averaging values within groups when the planted item is a singleton. Consequently, $A_* \xrightarrow{\mathbb{P}} 1$ implies $\widehat{v} \xrightarrow{\mathbb{P}} v_*$ and $\mathbb{E}\|\widehat{v} - v_*\|^2 \to 0$.

Proof. Since $\sum_i A_i = 1$, we have $\widehat{v} - v_* = \sum_{i \neq *} A_i (v_i - v_*)$. Therefore

$$\|\widehat{v} - v_*\| \le \sum_{i \neq *} A_i \|v_i - v_*\| \le 2R \sum_{i \neq *} A_i = 2R(1 - A_*).$$

An average of vectors with norm at most R also has norm at most R, and averaging leaves a singleton planted value unchanged, so the same calculation applies after group averaging. If $A_* \xrightarrow{\mathbb{P}} 1$, the displayed bound gives $\widehat{v} \xrightarrow{\mathbb{P}} v_*$. Moreover, $\|\widehat{v} - v_*\|^2 \le 4R^2$, so convergence in probability and boundedness imply $\mathbb{E}\|\widehat{v} - v_*\|^2 \to 0$. □

E A FINITE-ARRAY ENTROPY CHARACTERIZATION

Let $\widehat{\mu} = N^{-1} \sum_i \delta_{x_i}$ be the empirical score law. In the tail-probability convention, define the classical entropic value-at-risk (Ahmadi-Javid, 2012) by

$$\mathrm{EVaR}_q(\widehat{\mu}) = \inf_{\beta > 0} \frac{\log \int e^{\beta x} \, d\widehat{\mu}(x) - \log q}{\beta}.$$

Our optimized threshold is exactly $d^{-1} \mathrm{EVaR}_{1/M}(\widehat{\mu})$. Its entropy dual below is the standard relative-entropy dual: $D(p \| \mathrm{Unif}_N) \le \log M$ is equivalent to $S(p) \ge \log m$. We give a direct finite-array proof for clarity and do not claim this identity as new. The empirical limit in Theorem 3 concerns dependent triangular arrays with an exponentially vanishing tail probability; replacing their empirical law by a Gaussian population law requires that theorem, not just the identity.

Fix integers $M, m > 1$, put $N = Mm$, and fix distractor scores $x_1, \dots, x_N$. Define

$$Z(\beta) := \sum_{i=1}^N e^{\beta x_i}, \qquad p_i(\beta) := \frac{e^{\beta x_i}}{Z(\beta)}, \qquad S(p) := -\sum_{i=1}^N p_i \log p_i,$$

and

$$H(\beta) := \frac{\log Z(\beta) - \log m}{\beta d}.$$

Proposition 15 (EVaR identity and exact temperature calibration)**.** *Assume that the largest score is unique. Then H has a unique minimizer $\widehat{\beta} > 0$, characterized by $S(p(\widehat{\beta})) = \log m$. Moreover, if Δ_N is the probability simplex, then*

$$\min_{\beta>0} H(\beta) = \frac{1}{d}\,\mathrm{EVaR}_{1/M}(\widehat{\mu}) = \frac{1}{d} \max_{\substack{q\in\Delta_N \\ S(q)\geq \log m}} \sum_i q_i x_i. \tag{64}$$

For a planted score x_*, there exists $\beta > 0$ such that $A_*^{\mathrm{cnt}} > 1/2$ if and only if $x_*/d > \min_{\beta>0} H(\beta)$.

Proof. Differentiation gives

$$H'(\beta) = \frac{\log m - S(p(\beta))}{\beta^2 d}, \qquad \frac{d}{d\beta} S(p(\beta)) = -\beta\,\mathrm{Var}_{p(\beta)}(x) < 0.$$

The variance is positive because the scores are not all equal and every $p_i(\beta)$ is positive. Moreover, $S(p(\beta)) \to \log N$ as $\beta \downarrow 0$ and $S(p(\beta)) \to 0$ as $\beta \to \infty$, where the second limit uses uniqueness of the largest score. Since $0 < \log m < \log N$, the equation $S(p(\beta)) = \log m$ has one solution. The sign of H' is negative before this solution and positive after it, proving the unique-minimizer claim.

For any $q \in \Delta_N$, nonnegativity of $D(q\|p(\beta)) = \sum_i q_i \log(q_i/p_i(\beta)) = -S(q) - \beta \sum_i q_i x_i + \log Z(\beta)$ gives $\log Z(\beta) \geq \beta \sum_i q_i x_i + S(q)$ with equality when $q = p(\beta)$. If $S(q) \geq \log m$, then $H(\beta) \geq d^{-1} \sum_i q_i x_i$. At $\beta = \widehat{\beta}$, the choice $q = p(\widehat{\beta})$ satisfies $S(q) = \log m$ and gives equality, proving the last equality in (64). For the EVaR equality, use $N = Mm$:

$$\log \int e^{\beta x}\, d\widehat{\mu}(x) - \log(1/M) = \log Z(\beta) - \log N + \log M = \log Z(\beta) - \log m.$$

Finally,

$$A_*^{\mathrm{cnt}} > \frac{1}{2} \quad\Longleftrightarrow\quad e^{\beta x_*} > \frac{Z(\beta)}{m} \quad\Longleftrightarrow\quad \frac{x_*}{d} > H(\beta).$$

Such a temperature exists exactly when $x_*/d > \min_{\beta>0} H(\beta)$. □

This proposition does not require Gaussian scores, independence, or any particular correlation structure. It describes the best temperature after the distractor scores have been identified and observed, so it is not directly usable when the target location is unknown. In our Gaussian model the largest distractor score is unique almost surely, and the convergence and endpoint bounds proved in Theorem 3 give $\widehat{\beta} \xrightarrow{\mathbb{P}} \sqrt{2\alpha_1}$.

F THE SAME SEPARATION FOR ANY FIXED FINITE HIERARCHY

Fix $K \geq 2$, rates $a_1, \ldots, a_K > 0$, and variances $v_1, \ldots, v_K > 0$ with $\sum_{\ell=1}^K v_\ell = 1$. At level ℓ, every node has $n_{\ell,d} := \lceil e^{a_\ell d} \rceil$ children, and every level-ℓ edge has an independent $N(0, v_\ell d)$ score. For a leaf u, let X_u be the sum of the K edge scores on its path. Leaves are grouped by their first edge, so there are $n_{1,d}$ groups and each group contains $m_d := \prod_{\ell=2}^K n_{\ell,d}$ leaves.

Write $\mathcal{L}_d$ for the set of leaves and $\mathcal{G}_a$ for the leaves in first-level group a, and define

$$\mathcal{Z}_d^{(K)}(\beta) := \sum_{u\in\mathcal{L}_d} e^{\beta X_u}, \qquad H_d^{(K)}(\beta) := \frac{\log \mathcal{Z}_d^{(K)}(\beta) - \log m_d}{\beta d}.$$

For a singleton planted score X_*, the count-corrected mass is

$$A_{*,K}^{\mathrm{cnt}}(\beta) := \frac{e^{\beta X_*}}{e^{\beta X_*} + \mathcal{Z}_d^{(K)}(\beta)/m_d}.$$

For averaging, put

$$\bar{X}_a := \frac{1}{m_d} \sum_{u \in \mathcal{G}_a} X_u, \qquad \mathcal{Z}_d^{(K),\mathrm{avg}}(\beta) := \sum_{a=1}^{n_{1,d}} e^{\beta \bar{X}_a}.$$

Finally, write

$$a_{\le j} := \sum_{\ell=1}^{j} a_\ell, \qquad V_j := \sum_{\ell=1}^{j} v_\ell, \qquad a_{\mathrm{tot}} := a_{\le K} = \sum_{\ell=1}^{K} a_\ell.$$

Theorem 16 (Count correction and averaging in a finite hierarchy)**.** *For every fixed $\beta > 0$,*

$$\frac{1}{d} \log \mathcal{Z}_d^{(K)}(\beta) \xrightarrow{\mathbb{P}} F_K(\beta) := \max_{\substack{x \in \mathbb{R}^K \\ \sum_{\ell=1}^{j} x_\ell^2/(2v_\ell) \le a_{\le j} \ (1 \le j \le K)}} \left\{ a_{\mathrm{tot}} - \sum_{\ell=1}^{K} \frac{x_\ell^2}{2v_\ell} + \beta \sum_{\ell=1}^{K} x_\ell \right\}. \tag{65}$$

The convergence also holds in L^1. The count-corrected limiting threshold

$$h_K^{\mathrm{cnt}}(\beta) := \frac{F_K(\beta) - (a_{\mathrm{tot}} - a_1)}{\beta}$$

satisfies $h_K^{\mathrm{cnt}}(\beta) \ge \sqrt{2a_1}$ and $h_K^{\mathrm{cnt}}(\beta) = \sqrt{2a_1} \iff \beta = \sqrt{2a_1}$. Moreover, $\inf_{\beta>0} H_d^{(K)}(\beta) \xrightarrow{\mathbb{P}} \sqrt{2a_1}$, and if $X_*/d \xrightarrow{\mathbb{P}} s < \sqrt{2a_1}$, then $\sup_{\beta>0} A_{*,K}^{\mathrm{cnt}}(\beta) \xrightarrow{\mathbb{P}} 0$.

