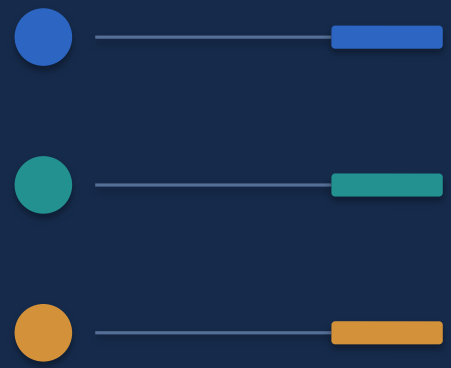
NEURIPS 2026 · 10-MINUTE PAPER PRESENTATION

Support Mismatch as a Benchmark Failure Mode for In-Context Prediction

Jie Liu[1]* · Lanlan Liang[1]* · ZiXi Yan[2] · Qinan Bao[3] · Linghao Meng[4] · Wenbo Gong[4]

[1] School of Information and Communication Engineering, Hainan University
[2] School of Computer Science and Technology, Hainan University · [3] School of Cyberspace Security, Hainan University
[4] School of Electrical and Information Engineering, Tianjin University



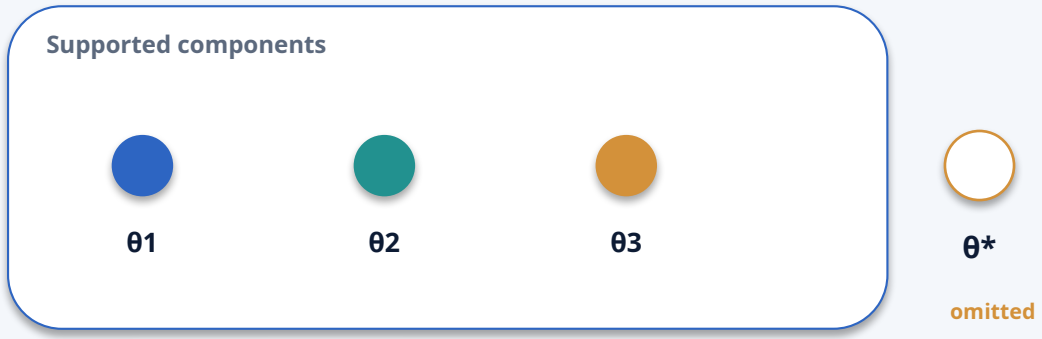
A finite-support benchmark for understanding what additional context can—and cannot—repair.

Poster #153575 · Equal contribution: first two authors · Corresponding author: Lanlan Liang

More context can reweight support, not expand it

The paper isolates a failure mode hidden by the usual well-specified-prior assumption.

FROZEN TASK SUPPORT π



The true task θ* lies outside the frozen finite support.

THE CENTRAL DISTINCTION

What context can do

Update posterior weights and reduce competition among supported tasks.

What context cannot do

Create a task that the frozen support does not contain.

Research questions: How fast does prediction self-correct? What non-vanishing floor remains?

Source: Paper §1, pp. 1–2.

A frozen finite prior induces a prediction floor

The formal object is the frozen-prior Bayes predictor—not a claim about a production LM's internal algorithm.

MODEL

$$\pi = \sum_{j=1}^{M} w_j \, \delta_{\theta_j}, \qquad \theta^* \notin \{\theta_1, \ldots, \theta_m\}$$

Context: $D_k = \{(x_t, y_t)\}_{t=1}^{k} \sim P_{\theta^*}$ iid

FROZEN-PRIOR BAYES PREDICTOR

$$\hat{f}^{\text{Bayes}}_{\pi}(D_k, x) = \sum_j w_j(D_k) \, f_{\theta_j}(x)$$

Posterior weights can move only among the M supported components.

KEY QUANTITIES

KL-closest supported component

$i^* = \arg\min_j \mathrm{KL}(P_{\theta^*} \,\|\, P_{\theta_j})$

KL separation

$\Gamma_j = \mathrm{KL}(P_{\theta^*} \,\|\, P_{\theta_j}) - \mathrm{KL}(P_{\theta^*} \,\|\, P_{\theta_i^*})$

Approximation floor

$\varepsilon^2_{\text{approx}} = \|f_{\theta_i^*} - f_{\theta^*}\|^2_{L^2}$

The floor is positive when the KL-closest supported regression function differs from the truth in L^2.

