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

{20233006003,lanlanliang,20233006455,20233006474}@hainanu.edu.cn
{linghaomeng,spacetime}@tju.edu.cn

Abstract

We analyze support mismatch in in-context prediction through a frozen-prior Bayes benchmark with finite discrete task support. The benchmark is diagnostic rather than literal: it does not identify a production language model with exact Bayesian updating. Under componentwise exponential-integrability of log-likelihood ratios, the predictor risk converges exponentially to the exact approximation floor, $R_k^\pi = \varepsilon^2_{\text{approx}} + O(e^{-k\rho})$; additional context removes posterior competition but cannot create a task omitted from frozen support. Scalar and aligned Gaussian location families characterize the exponent sharply and exhibit a regime change at $\mu_2 = 3\mu_1$. We examine exact Bayes, controlled learned predictors, explicit conditional-expert families, and matched language-task interventions. In the ASSET-structured Qwen2.5-7B-Instruct scenario, true restoration beats a wrong size-matched restoration at $k = 32$ by $1.22\,[1.09, 1.36]$ oracle-relative full-sequence excess-NLL units across three prompts and twelve context seeds; four-gram SARI improves by $0.80\,[0.58, 1.04]$, and Llama-3.1-8B-Instruct independently preserves the ordering. Twelve-seed subject-aligned multiple-choice controls, ambiguous-context null comparisons, nonce-label controls, and a controlled near-tie sweep support the same intervention-level picture while preserving the boundary that behavioral agreement does not recover an LM's internal prior, posterior, or KL geometry.

1 Introduction

In-context learning (ICL)—the ability of Transformer-based models to perform new tasks by conditioning on input-output examples in the prompt (Brown et al., 2020)—has emerged as a central capability of large language models. Unlike traditional supervised learning, ICL requires no gradient updates: the model implicitly infers the task from the context and generalizes to new queries. Because this behavior is often discussed as a form of rapid task adaptation, understanding its failure modes is as important as explaining its successes.

A growing body of work explains ICL as implicit Bayesian inference (Xie et al., 2022; Panwar et al., 2024). Under this view, pre-training endows the model with a prior over tasks, and each in-context example refines the posterior. This has been formalized for specific function classes (Garg et al., 2022; Akyürek et al., 2023; Zhang et al., 2024; Bai et al., 2023). However, a fundamental gap remains:

*Equal contribution.
†Corresponding author: lanlanliang@hainanu.edu.cn.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

virtually all results assume the pre-training prior is correctly specified—that the true task lies in the prior's support. In practice, that assumption is unlikely to hold exactly for deployed language models because of domain shift, task mismatch, and the impossibility of covering all downstream tasks during pre-training. This raises two questions:

> *When the pre-training prior is wrong, how quickly does the induced Bayes ICL predictor self-correct toward the KL-closest component? What is the resulting approximation floor?*

> *For a frozen discrete prior, misspecified Bayesian ICL exhibits a transient-to-floor behavior at the predictor-risk level: exponential self-correction toward the KL-closest support element until the support-mismatch floor dominates.*

Scope and object. Our formal object is the frozen-prior Bayes predictor, used as an analyzable benchmark rather than a literal model of production LMs. In the exact and controlled settings, support is a known finite task family. For production LMs, support is operational: it denotes transformation rules or hypotheses exposed through prompting, decoding, retrieval, or an external rule bank. We test intervention signatures—a plateau, selective gain under true rather than wrong size-matched restoration, and near-tie slowdown—without claiming to recover an LM-internal prior, posterior, or KL geometry.

Contributions.

1. Exact finite-sample transient-to-floor law (Theorem 3). Under componentwise exponential-integrability of log-likelihood ratios—without bounded observations—we prove $|R_k^\pi - \varepsilon^2_{\mathrm{approx}}| \leq Ce^{-k\rho_{\min}}$. Thus the posterior-mixing transient decays exponentially and the predictor risk converges to the exact support-mismatch floor $\varepsilon^2_{\mathrm{approx}}$. Appendix C gives a bounded-likelihood-ratio specialization.

2. Quantitative bridge to learned predictors and unrestricted LM generation. A controlled Transformer stays close to exact Bayes around crossover. The companion evaluation separates an externally specified conditional-expert family from a sequence-level intervention scenario. In the text-simplification scenario, Qwen2.5-7B-Instruct gives a pooled Wrong–True full-sequence excess-NLL difference of $1.22\,[1.09, 1.36]$ at $k = 32$, positive for all $12/12$ context seeds; Llama-3.1-8B-Instruct independently gives $1.60\,[1.44, 1.76]$. The sequence-scored generation arm and candidate-scored classification controls are analyzed separately.