For averaging,

$$\frac{1}{d} \log \mathcal{Z}_d^{(K),\mathrm{avg}}(\beta) \xrightarrow{\mathbb{P}} f(\beta; a_1, v_1),$$

so its limiting threshold has minimum $\sqrt{2v_1 a_1}$.

Proof. We first prove (65). This is the K-level extension of the bin-counting argument in Appendix A. Partition a bounded interval into finitely many bins. For a choice of bins $B_1, \dots, B_j$ along the first j levels, put $r_\ell := \inf_{x \in B_\ell} \frac{x^2}{2v_\ell}$. The expected number of depth-j paths whose normalized edge scores lie in these bins is

$$\exp\left\{ d \left(a_{\le j} - \sum_{\ell=1}^{j} r_\ell + o(1) \right) \right\}.$$

If $\sum_{\ell=1}^{j} r_\ell > a_{\le j}$, Markov's inequality shows that such a prefix is absent with probability tending to one. On the other hand, if $\sum_{\ell=1}^{j} r_\ell < a_{\le j}$ for $1 \le j \le K$, conditional on the selected paths at level $j-1$, the number of level-j extensions is binomial with an exponentially growing mean of rate $a_{\le j} - \sum_{\ell=1}^{j} r_\ell$. Applying a Chernoff bound successively for $j = 1, \dots, K$ gives the leaf-count exponent $a_{\mathrm{tot}} - \sum_{\ell=1}^{K} r_\ell$. Each such leaf contributes exponential weight $\exp\{\beta d \sum_{\ell=1}^{K} x_\ell + o(d)\}$. Summing over the finitely many bin choices and then shrinking the bins gives (65). Boundary points follow from replacing x by tx and sending $t \uparrow 1$. Gaussian union bounds restrict all edge scores to a fixed bounded interval after division by d, and the second-moment argument in Lemma 7 gives L^1 convergence.

Ignoring the constraints in (65), differentiation gives $x_\ell = \beta v_\ell$. For this point, $\sum_{\ell=1}^{j} \frac{x_\ell^2}{2v_\ell} = \frac{\beta^2 V_j}{2}$, so it satisfies every constraint whenever $0 < \beta \le b_0 := \min_{1 \le j \le K} \sqrt{2a_{\le j}/V_j}$. Because $K \ge 2$ and every $v_\ell > 0$, we have $v_1 < 1$ and hence $a_{\le 1}/V_1 = a_1/v_1 > a_1$. For $j \ge 2$, $a_{\le j} > a_1$ and $V_j \le 1$, so again $a_{\le j}/V_j > a_1$. Therefore $b_0 > c := \sqrt{2a_1}$.

For β in a neighborhood of c, the point $x_\ell = \beta v_\ell$ is feasible and $F_K(\beta) = a_{\mathrm{tot}} + \beta^2/2$. The variational formula is a maximum of affine functions of β, so F_K is convex. At $\beta = c$, we have $F_K(c) = a_{\mathrm{tot}} + a_1$ with $F_K'(c) = c$. Convexity therefore gives, for every $\beta > 0$,

$$F_K(\beta) \ge F_K(c) + F_K'(c)(\beta - c) = a_{\mathrm{tot}} - a_1 + c\beta.$$

After subtracting $a_{\mathrm{tot}} - a_1$ and dividing by β, $h_K^{\mathrm{cnt}}(\beta) \ge c$. At $\beta = c$ equality holds. If equality also held at some $\beta \ne c$, convexity would force F_K to equal the displayed tangent line on the segment joining β and c, contradicting $F_K(t) = a_{\mathrm{tot}} + t^2/2$ near c. Thus c is the unique minimizer.

It remains to control the empirical infimum over all β. Let

$$\overline{X}_d := \frac{1}{|\mathcal{L}_d|}\sum_{u\in\mathcal{L}_d} X_u, \qquad X_{\max,d} := \max_{u\in\mathcal{L}_d} X_u.$$

Each level-ℓ edge appears with the same coefficient in $\overline{X}_d$, and independence gives

$$\mathrm{Var}\left(\frac{\overline{X}_d}{d}\right) = \sum_{\ell=1}^{K} \frac{v_\ell}{d\prod_{j=1}^{\ell} n_{j,d}} \longrightarrow 0.$$

Hence $\overline{X}_d/d \xrightarrow{\mathbb{P}} 0$. Also, since $\frac{1}{\beta d}\log \mathcal{Z}_d^{(K)}(\beta) \in \left[\frac{X_{\max,d}}{d}, \frac{X_{\max,d}}{d} + \frac{\log|\mathcal{L}_d|}{\beta d}\right]$, using (65) and then sending $\beta\to\infty$ shows

$$\frac{X_{\max,d}}{d} \xrightarrow{\mathbb{P}} s_K := \lim_{\beta\to\infty}\frac{F_K(\beta)}{\beta}.$$

This limit is strictly larger than c: choose $c < \beta_0 < b_0$. Since $F_K(t) = a_{\mathrm{tot}} + t^2/2$ near β_0, we have $F_K'(\beta_0) = \beta_0 > c$, and convexity implies $s_K \geq \beta_0$.

Jensen's inequality and the largest-term bound give, for every $\beta > 0$,

$$H_d^{(K)}(\beta) \geq \frac{\log|\mathcal{L}_d| - \log m_d}{\beta d} + \frac{\overline{X}_d}{d}, \qquad H_d^{(K)}(\beta) \geq \frac{X_{\max,d}}{d} - \frac{\log m_d}{\beta d}. \tag{66}$$

The first bound controls all sufficiently small β, because $|\mathcal{L}_d|/m_d = n_{1,d}$ and $d^{-1}\log n_{1,d} \to a_1$. The second controls all sufficiently large β, because $X_{\max,d}/d \xrightarrow{\mathbb{P}} s_K > c$ and $d^{-1}\log m_d \to a_{\mathrm{tot}} - a_1$. On any fixed interval $[t,T] \subset (0,\infty)$, a Gaussian union bound gives $\max_u |X_u|/d = O_{\mathbb{P}}(1)$, so the functions $d^{-1}\log \mathcal{Z}_d^{(K)}(\beta)$ have bounded random Lipschitz constants with high probability. Fixed-grid convergence from (65) then gives uniform convergence on $[t,T]$. Combining this with (66) yields $\inf_{\beta>0} H_d^{(K)}(\beta) \xrightarrow{\mathbb{P}} c$.

Now suppose $X_*/d \xrightarrow{\mathbb{P}} s < c$, and put $\delta := (c-s)/4$. With probability tending to one,

$$H_d^{(K)}(\beta) \geq c - \delta \quad \text{for every } \beta > 0, \qquad \frac{X_*}{d} \leq s + \delta.$$

Choose $t > 0$ small enough that $t(|s|+2) < a_1/4$. For $\beta \geq t$,

$$\log\frac{1 - A_{*,K}^{\mathrm{cnt}}(\beta)}{A_{*,K}^{\mathrm{cnt}}(\beta)} = \beta d\left(H_d^{(K)}(\beta) - \frac{X_*}{d}\right) \geq \frac{td(c-s)}{2}.$$

For $0 < \beta < t$, Jensen's inequality gives

$$\frac{1}{d}\log\frac{1 - A_{*,K}^{\mathrm{cnt}}(\beta)}{A_{*,K}^{\mathrm{cnt}}(\beta)} \geq \frac{\log n_{1,d}}{d} + \beta\left(\frac{\overline{X}_d}{d} - \frac{X_*}{d}\right) \geq \frac{a_1}{4}$$

with probability tending to one. Therefore the log odds diverge uniformly over $\beta > 0$, proving $\sup_{\beta>0} A_{*,K}^{\mathrm{cnt}}(\beta) \xrightarrow{\mathbb{P}} 0$.