Source: Paper §3, p. 3.

Theorem 3: exponential convergence to the exact floor

A finite-sample predictor-risk result under a frozen, misspecified discrete prior.

ASSUMPTIONS

- A unique KL-closest component i^* exists: $\Gamma_{\min} = \min_{j \neq i^*} \Gamma_j > 0$.
- For each competitor, $E[\exp(-\lambda Z_t^{(j)})]$ is finite on a non-zero interval; choose $u_j \in (0, \Gamma_j)$, so $I_j(u_j) > 0$.

MAIN CONSEQUENCE

$$R_k\pi = \varepsilon^2_{\text{approx}} + O(e^{-k}\rho\min)$$

$$\rho\min = \min_{j \neq i^*} \min\{u_j, I_j(u_j)\} > 0$$

COMPONENTWISE RISK-DEVIATION ENVELOPE

$$|R_k^{\pi} - \varepsilon_{\text{approx}}^2| \leq \sum_{j \neq i^*} C_j \frac{w_j}{w_{i^*}} \left[e^{-k u_j} + e^{-k I_j(u_j)}\right]$$

$$C_j = 2|\langle a, d_j \rangle| + \|d_j\|^2_{L^2}$$

The theorem controls absolute deviation: in non-aligned geometry, finite-k risk can temporarily lie below its limiting floor.

Source: Theorem 3 and Corollary 4, §4.1, p. 4.

Why the transient decays—and why the floor remains

Posterior competition disappears exponentially; support mismatch does not.

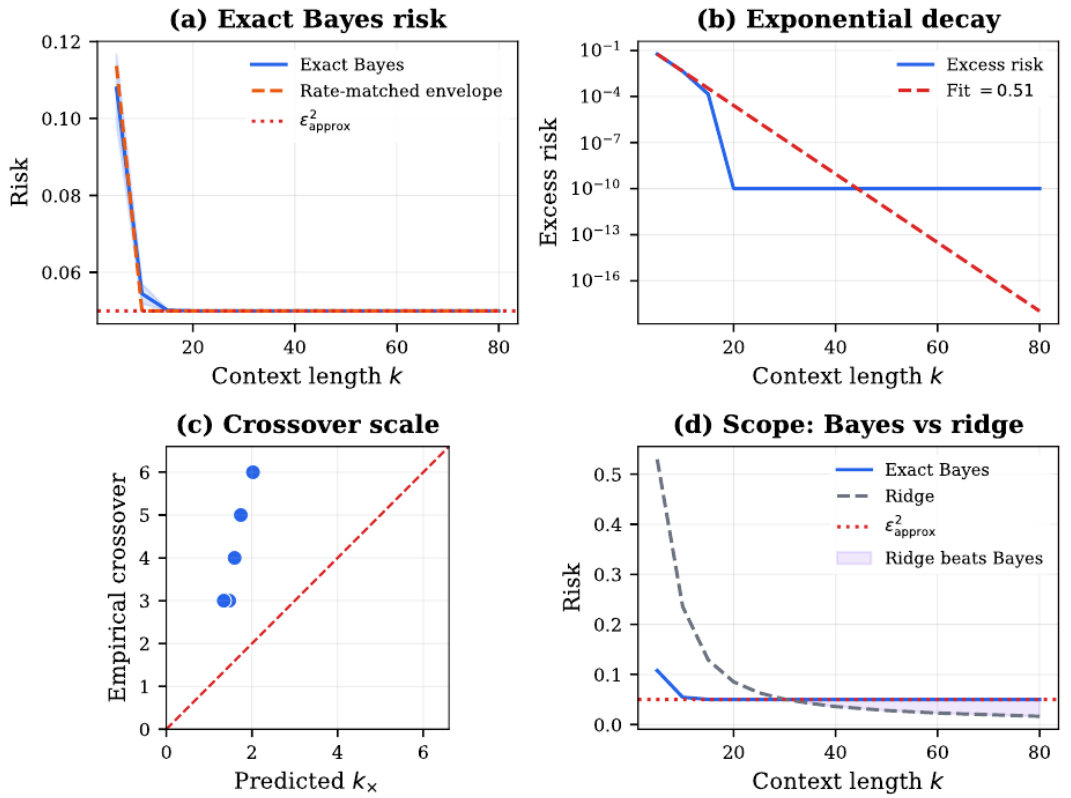
1 · GOOD LIKELIHOOD EVENT

$$\Sigma_t Z_t^{(j)} \geq k\, u_j \;\Rightarrow\; w_j(D_k)/w_i^*(D_k) \leq (w_j/w_i^*)\, e^{-ku_j}$$