3. Sharp Gaussian characterization and an unrestricted-adaptation boundary (Proposition 5, Corollaries 8–10, Proposition 11). In a two-point scalar Gaussian model we obtain a non-asymptotic risk sandwich with a regime-dependent exponent and critical boundary $\mu_2 = 3\mu_1$. Companion results extend this to aligned location pairs in any dimension, with exponents on the natural KL-gap scale. On aligned Gaussian pairs, the optimized upper exponent is within a factor 2 of the exact lower exponent, and an OLS baseline identifies when support-unrestricted adaptation becomes preferable.

Relation to classical misspecified Bayes. Classical misspecified-Bayes theory (Kleijn and van der Vaart, 2012; Ghosal and van der Vaart, 2017) identifies the KL minimizer asymptotically, but it does not by itself quantify the finite-sample prediction-risk transient, the exact nonvanishing support floor, or the crossover between them. Our predictor-risk bound couples posterior weights with function-space gaps, yielding $|R_k^\pi - \varepsilon^2_{\mathrm{approx}}| \leq Ce^{-k\rho_{\min}}$ and the corresponding crossover scale $k_\times$.

2 Related Work

ICL as Bayesian inference. Xie et al. (2022) first formalized ICL as implicit Bayesian inference. Panwar et al. (2024) extended this with PAC-Bayesian tools under well-specified priors, and Jeon et al. (2024) gave information-theoretic ICL bounds with $O(1/k)$ rates in the well-specified case. Our emphasis is different: we analyze support mismatch as a benchmark failure mode for LM-style in-context prediction and then test whether learned in-context predictors reproduce the same signatures.

ICL as gradient descent. von Oswald et al. (2023) and Dai et al. (2023) showed Transformers implicitly perform gradient descent. Akyürek et al. (2023) and Zhang et al. (2024) showed trained linear-attention Transformers implement ridge regression. These apply to single-component settings; our analysis targets multi-component misspecified posteriors.

Bayesian misspecification. Kleijn and van der Vaart (2012) and Ghosal and van der Vaart (2017) established asymptotic posterior concentration at the KL minimizer. Bhattacharya et al. (2019) developed finite-sample contraction for fractional posteriors. We study ordinary Bayesian updating with a frozen discrete prior; the object is the prediction rule and its approximation floor, not posterior contraction alone.

Prompt and label biases in NLP. On the NLP side, our direct language-model probe is closest in spirit to work on calibration, label bias, prompt-induced prediction artifacts, and verbalizer-sensitive prompt-based classification in few-shot and in-context settings (Schick and Schütze, 2021; Zhao et al., 2021; Fei et al., 2023; Jiang et al., 2023; Tan et al., 2025). Those works show that LM predictions can be strongly shaped by candidate labels, verbalizer choice, prompt format, and contextual bias. Our focus is complementary: we isolate one support-mismatch mechanism and analyze its finite-sample transient, floor, and crossover structure rather than proposing a new calibration procedure.

3 Problem Setup

Task and data model. The true task is parameterized by $\boldsymbol{\theta}^* \in \Theta \subseteq \mathbb{R}^d$. Context examples $D_k = \{(\mathbf{x}_t, y_t)\}_{t=1}^k$ are i.i.d. from $P_{\boldsymbol{\theta}^*}$, with $y_t = f_{\boldsymbol{\theta}^*}(\mathbf{x}_t) + \xi_t$, $\mathbb{E}[\xi_t \mid \mathbf{x}_t] = 0$ and $\mathbb{E}[\xi_t^2] = \sigma^2$. Each supported joint density $p_{\boldsymbol{\theta}_j}$ has regression function $f_{\boldsymbol{\theta}_j}$. All these functions and the true regression function are in L^2 of the true covariate law; the relevant KL divergences are finite. Gaussian results additionally assume the Gaussian noise or location model specified in their statements. The query covariate $\mathbf{x}_{k+1}$ is fresh, independent of D_k, and follows the true covariate law.

Pre-training prior (misspecified). Let $M \geq 2$ be the number of supported tasks, with positive prior weights w_j summing to one. Pre-training induces $\pi = \sum_{j=1}^M w_j \delta_{\boldsymbol{\theta}_j}$ with $\boldsymbol{\theta}^* \notin \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$.