Finally, fix a first-level group a. Its averaged score can be written as

$$\bar{X}_a = U_a + R_a,$$

where $U_a \sim N(0, v_1 d)$ is the score on its first edge. At level $\ell \geq 2$, the group average contains $\prod_{j=2}^{\ell} n_{j,d}$ independent edge scores with equal coefficients. Hence $R_a \sim N(0, dw_d)$, independently across groups and independently of $(U_a)_a$, where

$$w_d = \sum_{\ell=2}^{K} \frac{v_\ell}{\prod_{j=2}^{\ell} n_{j,d}} \leq (1-v_1)/n_{2,d} \longrightarrow 0.$$

For every $u > 0$, the union bound gives

$$\mathbb{P}\left\{\max_{1\leq a\leq n_{1,d}} |R_a| > ud\right\} \leq 2n_{1,d}\exp\left(-\frac{du^2}{2w_d}\right) \longrightarrow 0.$$

The log-sum-exp difference between $(\bar{X}_a)_a$ and $(U_a)_a$ is at most $\beta \max_a |R_a|$, so

$$\frac{1}{d}\log \mathcal{Z}_d^{(K),\mathrm{avg}}(\beta) - \frac{1}{d}\log \sum_{a=1}^{n_{1,d}} e^{\beta U_a} \xrightarrow{\mathbb{P}} 0.$$

The remaining sum contains $\exp(a_1 d + o(d))$ independent $N(0, v_1 d)$ scores, and therefore has limiting free energy $f(\beta; a_1, v_1)$. Dividing by β and using (4) gives the minimum $\sqrt{2v_1 a_1}$. □

G Beyond Gaussian scores: a cumulant characterization

The optimized separation extends to a class of non-Gaussian hierarchies. Non-Gaussian GREM free energies and level-dependent rate functions are classical (Jana & Rao, 2007). This extension uses exponential tilting and large deviations (Dembo & Zeitouni, 2010); it does not assert a new large-deviation principle or a new EVaR duality.

Keep the K-level tree, first-level groups, and notation of Appendix F. Replace the score on each level-ℓ edge e by $S_{\ell,e}^{(d)} := \sum_{r=1}^{d} Y_{\ell,e,r}$. All increments are independent, and for each fixed level ℓ they have the same centered, nondegenerate distribution. Assume

$$\Lambda_\ell(t) := \log \mathbb{E} e^{tY_{\ell,e,r}} < \infty \qquad (t \in \mathbb{R}),$$

and $\Lambda'_\ell(t) \to \pm\infty$ as $t \to \pm\infty$. Define

$$I_\ell(x) := \sup_{t\in\mathbb{R}} \{tx - \Lambda_\ell(t)\}, \qquad \Lambda(t) := \sum_{\ell=1}^{K} \Lambda_\ell(t).$$

Also write

$$a_{\le j} := \sum_{\ell=1}^{j} a_\ell, \qquad a_{\mathrm{tot}} := a_{\le K} = \sum_{\ell=1}^{K} a_\ell.$$

and write

$$L_{K,d}(\beta) := \frac{1}{d}\log \mathcal{Z}_d^{(K)}(\beta), \qquad A_{*,K}^{\mathrm{avg}}(\beta) := \frac{e^{\beta X_*}}{e^{\beta X_*} + \mathcal{Z}_d^{(K),\mathrm{avg}}(\beta)}.$$

Theorem 17 (Non-Gaussian free energy and count correction)**.** *The equation*

$$\beta_c \Lambda'(\beta_c) - \Lambda(\beta_c) = a_1 \tag{67}$$

has a unique positive solution. For every fixed $\beta > 0$,

$$L_{K,d}(\beta) \xrightarrow{\mathbb{P}} F_K(\beta) := \max_{\substack{x\in\mathbb{R}^K \\ \sum_{\ell=1}^{j} I_\ell(x_\ell) \le a_{\le j}\ (1\le j\le K)}} \left\{ a_{\mathrm{tot}} - \sum_{\ell=1}^{K} I_\ell(x_\ell) + \beta \sum_{\ell=1}^{K} x_\ell \right\}. \tag{68}$$

The convergence also holds in L^1. Define

$$h_K^{\mathrm{cnt}}(\beta) := \frac{F_K(\beta) - (a_{\mathrm{tot}} - a_1)}{\beta}, \qquad c_{\mathrm{cnt}} := \Lambda'(\beta_c).$$

Then $h_K^{\mathrm{cnt}}(\beta) \ge c_{\mathrm{cnt}}$ and $h_K^{\mathrm{cnt}}(\beta) = c_{\mathrm{cnt}} \iff \beta = \beta_c$. Moreover, $\inf_{\beta>0} H_d^{(K)}(\beta) \xrightarrow{\mathbb{P}} c_{\mathrm{cnt}}$. If $X_*/d \xrightarrow{\mathbb{P}} s < c_{\mathrm{cnt}}$, then $\sup_{\beta>0} A_{*,K}^{\mathrm{cnt}}(\beta) \xrightarrow{\mathbb{P}} 0$, while if $s > c_{\mathrm{cnt}}$, the fixed choice $\beta = \beta_c$ gives $A_{*,K}^{\mathrm{cnt}}(\beta_c) \xrightarrow{\mathbb{P}} 1$.

Proof. Rate functions. For each level ℓ and $t \in \mathbb{R}$, define the tilted law

$$\frac{d\mathbb{P}_{\ell,t}}{d\mathbb{P}}(y) = e^{ty - \Lambda_\ell(t)}.$$

Differentiating the moment generating function gives

$$\Lambda'_\ell(t) = \mathbb{E}_{\ell,t}[Y_{\ell,e,r}], \qquad \Lambda''_\ell(t) = \mathrm{Var}_{\ell,t}(Y_{\ell,e,r}) > 0.$$

The variance is positive because the tilted law has the same support as the original nonconstant distribution. Since the increments are centered, $\Lambda_\ell(0) = \Lambda'_\ell(0) = 0$. Thus Λ'_ℓ is strictly increasing. Because its range is $\mathbb{R}$, for every x there is a unique t_x with $\Lambda'_\ell(t_x) = x$, and $I_\ell(x) = t_x x - \Lambda_\ell(t_x)$. Also, for every $T > 0$,

$$I_\ell(x) \geq T|x| - \max\{\Lambda_\ell(T), \Lambda_\ell(-T)\},$$

so $I_\ell(x)/|x| \to \infty$.

Put $J_\ell(t) := t\Lambda'_\ell(t) - \Lambda_\ell(t)$. For $t > 0$, we have $J'_\ell(t) = t\Lambda''_\ell(t) > 0$. Moreover, for $t \geq 1$,

$$J_\ell(t) - J_\ell(1) = \int_1^t u\Lambda''_\ell(u)\, du \geq \Lambda'_\ell(t) - \Lambda'_\ell(1) \longrightarrow \infty.$$

Hence $\sum_\ell J_\ell(t) = t\Lambda'(t) - \Lambda(t)$ increases continuously from zero to infinity, proving existence and uniqueness of β_c.

Free energy. Let $\bar{Y}_d = d^{-1}\sum_{r=1}^d Y_{\ell,e,r}$ and fix $x \in \mathbb{R}$. Under the product tilted law with parameter t_x, the law of large numbers gives $\bar{Y}_d \to x$. For $\epsilon > 0$ and $B_\epsilon(x) = (x - \epsilon, x + \epsilon)$, change of measure gives

$$e^{-d[I_\ell(x)+|t_x|\epsilon]}\mathbb{P}_{\ell,t_x}\{\bar{Y}_d \in B_\epsilon(x)\} \leq \mathbb{P}\{\bar{Y}_d \in B_\epsilon(x)\} \leq e^{-d[I_\ell(x)-|t_x|\epsilon]}.$$

Therefore the probability that a level-ℓ edge has normalized score near x has exponential rate $-I_\ell(x)$.