2 · BAD EVENT IS RARE

$$P(\Sigma_t Z_t^{(j)} < k\, u_j) \leq e^{-kI_j(u_j)}$$

Together, both contributions to expected non-nearest posterior mass decay exponentially.



Exact Bayes diagnostics: risk approaches the floor; ridge eventually beats frozen support.

Source: Proof sketch, §4.1, p. 4; Figure 4, Appendix N, p. 25.

A sharp rate law with a phase transition at $\mu_2 = 3\mu_1$

Two-component scalar Gaussian location model: $y_t \sim N(0, \sigma^2)$, $\pi = \frac{1}{2}\delta_{\mu_1} + \frac{1}{2}\delta_{\mu_2}$, $0 < \mu_1 < \mu_2$.

ASYMPTOTIC EXCESS-RISK ORDER

SUBCRITICAL $\mu_2 < 3\mu_1$

$$\exp[-k\,\mu_1(\mu_2-\mu_1)/\sigma^2]$$

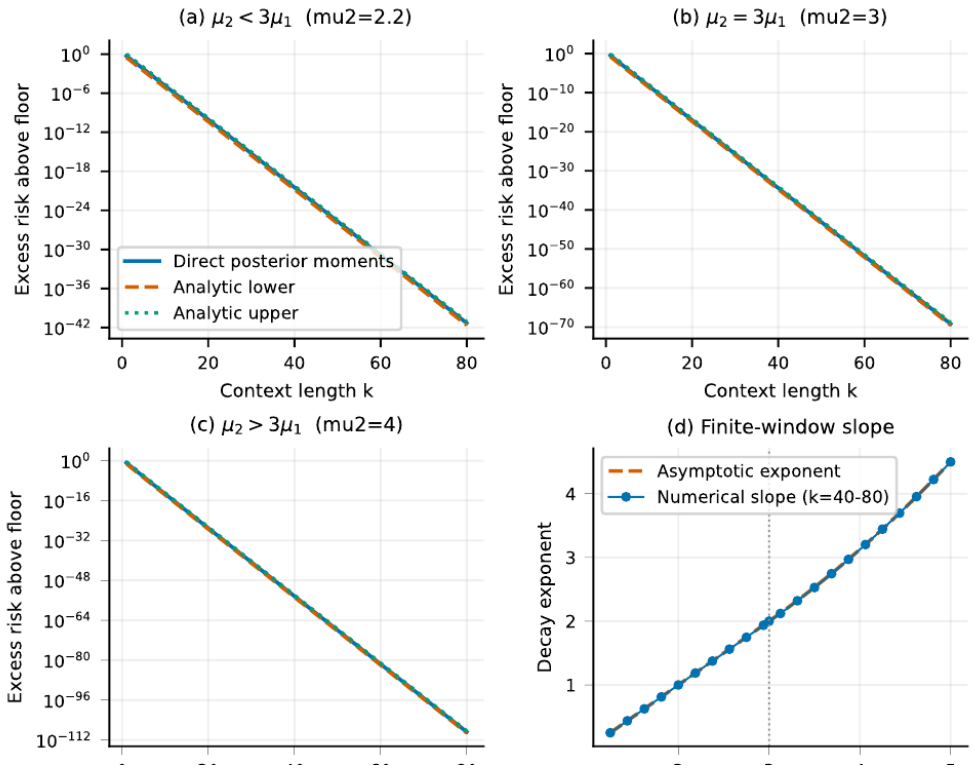
CRITICAL $\mu_2 = 3\mu_1$

$$\exp[-2k\,\mu_1^2/\sigma^2]$$

SUPERCRITICAL $\mu_2 > 3\mu_1$

$$k^{-1/2}\exp[-k(\mu_1+\mu_2)^2/(8\sigma^2)]$$

At the critical boundary, the posterior-weight and Gaussian-tail exponents coincide.