Bayesian ICL predictor. $f^{\pi}_{\mathrm{Bayes}}(D_k, \mathbf{x}_{k+1}) = \sum_{j=1}^M w_j(D_k) f_{\boldsymbol{\theta}_j}(\mathbf{x}_{k+1})$, where $w_j(D_k) \propto w_j \prod_{t=1}^k p_{\boldsymbol{\theta}_j}(\mathbf{x}_t, y_t)$.

Key quantities. For any measurable function h of the test covariate, we write $\|h\|_{L^2}^2 := \mathbb{E}_{\mathbf{x}_{k+1}}[h(\mathbf{x}_{k+1})^2]$, where the expectation uses the test-covariate law specified above. KL-closest component: $i^* = \arg\min_j \mathrm{KL}(P_{\boldsymbol{\theta}^*} \| P_{\boldsymbol{\theta}_j})$. KL gap: $\Gamma_j = \mathrm{KL}(P_{\boldsymbol{\theta}^*} \| P_{\boldsymbol{\theta}_j}) - \mathrm{KL}(P_{\boldsymbol{\theta}^*} \| P_{\boldsymbol{\theta}_{i^*}})$. Approximation floor: $\varepsilon^2_{\mathrm{approx}} = \|f_{\boldsymbol{\theta}_{i^*}} - f_{\boldsymbol{\theta}^*}\|_{L^2}^2$. The floor is positive when these regression functions differ in L^2; parameter-support mismatch alone does not ensure a positive prediction floor.

Prediction risk. Let $\hat{f}(D_k, \mathbf{x}_{k+1})$ be a prediction rule based on the context and fresh query. We work with the noiseless prediction risk

$$\begin{aligned} R_k(\hat{f}) &:= \mathbb{E}\big[(\hat{f}(D_k, \mathbf{x}_{k+1}) - f_{\boldsymbol{\theta}^*}(\mathbf{x}_{k+1}))^2\big] \\ &= \mathbb{E}_{D_k}\big[\|\hat{f}(D_k, \cdot) - f_{\boldsymbol{\theta}^*}\|_{L^2}^2\big], \end{aligned}$$

and write $R_k^{\pi} := R_k(f^{\pi}_{\mathrm{Bayes}})$ for the frozen-prior Bayes risk. The test-time MSE is $R_k(\hat{f}) + \sigma^2$. We separate the nonvanishing support mismatch from the finite-sample transient by defining

$$\mathrm{MG}_k := R_k^{\pi} - \varepsilon^2_{\mathrm{approx}}.$$

This decomposition is useful because $\varepsilon^2_{\mathrm{approx}}$ is an intrinsic support-mismatch term that cannot be removed by additional context under the frozen prior, whereas MG_k captures the finite-sample posterior-mixing transient.

Our main theorem bounds MG_k first and then returns to R_k^{π}.

Bayes imitation. For any predictor $\hat{f}$, a standard decomposition gives $R_k(\hat{f}) \le 2\,\mathbb{E}[\|\hat{f} - f^{\pi}_{\mathrm{Bayes}}\|^2_{L^2}] + 2R^{\pi}_k$ (Appendix J). When a controlled Transformer closely imitates Bayesian prediction, the first term is small and the benchmark law governs its risk. Native-LM interventions are interpreted behaviorally and do not use this transfer bound to identify an internal algorithm.

Assumption 1 (Separation)**.** $\Gamma_{\min} := \min_{j \neq i^*} \Gamma_j > 0$.

Assumption 2 (Bounded likelihood ratios and predictions)**.** *For the auxiliary specialization, there exist $B, H > 0$ such that $\|f_{\theta^*}\|_\infty, \max_j \|f_{\theta_j}\|_\infty \le B$ and $|\log(p_{\theta_{i^*}}/p_{\theta_j})| \le H$ almost surely under P_{θ^*} for every $j \neq i^*$.*

Assumption 1 means that one supported explanation is uniquely KL-best; near ties lengthen the transient. Finite frozen support means context can reweight existing components but cannot create an omitted task. Assumption 2 is used *only* in Appendix C. Nontrivial Gaussian location ratios are unbounded and use the exponential-integrability condition of the main theorem instead; bounded shift ratios can hold for Laplace observations even when observations themselves are unbounded.

Estimation note. In parametric settings such as linear-Gaussian families, KL gaps and the floor can be estimated from held-out likelihood and residual-risk statistics. We use such plug-in quantities only for the auxiliary switching diagnostic in Appendix B.