Now partition a bounded interval into finitely many bins. If the first j normalized edge scores lie near $x_1, \ldots, x_j$, the expected number of such depth-j paths has exponential rate $a_{\leq j} - \sum_{\ell=1}^j I_\ell(x_\ell)$. If this number is negative, Markov's inequality shows that no such depth-j partial path exists with probability tending to one. If the corresponding rate is positive at every depth, then conditional on the paths at level $j-1$, the number of level-j extensions is binomial with exponentially growing mean. Applying a Chernoff bound successively over the K levels gives the leaf-count rate $a_{\mathrm{tot}} - \sum_{\ell=1}^K I_\ell(x_\ell)$. Multiplying this count by the leaf weight $\exp\{\beta d\sum_\ell x_\ell + o(d)\}$ and maximizing over feasible x gives (68). Boundary points follow by replacing x with ux and sending $u \uparrow 1$, since $I_\ell(ux) \leq uI_\ell(x)$.

It remains only to remove the bounded-score restriction. For fixed $t, R > 0$, a union bound over all level-ℓ edges gives

$$\mathbb{P}\left\{\max_e |S^{(d)}_{\ell,e}| > dR\right\} \leq 2\Big(\prod_{j=1}^{\ell} n_{j,d}\Big)\exp\{d[\max(\Lambda_\ell(t), \Lambda_\ell(-t)) - tR]\}.$$

Choosing R large makes this tend to zero for every level. The same bound, integrated in R, gives a uniform second-moment bound for the largest leaf score divided by d, and hence the convergence also holds in L^1.

Count threshold. Without the path-existence constraints, each coordinate maximizes $\beta x - I_\ell(x)$ at $x_\ell = \Lambda'_\ell(\beta)$, with value $\Lambda_\ell(\beta)$. Its cost is $I_\ell(\Lambda'_\ell(\beta)) = \beta\Lambda'_\ell(\beta) - \Lambda_\ell(\beta)$. At $\beta = \beta_c$, the total cost is $\sum_\ell J_\ell(\beta_c) = a_1$. The level-1 cost is strictly smaller than a_1 because $J_\ell(\beta_c) > 0$ for $\ell \geq 2$, while for $j \geq 2$,

$$\sum_{\ell=1}^{j} J_\ell(\beta_c) \leq a_1 < a_{\leq j}.$$

Thus all constraints are strict at β_c and remain strict nearby. On this neighborhood,

$$F_K(\beta) = a_{\mathrm{tot}} + \sum_{\ell=1}^K \max_x\{\beta x - I_\ell(x)\} = a_{\mathrm{tot}} + \Lambda(\beta), \qquad F'_K(\beta) = \Lambda'(\beta).$$

Since F_K is a maximum of affine functions of β, it is convex. Its tangent at β_c gives

$$F_K(\beta) \geq F_K(\beta_c) + F'_K(\beta_c)(\beta - \beta_c) = a_{\mathrm{tot}} - a_1 + \beta\Lambda'(\beta_c),$$

where (67) gives the second equation. Therefore $h^{\mathrm{cnt}}_K(\beta) \geq \Lambda'(\beta_c) = c_{\mathrm{cnt}}$. Equality holds at β_c. Equality at another point would make F_K affine on the segment joining that point to β_c, contradicting $F''_K(\beta) = \Lambda''(\beta) > 0$ near β_c.

All temperatures. Let

$$\overline{X}_d := \frac{1}{|\mathcal{L}_d|} \sum_{u \in \mathcal{L}_d} X_u, \qquad X_{\max,d} := \max_{u \in \mathcal{L}_d} X_u.$$

Each level-ℓ edge appears with the same coefficient in $\overline{X}_d$, so

$$\mathrm{Var}\left(\frac{\overline{X}_d}{d}\right) = \frac{1}{d} \sum_{\ell=1}^{K} \frac{\Lambda_\ell''(0)}{\prod_{j=1}^{\ell} n_{j,d}} \longrightarrow 0.$$

Hence $\overline{X}_d/d \xrightarrow{\mathbb{P}} 0$. Also,

$$\frac{X_{\max,d}}{d} \le \frac{L_{K,d}(\beta)}{\beta} \le \frac{X_{\max,d}}{d} + \frac{\log|\mathcal{L}_d|}{\beta d}.$$

Using $L_{K,d}(\beta) \xrightarrow{\mathbb{P}} F_K(\beta)$ and then sending $\beta \to \infty$ gives

$$\frac{X_{\max,d}}{d} \xrightarrow{\mathbb{P}} v := \lim_{\beta \to \infty} \frac{F_K(\beta)}{\beta}.$$

Moreover $v > c_{\mathrm{cnt}}$: choose $\beta_0 > \beta_c$ in the neighborhood where $F_K = a_{\mathrm{tot}} + \Lambda$. Then

$$F_K'(\beta_0) = \Lambda'(\beta_0) > \Lambda'(\beta_c) = c_{\mathrm{cnt}},$$

and convexity implies $v \ge F_K'(\beta_0)$.

Jensen's inequality and the largest-term bound give, for every $\beta > 0$,

$$H_d^{(K)}(\beta) \ge \frac{\log n_{1,d}}{\beta d} + \frac{\overline{X}_d}{d}, \qquad H_d^{(K)}(\beta) \ge \frac{X_{\max,d}}{d} - \frac{\log m_d}{\beta d}.$$

The first inequality keeps the minimum away from $\beta = 0$, and the second keeps it away from $\beta = \infty$, because $d^{-1}\log n_{1,d} \to a_1$, $d^{-1}\log m_d \to a_{\mathrm{tot}} - a_1$, and $v > c_{\mathrm{cnt}}$. On every fixed interval $[t, T] \subset (0, \infty)$, the derivative of $L_{K,d}$ is bounded in absolute value by $\max_u |X_u|/d = O_{\mathbb{P}}(1)$. Fixed-grid convergence therefore gives uniform convergence on $[t, T]$. Combining the three temperature ranges and evaluating at β_c yields

$$\inf_{\beta > 0} H_d^{(K)}(\beta) \xrightarrow{\mathbb{P}} c_{\mathrm{cnt}}.$$

Finally suppose $X_*/d \xrightarrow{\mathbb{P}} s < c_{\mathrm{cnt}}$ and put $\delta = (c_{\mathrm{cnt}} - s)/4$. With probability tending to one,

$$H_d^{(K)}(\beta) \ge c_{\mathrm{cnt}} - \delta \quad \text{for every } \beta > 0, \qquad \frac{X_*}{d} \le s + \delta.$$

Choose $t > 0$ small enough that $t(|s| + 2) < a_1/2$. For $\beta \ge t$,

$$\log \frac{1 - A_{*,K}^{\mathrm{cnt}}(\beta)}{A_{*,K}^{\mathrm{cnt}}(\beta)} = \beta d \left(H_d^{(K)}(\beta) - \frac{X_*}{d} \right) \ge \frac{td(c_{\mathrm{cnt}} - s)}{2}.$$

For $0 < \beta < t$, Jensen's inequality gives

$$\frac{1}{d} \log \frac{1 - A_{*,K}^{\mathrm{cnt}}(\beta)}{A_{*,K}^{\mathrm{cnt}}(\beta)} \ge \frac{\log n_{1,d}}{d} + \beta \left(\frac{\overline{X}_d}{d} - \frac{X_*}{d} \right),$$

whose right-hand side is bounded away from zero with probability tending to one. Thus the log odds diverge uniformly over all $\beta > 0$, proving

$$\sup_{\beta > 0} A_{*,K}^{\mathrm{cnt}}(\beta) \xrightarrow{\mathbb{P}} 0.$$

If $s > c_{\mathrm{cnt}}$, then at $\beta = \beta_c$,

$$\frac{1}{\beta_c d} \log \frac{1 - A_{*,K}^{\mathrm{cnt}}(\beta_c)}{A_{*,K}^{\mathrm{cnt}}(\beta_c)} \xrightarrow{\mathbb{P}} c_{\mathrm{cnt}} - s < 0,$$

so $A_{*,K}^{\mathrm{cnt}}(\beta_c) \xrightarrow{\mathbb{P}} 1$. □

Theorem 18 (Averaging beyond Gaussian scores)**.** *The equation*

$$\beta_r \Lambda_1'(\beta_r) - \Lambda_1(\beta_r) = a_1 \tag{69}$$

has a unique positive solution. Define

$$F_{\mathrm{avg}}(\beta) := \max_{I_1(x) \le a_1} \{a_1 - I_1(x) + \beta x\}, \qquad c_{\mathrm{avg}} := \Lambda_1'(\beta_r).$$

Then, for every fixed $\beta > 0$,

$$\frac{1}{d} \log \mathcal{Z}_d^{(K),\mathrm{avg}}(\beta) \xrightarrow{\mathbb{P}} F_{\mathrm{avg}}(\beta).$$

The threshold $F_{\mathrm{avg}}(\beta)/\beta$ has minimum c_{avg} and equals c_{avg} for every $\beta \ge \beta_r$. Moreover, $c_{\mathrm{avg}} < c_{\mathrm{cnt}}$. Hence for every $s \in (c_{\mathrm{avg}}, c_{\mathrm{cnt}})$, averaging retrieves at every fixed $\beta \ge \beta_r$, whereas count correction fails for every temperature choice. The groups must still be known and the planted item must still be a singleton.