Source: Proposition 5 and Remark 7, §4.2, p. 5; Figure 5, Appendix N, p. 25.

When can leaving the frozen support become preferable?

In isotropic linear-Gaussian regression, compare an unrestricted OLS estimator with the frozen-support floor.

OLS PREDICTION RISK (k > d + 1)

$$R_k^{OLS} = d\sigma^2 / (k - d - 1)$$

SUFFICIENT CROSSING CONDITION

$$k > d + 1 + d\sigma^2 / \varepsilon^2_{approx}$$

Then OLS risk falls below the frozen-component approximation floor.

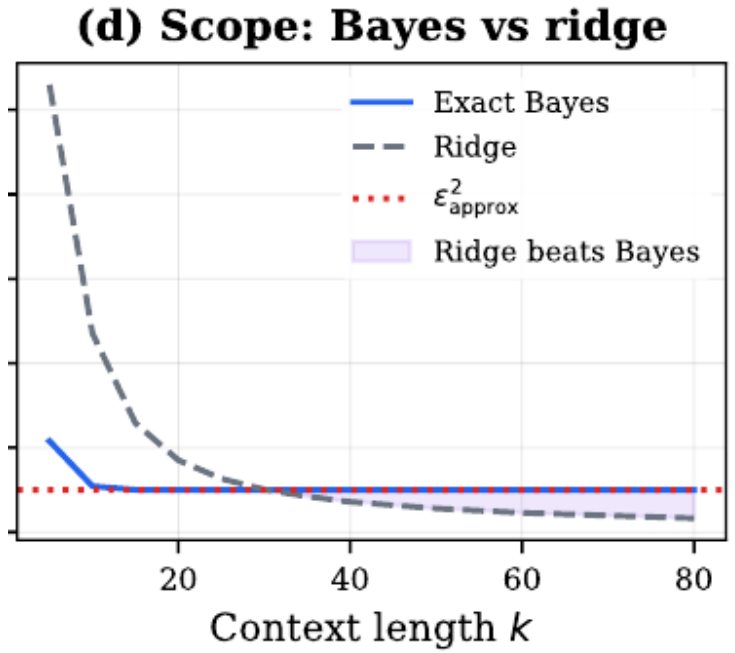


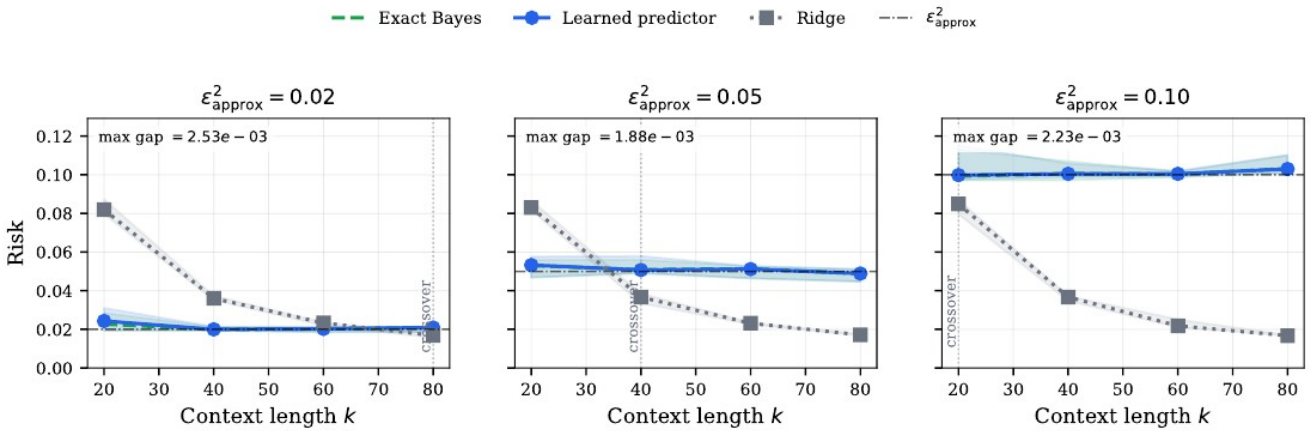
Figure 4(d) shows ridge eventually beating frozen-prior Bayes.