4 Main Results

We focus on three objects: the transient-to-floor upper law (Theorem 3), Gaussian sharpness for that law (Proposition 5 and corollaries), and the regime boundary beyond which support-unrestricted adaptation is preferable (Proposition 11).

4.1 Exponential Self-Correction

Classical posterior contraction theory identifies the KL-closest component asymptotically but does not directly yield a finite-sample bound on the *prediction risk* of the frozen-prior Bayes rule. At the predictor level, the transient comes from posterior competition among support elements, while the floor comes from support mismatch itself. The following theorem quantifies the joint effect of these two mechanisms, together with the resulting crossover, at the predictor-risk level.

Theorem 3 (Exact transient-to-floor bound)**.** *Assume Assumption 1. Let*

$$a = f_{\theta_{i^*}} - f_{\theta^*}, \qquad d_j = f_{\theta_j} - f_{\theta_{i^*}}, \qquad Z_t^{(j)} = \log \frac{p_{\theta_{i^*}}(\mathbf{x}_t, y_t)}{p_{\theta_j}(\mathbf{x}_t, y_t)},$$

so $\mathbb{E}_{\theta^*} Z_t^{(j)} = \Gamma_j$. Suppose $\mathbb{E}e^{-\lambda Z_t^{(j)}} < \infty$ for $0 \le \lambda \le \lambda_{0,j}$ with some $\lambda_{0,j} > 0$. For each $j \neq i^*$, choose $u_j \in (0, \Gamma_j)$. Define

$$I_j(u) := \sup_{\lambda \ge 0} \left\{ -\lambda u - \log \mathbb{E}_{\theta^*}\left[e^{-\lambda Z_t^{(j)}} \right] \right\}.$$

Then, for all $k \ge 1$,

$$\left| R^{\pi}_k - \varepsilon^2_{\mathrm{approx}} \right| \le \sum_{j \neq i^*} \left(2|\langle a, d_j \rangle| + \|d_j\|^2_{L^2} \right) \left[\frac{w_j}{w_{i^*}} e^{-k u_j} + e^{-k I_j(u_j)} \right]. \tag{1}$$

Consequently, $|R^{\pi}_k - \varepsilon^2_{\mathrm{approx}}| \le C e^{-k\rho_{\min}}$ for $\rho_{\min} := \min_{j \neq i^*} \min\{u_j, I_j(u_j)\} > 0$ and some finite $C > 0$.

The theorem identifies $\varepsilon^2_{\mathrm{approx}}$ as the exact limiting floor. Absolute deviation is needed because, in non-aligned geometry, a finite-k mixture can temporarily have risk below its limiting nearest-component risk. The rate combines posterior-odds decay with a lower-tail Chernoff rate. The exponential-moment assumption controls misleading likelihood evidence in the tails without requiring bounded observations.

Proof sketch. Set $b_k = \sum_{j \neq i^*} w_j(D_k) d_j$. Exact expansion gives

$$R^{\pi}_k - \varepsilon^2_{\mathrm{approx}} = \mathbb{E}\left[2\langle a, b_k \rangle + \|b_k\|^2_{L^2} \right].$$

The triangle inequality and convexity (including the unused mass at the zero function) bound the absolute integrand by $\sum_{j\neq i^*} w_j(D_k)(2|\langle a, d_j\rangle| + \|d_j\|_{L^2}^2)$. On $\{\sum_t Z_t^{(j)} \geq ku_j\}$, posterior odds are at most $(w_j/w_{i^*})e^{-ku_j}$; the complement has probability at most $e^{-kI_j(u_j)}$. Taking expectations and summing proves (1); Appendix D gives the full argument. □

Appendix C records the bounded-likelihood-ratio/Hoeffding specialization.

Corollary 4 (Crossover and posterior self-correction)**.** *The same componentwise odds bound gives* $\mathbb{E}[1 - w_{i^*}(D_k)] \leq C'e^{-k\rho_{\min}}$. *If* $\varepsilon_{\text{approx}}^2 > 0$*, then for integer* $k \geq \max\{1, \lceil k_\times \rceil\}$*, where* $k_\times := \rho_{\min}^{-1} \log(C/\varepsilon_{\text{approx}}^2)$*, the deviation is at most the floor and* $R_k^\pi \leq 2\varepsilon_{\text{approx}}^2$.

The same envelope suggests a model-dependent switch from support-restricted to unrestricted adaptation; Appendix B gives the auxiliary diagnostic and robustness guarantee.