Proof. Put $J_1(t) := t\Lambda_1'(t) - \Lambda_1(t)$. Since $J_1'(t) = t\Lambda_1''(t) > 0$, $J_1(0) = 0$, and $J_1(t) \to \infty$, equation (69) has a unique positive solution.

For group a, write $\bar{X}_a = S_{1,a}^{(d)} + R_a$, where R_a is the averaged contribution from levels $2, \dots, K$. Put $q_{\ell,d} = \prod_{j=2}^{\ell} n_{j,d}$ for $\ell \ge 2$; then R_a is the sum, over these levels, of averages of $q_{\ell,d}$ independent edge scores. Hence

$$\frac{1}{d} \log \mathbb{E} e^{tR_a} = \sum_{\ell=2}^{K} q_{\ell,d} \Lambda_\ell(t/q_{\ell,d}) \longrightarrow 0$$

for every fixed real t, since $\Lambda_\ell(z) = \Lambda_\ell''(0) z^2/2 + o(z^2)$ at zero. For $u > 0$, choose a fixed $t > a_1/u$. A union bound and exponential Markov inequalities at t and $-t$ give

$$\mathbb{P}\{\max_a |R_a| > du\} \le 2n_{1,d} \exp\{-dtu + o(d)\} \longrightarrow 0.$$

Averaging therefore leaves the one-level root free energy F_{avg} defined in the theorem. For positive β, the optimizer is $\Lambda_1'(\beta)$ until β_r and the rightmost feasible point $c_{\mathrm{avg}} = \Lambda_1'(\beta_r)$ thereafter. Thus the threshold decreases to c_{avg} and equals it for $\beta \ge \beta_r$. In particular

$$c_{\mathrm{avg}} = \min_{\beta > 0} \frac{a_1 + \Lambda_1(\beta)}{\beta}.$$

At the count optimum,

$$c_{\mathrm{cnt}} = \frac{a_1 + \Lambda_1(\beta_c)}{\beta_c} + \frac{\sum_{\ell=2}^{K} \Lambda_\ell(\beta_c)}{\beta_c} > c_{\mathrm{avg}},$$

because the second summand is strictly positive. This proves the gap and all retrieval assertions. □

Concrete non-Gaussian example. Let $Y_\ell = b_\ell(P_\ell^+ - P_\ell^-)$ for independent Poisson variables of mean $\lambda_\ell/2$, with $b_\ell, \lambda_\ell > 0$. Then

$$\Lambda_\ell(t) = \lambda_\ell(\cosh(b_\ell t) - 1), \qquad I_\ell(x) = \frac{x}{b_\ell} \operatorname{arsinh} \frac{x}{\lambda_\ell b_\ell} - \sqrt{(x/b_\ell)^2 + \lambda_\ell^2} + \lambda_\ell.$$

These distributions meet every hypothesis and allow direct numerical optimization of (68). For Gaussian increments $\Lambda_\ell(t) = v_\ell t^2/2$ with $\sum_\ell v_\ell = 1$, Theorems 17 and 18 recover exactly $\beta_c = c_{\mathrm{cnt}} = \sqrt{2a_1}$ and $c_{\mathrm{avg}} = \sqrt{2v_1 a_1}$. The dependence on the leaf marginal cumulant, rather than its variance alone, matters beyond Gaussian scores. Heavy-tailed increments without exponential moments and bounded-support saturation regimes are not covered by these theorems.

H Extensions and boundary cases

H.1 Partial correction and a second optimization boundary

Fix $0 < \gamma \le 1$. Partial count correction divides every distractor weight by m_d^γ, so $\gamma = 1$ is full count correction and smaller γ leaves a factor $m_d^{1-\gamma}$ of the within-cluster multiplicity. Its distractor denominator and limiting threshold are

$$D_{\gamma,d}(\beta) := \frac{\mathcal{Z}_d(\beta)}{m_d^\gamma}, \qquad h_\gamma(\beta) := \frac{F(\beta) - \gamma\alpha_2}{\beta}.$$

Put $c_\gamma := \alpha_1 + (1-\gamma)\alpha_2$, the total distractor growth rate left after the partial correction.

Proposition 19 (Optimal partial correction)**.** *Assume $0 < \rho < 1$. If $\rho c_\gamma \le \alpha_1$, then*

$$\beta_\gamma = \sqrt{2c_\gamma}, \qquad \min_{\beta>0} h_\gamma(\beta) = \sqrt{2c_\gamma}. \tag{70}$$

If $\rho c_\gamma > \alpha_1$, then

$$\beta_\gamma = \sqrt{\frac{2(1-\gamma)\alpha_2}{1-\rho}}, \qquad \min_{\beta>0} h_\gamma(\beta) = \sqrt{2\rho\alpha_1} + \sqrt{2(1-\rho)(1-\gamma)\alpha_2}. \tag{71}$$

The formulas agree at equality. The minimizer is unique for $\gamma > 0$.

Proof. In the merged regime the first branch is $c_\gamma/\beta + \beta/2$, whose optimum lies strictly before $\sqrt{2\alpha}$. The last branch is $\sqrt{2\alpha} - \gamma\alpha_2/\beta$, strictly increasing. Also $\rho \le \alpha_1/\alpha$ implies $\rho c_\gamma < \alpha_1$. In the split regime, the first branch has the same expression. Its optimum lies at or before b_1 precisely when $\rho c_\gamma \le \alpha_1$. The middle branch is

$$\sqrt{2\rho\alpha_1} + \frac{(1-\gamma)\alpha_2}{\beta} + \frac{\beta(1-\rho)}{2}.$$

Under the reverse strict inequality its optimum in (71) lies strictly between b_1 and b_2. Otherwise its derivative is nonnegative from b_1 onward. The last branch $v - \gamma\alpha_2/\beta$ is strictly increasing. The derivative signs and continuity prove the formulas and uniqueness. The case $\gamma = 1$ always satisfies (70). □

H.2 Unequal multiplicities and a relevant group

Fix a finite number T of distractor types. Type t has

$$M_{t,d} := \lceil e^{a_{1t}d} \rceil \quad \text{clusters}, \qquad m_{t,d} := \lceil e^{a_{2t}d} \rceil \quad \text{tokens per cluster},$$

correlation $\rho_t \in (0,1)$, and scores $\mu_t d + \sigma_t X_{aj}^{(t)}$ with $\sigma_t > 0$ where $X^{(t)}$ follows (2) with rates (a_{1t}, a_{2t}) and correlation ρ_t. Let F_t denote its free energy from Theorem 1, and let $\mathcal{Z}_{t,d}$ be its partition function including the shift and scale above. Applying exponent-γ count correction gives the total distractor denominator

$$D_d^{\text{types}}(\beta) := \sum_{t=1}^{T} \frac{\mathcal{Z}_{t,d}(\beta)}{m_{t,d}^\gamma}.$$

Suppose the relevant group has r_d identical copies of score X_*, where $d^{-1}\log r_d \to a_* \ge 0$ and $X_*/d \xrightarrow{\mathbb{P}} s$. Its corrected numerator and total attention mass are

$$r_d^{1-\gamma} e^{\beta X_*}, \qquad A_{\text{rel}} := \frac{r_d^{1-\gamma} e^{\beta X_*}}{r_d^{1-\gamma} e^{\beta X_*} + D_d^{\text{types}}(\beta)}.$$

Proposition 20 (Finitely many distractor types)**.** *Under the setup above,*

$$\frac{1}{d}\log\frac{1-A_{\text{rel}}}{A_{\text{rel}}} \xrightarrow{\mathbb{P}} \max_{1\le t\le T}\{\beta\mu_t + F_t(\beta\sigma_t) - \gamma a_{2t}\} - \beta s - (1-\gamma)a_*.$$

A strictly negative limit implies success and a strictly positive limit implies failure. Cross-type independence is unnecessary if each type has the stated marginal model.