Scope note: the formula is Proposition 11's OLS result; the plotted comparison in Figure 4(d) is ridge.

Source: Proposition 11, §4.3, p. 6; Figure 4(d), Appendix N, p. 25.

A trained Transformer tracks the exact Bayes benchmark

Same M = 4 Gaussian support family; learned softmax Transformer with L = 8.



Worst trained–Bayes gap

$< 2.53 \times 10^{-3}$

Near ridge crossovers

Median gap is 1.2%–4.8% of the Bayes–ridge separation.

Benchmark agreement in a controlled architecture—not evidence of exact Bayesian updating in production LLMs.

Source: Figure 1 and §5.1, p. 7; Appendix M.2, p. 19.

A matched design isolates true support restoration

Open-sequence text simplification with ASSET-style inputs and reference structure.

A · Restricted

Expose only the restricted rule support.

B · Wrong, size-matched

Add a wrong rule while matching support size.

C · True-restored

Restore the omitted true transformation rule.

Primary estimand: oracle-relative full-sequence excess NLL

PROTOCOL

- Qwen2.5-7B-Instruct; three instruction variants and four matched context draws each.
- Variant v1 sweeps $k \in \{0, 1, 2, 4, 8, 16, 32\}$; v2/v3 are evaluated at $k = 32$.