Appendix A records many-point corollaries. Nearest-competitor reduction requires separation of the *effective* componentwise rates (or retention of a leading competitor cluster), not only a unique smallest mean KL gap.

4.2 Gaussian Lower Bounds and Factor-Two Matching

The general theorem gives the structural transient-to-floor law; the Gaussian pair analysis gives its sharpest quantitative picture. In particular, this section identifies when pure exponential excess decay does and does not emerge, and shows that the general upper law is on the correct scale in Gaussian location families.

Proposition 5 (Two-component Gaussian sandwich)**.** *Consider the scalar Gaussian location model* $y_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ *with frozen prior* $\pi = \frac{1}{2}\delta_{\mu_1} + \frac{1}{2}\delta_{\mu_2}$, $0 < \mu_1 < \mu_2$. *Write* $\Delta = \mu_2 - \mu_1$ *and define*

$$A_k := \Phi\left(-\frac{\sqrt{k}(\mu_1 + \mu_2)}{2\sigma}\right),$$
$$B_k := \exp\left(-\frac{k\mu_1\Delta}{\sigma^2}\right)\Phi\left(\frac{\sqrt{k}(3\mu_1 - \mu_2)}{2\sigma}\right).$$

Then the Bayes risk satisfies

$$\mu_1^2 + \mu_1\Delta(A_k + B_k) \leq R_k^\pi,$$
$$R_k^\pi \leq \mu_1^2 + (2\mu_1\Delta + \Delta^2)(A_k + B_k),$$

and the excess $R_k^\pi - \varepsilon_{\text{approx}}^2$ *(with* $\varepsilon_{\text{approx}}^2 = \mu_1^2$*) satisfies, for large* k*,*

$$c_-\psi_k \;\leq\; R_k^\pi - \mu_1^2 \;\leq\; c_+\psi_k,$$

where ψ_k *depends on the regime:*

$$\psi_k = \begin{cases} e^{-k\mu_1\Delta/\sigma^2}, & \mu_2 < 3\mu_1, \\ e^{-2k\mu_1^2/\sigma^2}, & \mu_2 = 3\mu_1, \\ k^{-1/2}e^{-k(\mu_1+\mu_2)^2/(8\sigma^2)}, & \mu_2 > 3\mu_1. \end{cases}$$

Proof sketch. The log-odds $R_k = \log(w_1/w_2)$ are Gaussian under the true model. The Bayes predictor is $\hat{\mu}_k^\pi = \mu_1 + \Delta w_2(D_k)$. The key technical step is that $B_k = \mathbb{E}[e^{-R_k}\mathbf{1}\{R_k > 0\}]$ is a truncated Gaussian-Laplace integral whose exact evaluation exposes the regime change at $\mu_2 = 3\mu_1$, where the Gaussian-tail exponent equals the residual-weight exponent. Full proof in Appendix E. □

Corollary 6 (Gaussian pure-excess decay in the two-point model)**.** *In Proposition 5, with* $\varepsilon_{\text{approx}}^2 = \mu_1^2$*, the excess obeys*

$$c_-\psi_k \;\leq\; \text{MG}_k = R_k^\pi - \varepsilon_{\text{approx}}^2 \;\leq\; c_+\psi_k$$

for all large k*, with the same regime-dependent* ψ_k*. In particular, no additive* $O(\varepsilon_{\text{approx}}^2)$ *residual appears in this exact Gaussian pair model.*

Proof. This is immediate from Proposition 5 after substituting $\varepsilon^2_{\mathrm{approx}} = \mu_1^2$ and rearranging. □

Theorem 3 already gives exponential convergence to the exact approximation floor. The scalar Gaussian pair analysis additionally identifies the sharp excess-risk exponent and its leading prefactor.

Remark 7 (Critical boundary)**.** *The boundary $\mu_2 = 3\mu_1$ arises from competition between two mechanisms: $A_k = \Pr(R_k \le 0)$ has exponent $(\mu_1+\mu_2)^2/(8\sigma^2)$, while B_k has exponent $\mu_1(\mu_2-\mu_1)/\sigma^2$. These coincide when $\mu_2 = 3\mu_1$.*