Proof. Each corrected type partition has the asserted exponent by Theorem 1. The logarithm of a sum of T positive numbers differs from their maximum log by at most $\log T = o(d)$. A union bound gives simultaneous component convergence. The relevant numerator is $r_d^{1-\gamma} e^{\beta X_*}$. Subtract its logarithm and use the logistic identity. □

Full correction removes the relevant group's multiplicity advantage as well as distractor multiplicities. Individual relevant-token mass is A_{rel}/r_d, so semantic-group retrieval and retrieval of one specific token are different objectives.

H.3 Exact and approximate replication

Consider a scalar count rule that multiplies every item in a group of size m by the same factor $w(m) > 0$. A group with scores $x_1, \ldots, x_m$ then contributes

$$C_w(x_1, \ldots, x_m) := w(m) \sum_{j=1}^{m} e^{\beta x_j}.$$

Call the rule replication invariant if replacing every score by k identical copies leaves this contribution unchanged for every integer $k \geq 1$.

Proposition 21 (Uniqueness within scalar count rules)**.** *Replication invariance holds if and only if $w(m) = w(1)/m$. With this rule, replicating an entire score list leaves $m^{-1} \sum_j e^{\beta x_j}$ unchanged, whereas replicating only selected nonidentical scores need not preserve it.*

Proof. Replicate a singleton m times to get $mw(m) = w(1)$; the converse is immediate. Replication of a whole list multiplies its exponential sum and its size by the same factor. Selective replication changes the empirical distribution of scores. □

For approximate replication, let $\bar{x} = m^{-1} \sum_j x_j$. If $|x_j - \bar{x}| \leq u$ for all group members, Jensen's inequality and the termwise upper bound give

$$0 \leq \log \left(m^{-1} \sum_j e^{\beta x_j} \right) - \beta \bar{x} \leq \beta u.$$

Thus uniformly $o(d)$ score spread makes count correction and averaging equivalent at the exponential scale. With fixed $\rho < 1$ and exponentially many children, that uniform spread need not be $o(d)$.

H.4 No universal effective context length

An i.i.d. Gaussian context with $\exp(a_{\mathrm{eff}} d + o(d))$ distractors has limiting free energy $f(\beta; a_{\mathrm{eff}}, 1)$. A temperature-independent effective context length would therefore be a single rate a_{eff} satisfying

$$F(\beta) = f(\beta; a_{\mathrm{eff}}, 1) \qquad \text{for every } \beta > 0.$$

The question is whether the correlated two-level model can be replaced by such an i.i.d. context at all temperatures.

Corollary 22. *If $\rho > \alpha_1/\alpha$, no such rate a_{eff} exists.*

Proof. Sending $\beta \downarrow 0$ forces $a_{\mathrm{eff}} = \alpha$. But $b_1 < \min(b_2, \sqrt{2\alpha})$, and on the nonempty interval between them,

$$f(\beta; \alpha, 1) - F(\beta) = \frac{\rho}{2}(\beta - b_1)^2 > 0.$$

This contradicts equality. The strict split condition is essential; in the merged regime the iid free energy holds. □

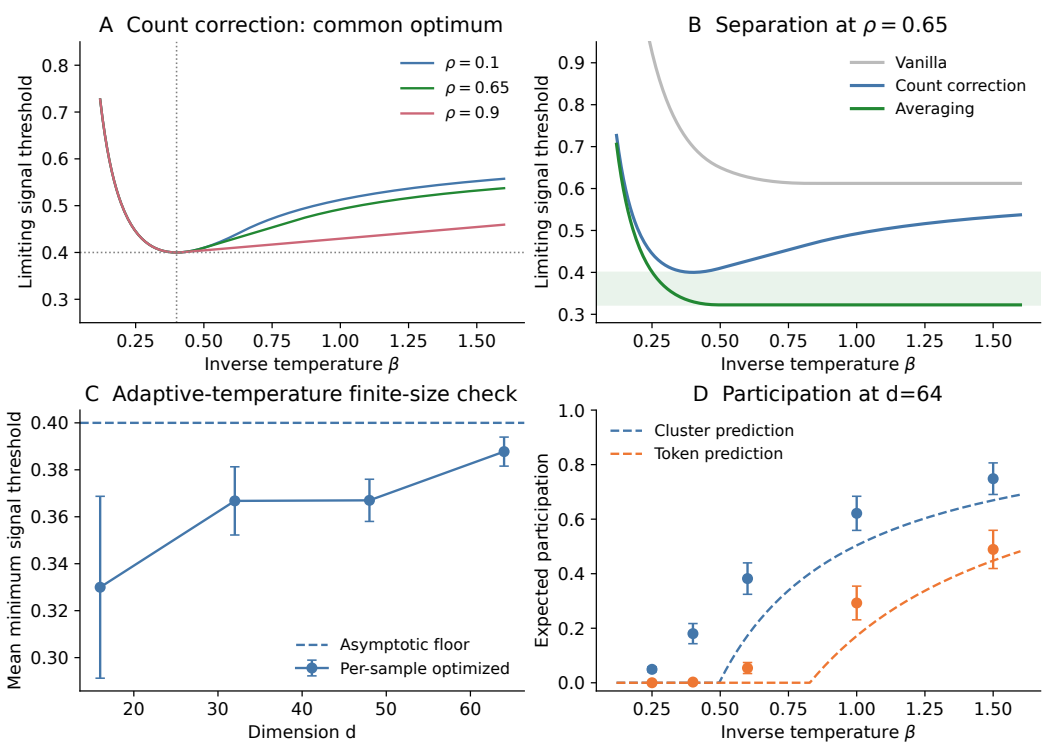


Figure 1: Analytical retrieval thresholds and finite-size Gaussian checks. Dashed lines in the simulation panels are infinite-size predictions; markers are empirical means with 1.96 standard errors. These are synthetic model checks.

H.5 Endpoint and scaling checks

At $\rho = 0$, all scores are i.i.d. and $F = f(\beta; \alpha, 1)$. Count correction still has unique optimum $\sqrt{2\alpha_1}$. Averaging has limiting free energy α_1, hence threshold α_1/β and infimum zero as $\beta \to \infty$. At $\rho = 1$,

$$\mathcal{Z}_d = m_d \sum_a e^{\beta\sqrt{d}G_a}, \qquad F = \alpha_2 + f(\beta; \alpha_1, 1).$$

Count correction and averaging coincide exactly. Their minimum is $\sqrt{2\alpha_1}$, attained by every $\beta \geq \sqrt{2\alpha_1}$, so uniqueness in Theorem 3 excludes this endpoint.

At fixed β, a correction with uniformly $o(d)$ log weight cannot change an off-boundary threshold. Fixed or subexponential multiplicities therefore cannot generate the order-d correction studied here. The two-level model assumes $\alpha_2 > 0$ rather than a fixed number of duplicates.

For conventional scaled dot-product pre-softmax scores $t_d X/\sqrt{d}$, a finite-size plug-in choice is

$$t_{\mathrm{opt},d} = \sqrt{2 \log M_d}, \qquad \beta_{\mathrm{opt},d} = \sqrt{2 \log M_d / d} \longrightarrow \sqrt{2\alpha_1}.$$

Compact uniform convergence justifies this plug-in choice. The scale depends on marginal Gaussian variance and on supplied groups; it is not a universal scaling law for bounded or normalized attention models.

I Numerical protocol and reproducibility

Figure 1 retains the four Gaussian diagnostics. Panel A shows the common count-correction optimum as correlation changes. Panel B compares the required signals for ordinary attention, count correction, and averaging. Panel C optimizes the temperature for each finite array and compares

the required signal to the asymptotic limit. Panel D distinguishes cluster participation from token participation.