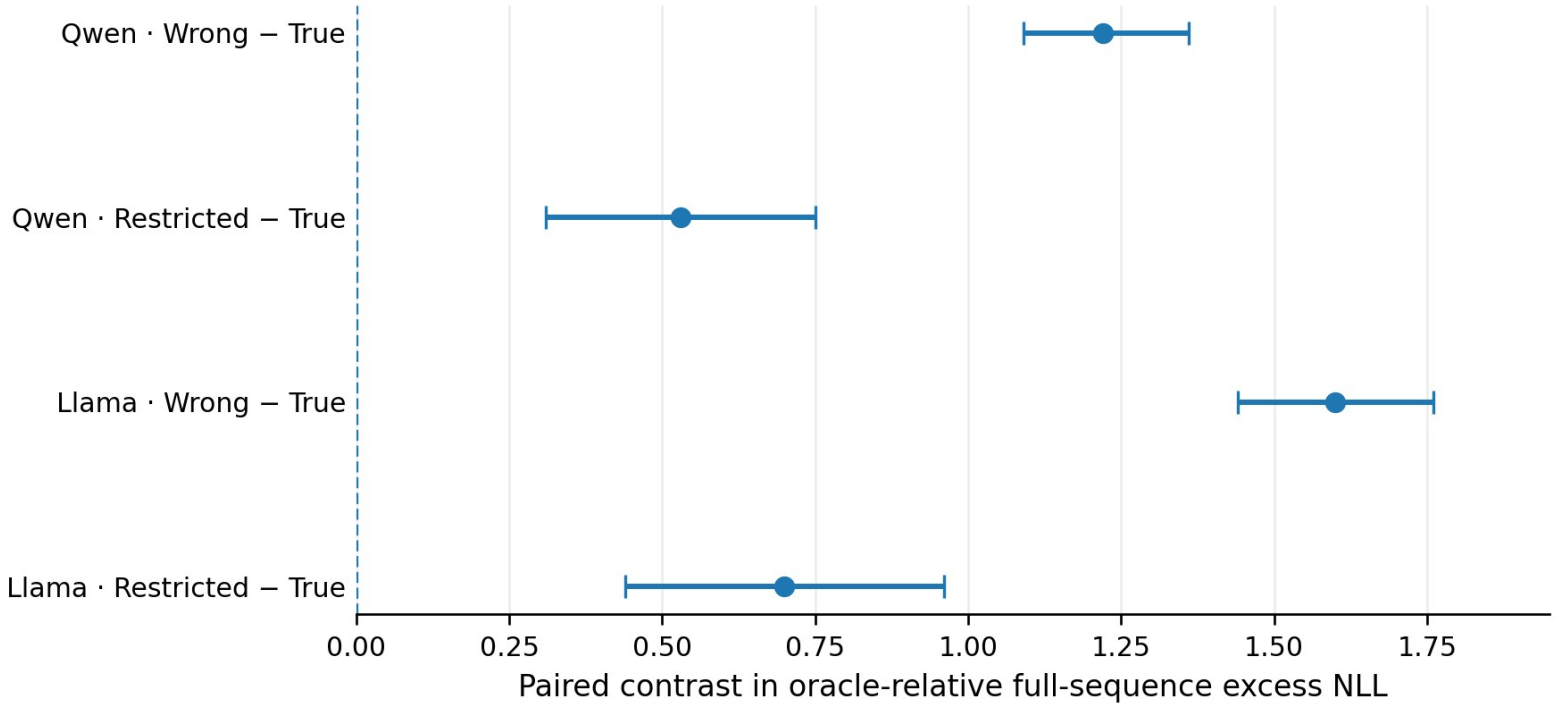
SCORING DISCIPLINE

- Full-sequence scoring; no candidate scoring, token mask, grammar constraint, or logits processor.
- Llama-3.1-8B-Instruct provides a second-model confirmation.

Source: §5.2, p. 7; Appendix M.3, p. 19.

True restoration beats both restricted and wrong support

Positive contrasts favor True restoration; all intervals shown are paired 95% confidence intervals.



12 / 12

Qwen context seeds show positive Wrong–True contrast at k = 32.

0.80

True–Wrong four-gram SARI contrast [0.58, 1.04] (auxiliary metric).

Second-model confirmation

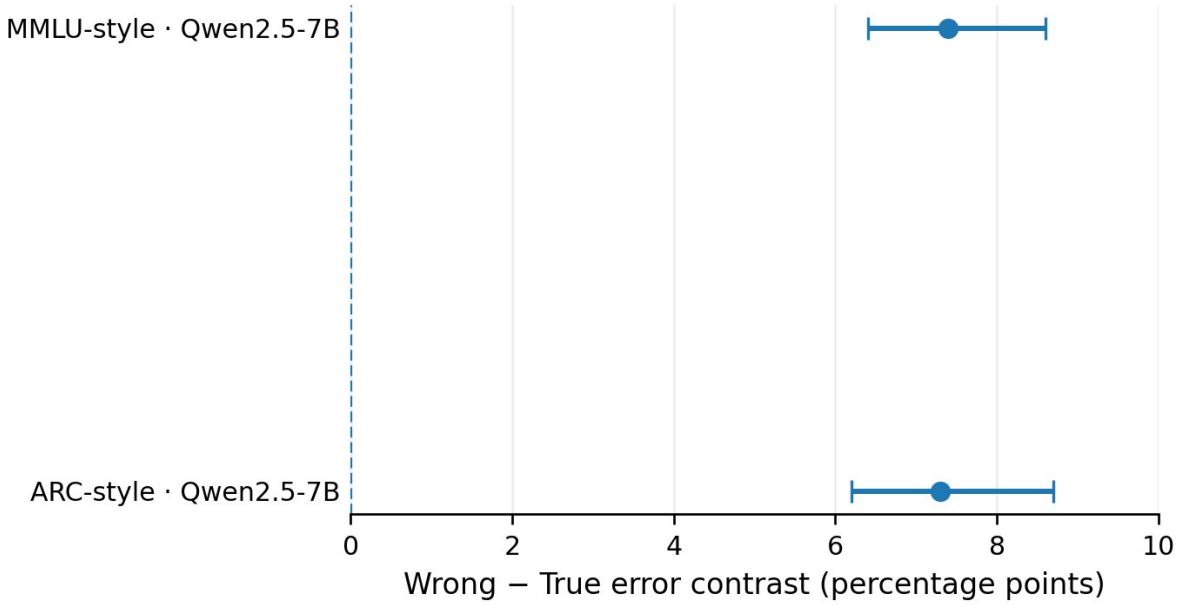
Llama Wrong–True = 1.60 [1.44, 1.76].

At k = 32. NLL is per full sequence and oracle-relative; values are not normalized by target length.

Source: Table 1, §5.2, p. 8; Appendix M.3, p. 19.