Corollary 8 (Uniform scalar lower bound)**.** *Let $\lambda(\mu_1,\mu_2)$ denote the regime-dependent exponent. For any compact $\mathcal{K}$ within one regime, there exist $c_{\mathcal{K}} > 0$ and $K_{\mathcal{K}}$ such that for all $(\mu_1,\mu_2) \in \mathcal{K}$ and $k \ge K_{\mathcal{K}}$,*

$$R_k^{\pi} - \mu_1^2 \;\ge\; c_{\mathcal{K}}\, \omega_k\, e^{-k\beta_{\mathcal{K}}},$$

where $\beta_{\mathcal{K}} := \sup_{\mathcal{K}} \lambda$ and $\omega_k \in \{1, k^{-1/2}\}$. Moreover, $\frac{1}{4}\Gamma_2 \le \lambda \le \Gamma_2$ for all $0 < \mu_1 < \mu_2$, where $\Gamma_2 = (\mu_2^2 - \mu_1^2)/(2\sigma^2)$ is the KL gap.

Corollary 9 (Aligned non-scalar lower bound)**.** *For the two-point model $y_t \sim \mathcal{N}(0, \sigma^2 I_d)$ with $\pi = \frac{1}{2}\delta_{\mu_1} + \frac{1}{2}\delta_{\mu_2}$, $a := \mu_1^\top(\mu_2 - \mu_1) > 0$, and $\eta := 2a/\|\mu_2 - \mu_1\|^2$, the same type of compact-family lower bound holds with*

$$\lambda(\mu_1,\mu_2) = \begin{cases} a/\sigma^2, & \eta \ge 1, \\ (2a + \|\Delta\|^2)^2/(8\sigma^2\|\Delta\|^2), & \eta < 1, \end{cases}$$

satisfying $\frac{1}{4}\Gamma_2 \le \lambda \le \Gamma_2$. The alignment condition $a > 0$ ensures that residual posterior mass on μ_2 pushes risk above the floor $\|\mu_1\|^2$; when $a \le 0$, posterior mixing can temporarily improve on $\|\mu_1\|^2$ and the relevant geometry changes.

When $a \le 0$, the second support point points sufficiently against the misspecification direction that a small amount of posterior mass on μ_2 can partially debias the prediction before eventual support restriction dominates. That is why the lower-side geometry in this regime is qualitatively different rather than a minor extension of the aligned case.

The next result shows that, in the Gaussian family, the gap between upper and lower exponents is at most a factor 2 once the Hoeffding step is replaced by the optimized Cramér rate.

Corollary 10 (Factor-two comparison in Gaussian families)**.** *In the setting of Corollary 9, the optimized Cramér upper exponent from Theorem 3 is*

$$\rho_{\mathrm{G}} = \frac{\|\Delta\|^2}{2\sigma^2}\big(\eta + 3 - 2\sqrt{\eta+2}\big),$$

and satisfies $\frac{1}{2}\lambda(\mu_1,\mu_2) \le \rho_{\mathrm{G}} \le \lambda(\mu_1,\mu_2)$ for all aligned pairs, with $\rho_{\mathrm{G}} = \lambda/2$ at $\eta = 2$. Thus, within the good-event/bad-event argument of Theorem 3, the Cramér upper exponent and the exact lower exponent differ by at most a factor 2 on aligned Gaussian location pairs, and this factor is itself tight.

The equality point $\eta = 2$ is not a degenerate boundary case: in the scalar positive-support model it corresponds to $\mu_2 = 2\mu_1$. Methodologically, this indicates that the good-event/bad-event proof template is already near-optimal on the Gaussian benchmark, while broader non-Gaussian families are better interpreted in terms of rate-scale agreement than constant-factor matching; Appendix H gives one explicit Laplace illustration.

4.3 Adaptation Boundary

Proposition 11 (OLS benchmark)**.** *In isotropic linear-Gaussian regression with $x_t \sim \mathcal{N}(0, I_d)$, the OLS prediction risk is $R_k^{\mathrm{OLS}} = d\sigma^2/(k-d-1)$ for $k > d+1$. Thus $R_k^{\mathrm{OLS}} < \varepsilon^2_{\mathrm{approx}}$ whenever $k > d + 1 + d\sigma^2/\varepsilon^2_{\mathrm{approx}}$.*

OLS uses the full ambient parameter class, whereas Bayes is restricted to frozen support. The comparison locates an adaptation boundary: sufficiently many samples can offset the OLS estimation cost and lower risk below the support-mismatch floor. Appendix I gives the fixed-subspace analogue, where inaccessible error remains while the accessible component can be estimated.