The scripts use NumPy, SciPy, and Matplotlib on CPU. Script `scripts/verify_theory.py` uses seed 20260910 and 120 parameter configurations: α_1, α_2 log-uniform on $[e^{-3}, e^{0.5}]$, ρ uniform on $[0.015, 0.985]$, and β log-uniform on $[e^{-3}, e^{2}]$. SLSQP optimizes the variational problem from two starting points, independently of its closed-form case distinction. Scalar optimization checks the count and partial-count optima. Additional checks cover covariance monotonicity, Jensen comparisons, replication, bounded count errors, and the finite-sample interpolation identity. The last check has Monte Carlo sampling error and is not symbolic verification.

Script `scripts/simulate.py` uses seed 271828 with $(\alpha_1, \alpha_2, \rho) = (0.08, 0.12, 0.65)$, $d \in \{16, 32, 48, 64\}$, and 64 independent arrays at each dimension. It evaluates $\beta \in \{0.25, 0.4, 0.6, 1, 1.5\}$ on each shared array. The largest array has $168 \times 2165 = 363720$ distractor scores. Each per-array optimum is searched on $[0.025, 4]$; this bounded search is not the all-temperature proof. Plot error bars are 1.96 times the sample standard error, interpreted as descriptive normal approximations. Raw trials, optimized values, seeds, and runtime metadata are saved under `results/`.

Script `scripts/verify_extensions.py` uses seed 314159 to compare entropy-root calibration with scalar optimization and a separately constrained 20-dimensional entropy optimization on 20 arrays. It also checks the high-temperature tangent bound by directly optimizing the prefix-constrained variational problem for 40 hierarchies of depths three through five. The maximum entropy-duality discrepancy was about 5.1×10^{-6}; numerical tolerances and raw discrepancies are recorded, and solver success is not a proof.

Script `scripts/verify_cramer.py` uses seed 161803 and 30 hierarchies of depths two through five, with independent level-specific Poisson rates and jump sizes. It compares the cumulant optimum with direct prefix-constrained optimization and checks the averaged root floor through a separate rate-function inversion. Full parameters and discrepancies are saved in `results/cramer_verification.json`.

Agreement with a Gaussian simulator checks the stipulated model's implementation, not its validity for trained heads. The separate frozen-sentence experiment is described in Appendix J; no training experiment or implemented LUCID comparison is claimed.

J Frozen-encoder experiment: protocol and additional results

Data and representation. BANKING77 contains 10,003 training and 3,080 test sentences across 77 intents (Casanueva et al., 2020). We process the official training split first and test split second, discarding repeated texts after lowercasing and collapsing whitespace, including test texts already encountered in training. This removes four training and eight test rows. With seed 20260913, we select 32 training sentences per intent as a key pool, then four validation and 12 evaluation queries per intent from the test split. Validation and evaluation share the intent vocabulary and key pool, but no normalized query text. We retain all selected queries regardless of retrieval outcome. The resulting scores are not an official BANKING77 benchmark result. Dataset labels construct the controlled contexts and evaluate outputs; the retrieval functions receive scores, groups, and values, without the query label.

The encoder is `sentence-transformers/all-MiniLM-L6-v2`, revision `1110a243fdf4706b3f48f1d95db1a4f5529b4d41`, using its official float32 ONNX export on CPU. Following the model card (Sentence Transformers, 2021), we apply attention-mask mean pooling and normalize the resulting 384-dimensional sentence vectors. The maximum input length is 256 tokens. The 3,696 selected sentences are encoded in batches of 32; no parameters are trained. Cosine scores are computed from these vectors. A group score is the arithmetic mean of its member scores, equivalently the dot product with the mean of the normalized key vectors, without normalizing that mean again. BANKING77 is distributed under CC BY 4.0 and the encoder under Apache 2.0. Model pretraining overlap has not been audited. Revision identifiers, download URLs, and SHA256 hashes are recorded in `results/learned/assets.json`.

Paired contexts. For each seed in $\{11, 23, 47\}$, we randomly order each intent's 32 keys. For each target intent, we choose 38 of the other 76 intents to form larger distractor groups; the remaining 38 are singleton distractors. Queries of the same intent share these draws. The six conditions use the same ordering: one key per intent; eight copies of the first key in the selected distractor groups; four distinct keys; eight distinct keys; eight distractor keys with four relevant keys; and a mixed relevant group. In the last condition, three keys are moved from the first selected distractor group into the relevant singleton's group. This preserves the token set and leaves five keys in the donor group. Context sizes are respectively 77, 343, 191, 343, 346, and 343 keys. This construction deliberately controls whether relevant keys are averaged; it does not estimate a partition or sample naturally occurring context frequencies.

Methods and values. With cosine score c_i, multiplier t, and supplied group size m_i, ordinary and corrected token weights are proportional to e^{tc_i} and e^{tc_i}/m_i. Averaging assigns weight $e^{t\bar{c}_g}$ and value $\bar{v}_g$ to group g. Individual values are the 77-dimensional one-hot intent vectors. Thus the output's relevant coordinate sums the attention on relevant values; with several relevant keys it is not a single-token mass. A fourth baseline selects one member uniformly per group and applies softmax with equal group priors. Its selection uses the context seed, not query scores. In the mixed condition it may discard the sole relevant key, yielding zero relevant mass and infinite negative log mass. It therefore has no finite validation objective there; its deterministic smallest-grid fallback is archived but excluded from Table 2.

Temperature and uncertainty. Each method and condition selects one t from 81 logarithmically spaced values between 0.25 and 1024 by minimizing mean validation negative log relevant value mass over the 308 queries and three seeds. The same chosen value is applied to all evaluation queries and context seeds. All three primary methods choose an interior grid point. The multiplier t acts on cosine scores; we do not identify it with the Gaussian β or fit the asymptotic square-root thresholds to these bounded scores. Table 2 reports top-1 accuracy of the output intent at the same selected multiplier. Figure 2 shows evaluation curves descriptively, without using them to select temperatures. Bootstrap intervals use 2,000 paired resamples of the 77 intent blocks (seed 8831), after averaging the three context seeds within each query. They describe variation across these intents conditional on this encoder, corpus, and key draws; the 2,772 query-context pairs are not treated as independent queries.

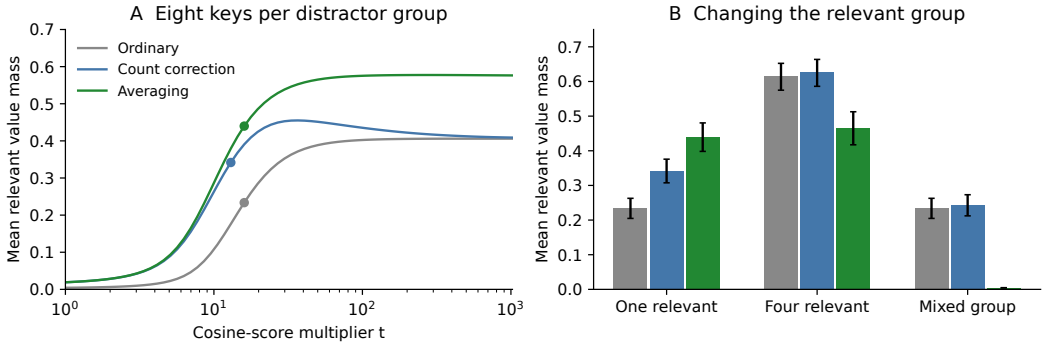


Figure 2: Frozen sentence scores. A: evaluation temperature curves for eight-key distractor groups; dots mark independently validation-selected temperatures. B: relevant value mass after changing the relevant group, with 95% intent-block bootstrap intervals. The reversal for four relevant keys occurs even though all four carry the correct intent label.

An all-temperature oracle. For singleton-target conditions we additionally give count correction the true relevant position and optimize its relevant mass separately for each evaluation query. With distractor scores c_i and target score c_*, minimize the convex log-odds function $\log \sum_{i \neq *} \exp[t(c_i - c_*)]/m_i$ over $t \geq 0$. Its derivative is monotone, so bracketing and scalar root finding locate an interior minimum; the implementation also handles the limits at zero and infinity. This objective maximizes mass for a known target and differs from minimizing the signal threshold in Proposition 15. With eight-key distractor groups, oracle count correction has mean mass 0.511,

above calibrated averaging's 0.440. Only seven of the 2,772 evaluation query-context pairs have averaging mass above 0.9 while the count oracle cannot reach 0.9. Moreover, the oracle optimum is at infinity for 40.7% of these pairs. Finite optimal calibration across queries must therefore not be confused with the theorem's unique finite threshold optimum. The experiment does not establish an asymptotic or uniform all-temperature lower bound on learned scores.