The intervention ordering also appears in subject-structured tasks

Qwen2.5-7B, k = 8; classification contrasts are percentage points.



IDENTIFICATION CONTROLS

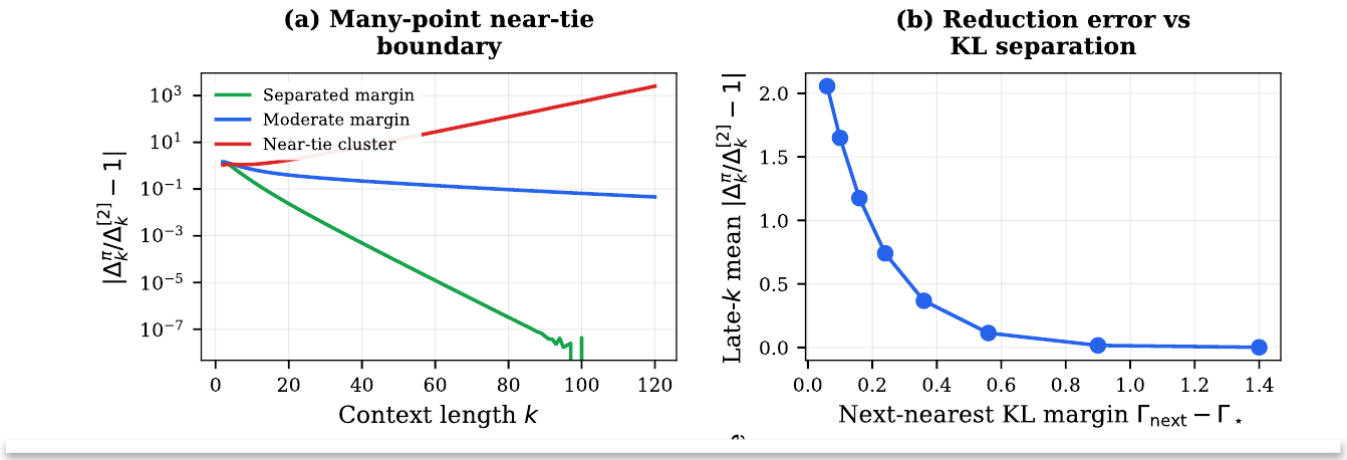
- 12 paired seeds; five-subject MMLU-style and general-science ARC-style panels.
- Fixed nonce verbalizers and four Latin-square mappings.
- Raw, contextual, and batch calibration preserve approximately seven-point Wrong–True gains.
- Demonstration-disjoint sensitivity analyses preserve positive point estimates.

Interpretation: controls constrain label, position, and calibration explanations; they do not establish robustness to arbitrary prompts.

Source: Table 1 and §5.3, p. 8; Appendix M.4–M.7, pp. 20–21.

Smaller KL gaps lengthen the transient

A controlled exact sweep links the per-sample KL gap to the posterior half-life.



Separated margins decay faster; a near-tie competitor cluster slows posterior separation.

EXACT GAP–HALF-LIFE SWEEP

KL gap	Half-life
0.8789	4
0.4394	8
0.2197	16
0.1099	32

At k = 128 with the target absent: nearest-wrong posterior mass = 0.9860 [0.9707, 0.9965]; excess NLL = 0.1108 [0.1099, 0.1123], matching the 0.1099 plateau.

Near-tie diagnostic: slower posterior separation is a benchmark boundary, not an internal-LM KL measurement.

Source: §5.4, p. 8; Figure 15, Appendix N, p. 30; Appendix M.8, p. 21.

Support mismatch is a transient-to-floor phenomenon.

The theory separates finite-context posterior competition from the irreducible error caused by omitted support.

1 · Diagnose

Track prediction as context grows; distinguish transient improvement from the remaining floor.

2 · Intervene

Compare true restoration with a wrong but size-matched support change.

3 · Adapt

When the floor dominates, consider support expansion or unrestricted adaptation under an appropriate model.

SCOPE

Exact theory applies to the stated finite-support benchmark. LM experiments test observable intervention signatures; they do not identify a production model's latent prior, posterior, or KL geometry. Evidence is limited to the reported models, prompts, and tasks.

Thank you

Source: Paper §§6–8, pp. 8–9.