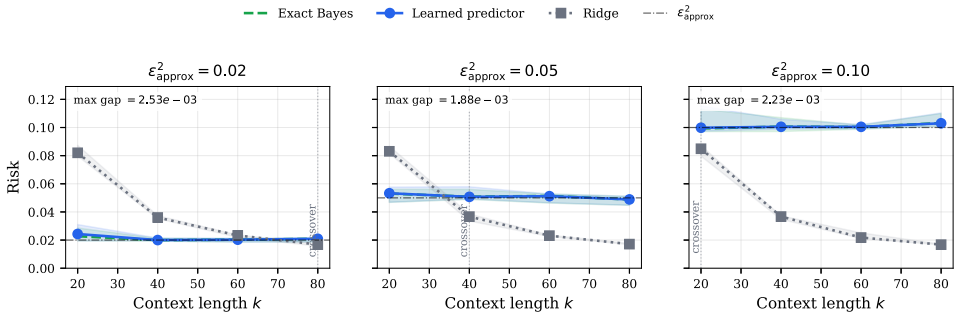


Figure 1: **Benchmark agreement across misspecification regimes.** For $\varepsilon^2_{\text{approx}} \in \{0.02, 0.05, 0.10\}$, the learned softmax Transformer (L=8), exact frozen-prior Bayes predictor, and unrestricted ridge baseline are compared with bands over anchor tasks. The learned predictor reproduces the same transient drop, floor-limited regime, and crossover ordering while staying close to exact Bayes across regimes. This is benchmark agreement in a controlled learned setting, not a validation of exact Bayesian updating in production LLMs.

5 Empirical Evaluation

Our empirical analysis has four complementary components: exact frozen-support Bayes, a controlled Transformer comparison, a finite conditional-expert calculation, and matched language-task interventions. Exact support and posterior calculations apply to the defined task families; behavioral comparisons test observable consequences rather than identifying a pretrained model's latent task prior.

5.1 Benchmark Agreement with Learned In-Context Predictors

Learned-model bridge. Figure 1 uses the same M=4 Gaussian support family while sweeping $\varepsilon^2_{\text{approx}} \in \{0.02, 0.05, 0.10\}$. The learned softmax Transformer tracks exact Bayes: the worst trained-vs-Bayes gap is below 2.53×10^{-3}, and around ridge crossovers the median gap is only 1.2%–4.8% of the Bayes-vs-ridge separation. This is finite-range agreement for a controlled architecture. Appendix M.2 specifies the recorded training configuration, architecture defaults, task generator, and evaluation grid.

5.2 Unrestricted Generation and Matched Support Interventions

Protocol. In the text-simplification scenario (with ASSET-style input and reference structure), the Qwen2.5-7B-Instruct protocol uses full-sequence scores, with no candidate scoring, token mask, grammar constraint, or logits processor. We use three instruction variants and four matched context draws each; v1 covers $k \in \{0, 1, 2, 4, 8, 16, 32\}$, while v2/v3 cover $k = 32$. Demonstrations and queries are matched while exposed rule support is Restricted, augmented with a Wrong size-matched rule, or True-restored. The primary estimand is oracle-relative full-sequence excess NLL. An explicit 1–4-gram SARI implementation is the auxiliary metric for the per-source analysis. Response-stage quality summaries are reported separately in Appendix M.10.

Results and auditable intermediate arm. At $k = 32$, pooled Qwen Wrong–True is positive for all $12/12$ context seeds (Table 1); True versus Wrong also improves sample-index-0 n-gram SARI by $0.80\,[0.58, 1.04]$. The response-stage seed-level summary additionally reports a BERTScore F1 contrast of $0.003\,[0.001, 0.005]$ under normal-reference intervals (Appendix M.10). Llama provides a second-model confirmation. In the explicitly defined conditional-expert family, Restricted and Wrong each have excess NLL 39.6166, while True restoration has 0.0000. The conditional-expert posterior is explicitly specified and is not identified with latent states of the language model.

Table 1: **Matched support-restoration effects.** Positive contrasts favor True restoration. NLL is per full sequence; classification contrasts are percentage points. Paired 95% bootstrap CIs resample context-seed means for text simplification and use the stated hierarchies for classification.