Failures and checks. With four relevant keys, averaging loses 0.161 mean relevant mass against count correction, with paired 95% interval $[-0.181, -0.142]$. High-scoring individual matches can be useful on the relevant side even with correct intent groups. The mixed condition also changes the value: only one quarter of the averaged target-group value is relevant, so its relevant mass cannot exceed $1/4$. The three contaminating keys share one other intent; its output coordinate is always at least three times the target's, making zero top-1 accuracy structural in this stress test. We do not interpret it as a separate empirical discovery. For singleton-target contexts, Jensen ordering holds at every evaluated temperature; exact-copy invariance and equality of all methods on all-singleton contexts hold to 1.1×10^{-13}. Independent small-array tests compare mixed values with direct softmax sums and check interior and endpoint oracle solutions against dense temperature search.

Table 2: Temperature t, mean relevant value mass V, and top-1 output-intent accuracy (%) at that temperature. Temperatures minimize validation negative log relevant value mass, not top-1 error.

Context	Method	t	V	Accuracy (%)
One key per intent	Ordinary	14.42	0.379	51.8
One key per intent	Count correction	14.42	0.379	51.8
One key per intent	Averaging	14.42	0.379	51.8
One key per intent	One representative	14.42	0.379	51.8
Exact copies (8)	Ordinary	17.75	0.288	36.3
Exact copies (8)	Count correction	14.42	0.379	51.8
Exact copies (8)	Averaging	14.42	0.379	51.8
Exact copies (8)	One representative	14.42	0.379	51.8
Same intent (4)	Ordinary	14.42	0.267	39.4
Same intent (4)	Count correction	13.00	0.345	53.5
Same intent (4)	Averaging	14.42	0.406	56.7
Same intent (4)	One representative	14.42	0.381	53.3
Same intent (8)	Ordinary	16.00	0.234	33.1
Same intent (8)	Count correction	13.00	0.342	53.2
Same intent (8)	Averaging	16.00	0.440	57.6
Same intent (8)	One representative	14.42	0.375	51.6
Four relevant keys	Ordinary	21.86	0.614	72.1
Four relevant keys	Count correction	21.86	0.626	72.8
Four relevant keys	Averaging	19.70	0.465	59.1
Four relevant keys	One representative	14.42	0.366	50.3
Mixed relevant group	Ordinary	16.00	0.234	33.1
Mixed relevant group	Count correction	16.00	0.243	34.2
Mixed relevant group	Averaging	4.59	0.005	0.0

Reproduction. The supplementary material contains `requirements-experiments.txt` and the scripts `download_learned_assets.py`, `learned_retrieval.py`, and `report_learned_retrieval.py`. The first script downloads only the pinned public assets. Running the three scripts regenerates exact row IDs and texts, frozen scores, every context, all temperature curves, per-query values, selected temperatures, uncertainty intervals, and package versions under `results/learned/`. Generated outputs, downloaded model weights, and raw source data are not bundled. The protocol was written before inspecting retrieval results; the oracle rescue fraction is an additional descriptive analysis. This single-encoder, single-dataset study neither isolates correlation at fixed marginal score distributions nor tests internal attention heads, context-length scaling, or unknown-group recovery.

Table 3: Frozen DistilBERT layer-6 attention intervention. Means average three context seeds and two reversed orders within 80 queries; brackets are paired 95% bootstrap intervals. Δ is averaging minus count correction.

Condition	$\tau_{\mathrm{cnt}}/\tau_{\mathrm{avg}}$	Δ relevant mass	Δ candidate prob.
All singleton	$1/1$	.00000	.00000
Repeated facts (4)	$1/1$	.00032 [.00030,.00035]	-.00215 [-.00253,-.00181]
Related (4), selected	$2/1$	-.00139 [-.00291,.00001]	-.02042 [-.03285,-.00946]
Related (4), native	$1/1$	.00018 [.00017,.00020]	-.00159 [-.00186,-.00133]
Four relevant	$1/1$	-.00083 [-.00093,-.00074]	.00054 [.00014,.00098]

K FROZEN-TRANSFORMER INTERNAL-ATTENTION DIAGNOSTIC

This diagnostic asks whether count correction and averaging remain mechanically distinct inside an actual pretrained transformer. It is deliberately secondary to the frozen sentence-retrieval study in Section 8: the intervention is causal, but the model is weak on the controlled task and the downstream results do not support a general advantage for averaging.

Model and task. We freeze the masked-language model `distilbert-base-uncased` (Sanh et al., 2019) at revision `12040accade4e8a0f71eabdb258fecc2e7e948be`. Synthetic prompts define an eight-color associative retrieval task. Marked answer positions form the supplied groups and every other active token is a singleton. The locked evaluation contains 80 query specifications, three context seeds, and paired forward and reversed group orders, giving 1,920 prompt instances. The all-singleton control, four repeated exact facts, four distinct related facts, and four relevant answer positions test identity, singleton, and multiple-relevant regimes.

Internal intervention. At the masked-token query row, we replace the actual softmax computation in all 12 heads of validation-selected layer 5, the final DistilBERT layer. Ordinary attention applies softmax to the token logits. Count correction subtracts the log supplied group size from each member logit. Averaging applies softmax to group-mean logits and combines group-mean projected values. Every other model operation and parameter remains unchanged. Layer and temperature choices were fixed from 24 validation query specifications before the 80 evaluation specifications were generated and hash-locked. Intervals below use 5,000 paired bootstrap resamples of query specifications after averaging context seeds and the two orders.

Mechanistic result. At the native temperature, averaging raises singleton relevant attention mass above count correction for every prompt in both repeated and related groups. The mean gains are small but resolved: 3.21×10^{-4} for repeated facts and 1.82×10^{-4} for related facts. Related groups also have larger within-group extremes: the mean maximum-minus-mean logit is 0.443, compared with 0.391 for repeated facts, while the corresponding log-mean-exp-minus-mean gaps are 0.0811 and 0.0676. Thus the score variation targeted by averaging is present in actual self-attention logits.

Repeated textual facts are not identical score–value copies inside the transformer. Absolute positions and prior contextualization give their marked tokens different logits and projected values. Consequently, the exact-copy equality from the abstract finite-array model applies only to an artificial control with identical logits and values; that control passes to floating-point tolerance, but repetition in a transformer need not satisfy it.

Downstream result and scope. Increased relevant attention mass does not improve the frozen model's masked-token prediction. At the common native temperature, averaging lowers the correct candidate probability relative to count correction for both repeated and related singleton conditions. After validation-only calibration, related-group count correction selects temperature 2 while averaging selects 1, and averaging is lower by 0.0204 with paired 95% interval $[-0.0329, -0.00946]$. Candidate top-1 accuracy is inconclusive, and secondary full-vocabulary probability and rank change direction under separate calibration. Target position also changes the selected-temperature mass comparison. With four relevant positions, averaging lowers total relevant mass, while its candidate-probability increase is only 5.42×10^{-4}; ordinary attention has the highest candidate probability in that condition.

The model is barely above the eight-way random baseline: its all-singleton candidate accuracy is 0.131, versus 0.125 at chance. Only one small English masked-language model, one synthetic task, the final layer, and oracle-supplied groups are tested, with all heads intervened together. These facts limit what either the downstream nulls or reversals imply about capable retrieval heads. Finite bounded logits on these prompts do not verify the Gaussian phase transition, the all-temperature separation, or the hierarchical extension. The diagnostic establishes only that the two operations are internally distinct and that their downstream effects here are small, metric-dependent, and often adverse.

Reproduction and audit. The locked evaluation used 6,240 forward passes in 165 seconds on a GTX 1080 and produced 11,520 metric rows. The supplementary material contains the preregistered protocol, pinned dependencies, intervention code, tests, and complete commands. Running them regenerates model and asset hashes, exact prompts, the protocol lock, temperatures, per-prompt outputs, and audit results under `results/transformer/`; generated outputs and model weights are not bundled. The audit checks source and protocol hashes, row uniqueness, probability bounds, temperature locks, singleton equality, native Jensen inequalities, and deterministic reanalysis.