Model/task	Estimand	k	Wrong–True	Restricted–True
Qwen2.5-7B / text-simplification	excess NLL	32	$1.22\,[1.09, 1.36]$	$0.53\,[0.31, 0.75]$
Llama-3.1-8B / text-simplification	excess NLL	32	$1.60\,[1.44, 1.76]$	$0.70\,[0.44, 0.96]$
Qwen2.5-7B / MMLU-style	error (pp)	8	$7.4\,[6.4, 8.6]$	$7.6\,[6.5, 9.0]$
Qwen2.5-7B / ARC-style	error (pp)	8	$7.3\,[6.2, 8.7]$	$7.5\,[6.3, 8.9]$

5.3 Statistical and Identification Controls

Seed-level uncertainty and label artifacts. We evaluated a five-subject MMLU-style and general-science ARC-style question suite, using 12 paired seeds and hierarchical intervals, clustering MMLU by seed and subject (Table 1). Fixed nonce verbalizers, four Latin-square mappings, and raw/contextual/batch calibration preserve approximately seven-point Wrong–True gains at $k = 8$ on both tasks. A post hoc demonstration-disjoint and de-duplicated evaluation subset preserves the sign of the paired effect (Appendix M.7, Table 6); its tighter template-based variant has smaller cells and less certain science-panel estimates. At $k = 8$, the five-subject nonce-minus-semantic interaction is -0.49 percentage points, with paired hierarchical intervals crossing zero under each calibration; science-panel intervals also cross zero. A small nonzero five-subject interaction remains at $k = 2$, so label artifacts are bounded rather than eliminated. These controls reduce label-semantic, position, and fixed-calibration explanations, but do not establish robustness to arbitrary prompts.

Ambiguous contexts. To avoid weak identification after posterior collapse, we restrict to contexts with $\max_c p_{\mathrm{Bayes}}(c \mid D_k, x) < 0.8$. LM-to-Bayes KL decreases from 0.089 to 0.071 over $k = 2, 4, 8$ and is below a temperature-fitted count null by 0.017–0.212, with all paired CIs excluding zero. Brier score similarly improves over a Bayes-MAP one-hot null. This tests distributional agreement with the specified nulls. Identifying an internal decay rate requires an identified latent task and likelihood family.

5.4 Exact Diagnostics and Failure Boundaries

Near ties and the exact floor. In a controlled exact sweep, per-sample KL gaps $0.8789, 0.4394, 0.2197, 0.1099$ produce posterior half-lives $4, 8, 16, 32$. At $k = 128$ with the target absent, nearest-wrong posterior mass is $0.9860\,[0.9707, 0.9965]$, and excess NLL $0.1108\,[0.1099, 0.1123]$ matches the exact plateau 0.1099. This validates the gap–half-life–floor relationship without treating semantic cosine as an internal LM KL gap.

Auxiliary closed-set probes. Appendix M presents candidate-label restriction, verbalizer, retrieval, scale/family, and calibration studies as auxiliary ablations. In a balanced three-way task, omitting one true label creates a mechanical $1/3$ error floor for that class prevalence; these probes therefore cannot by themselves identify support mismatch. Their role is to test matched controls and robustness, while the open-sequence evaluation supplies a complementary test beyond candidate-set mechanics. The exact analyses establish the transient, crossover, and Gaussian regime change; the appendix retains prior-weight and many-competitor diagnostics.

6 Discussion

What the paper settles. For frozen discrete priors, misspecified Bayesian ICL is a transient-to-floor phenomenon rather than merely an asymptotic identification problem. Theorem 3 shows exponential convergence to the exact support-mismatch floor. Gaussian results establish sharpness on the natural KL-gap scale. Controlled learned predictors, auditable conditional experts, and language-task interventions then test progressively weaker but more realistic implications while keeping mechanism-identification boundaries explicit.

From diagnosis to correction. The benchmark suggests a practical workflow: detect a plateau as context grows; compare True restoration with a Wrong size-matched restoration; interpret a selective

True gain as a support-sensitive signature; and then expand or leave exposed support through retrieval, tools, adapters, or fine-tuning when estimated adaptation risk is below the remaining floor. Calibration can reweight existing support but cannot create an omitted rule. This is a diagnostic-and-remediation workflow, not a universal deployment guarantee.

7 Limitations

Our finite discrete prior is a scope benchmark for support-restricted adaptation, not a literal model of pretraining; Proposition 11 marks when adaptation that leaves support becomes preferable. Ties require a pseudo-true set, infinite-LLR cases need separate treatment, and continuous-support and $k \to \infty$ limits need not commute. Uniformly controlled large-M families incur an $O(\log M/\rho_{\min})$ context shift, while densification can also reduce the effective rate.

For production language models, prompt-level support is an operational construct; it does not identify pretrained-weight support, and prompt artifacts or label bias may coexist with the measured contrasts. Open-generation evidence covers controlled text simplification, two model families, and finitely many prompts; a compositional LM may synthesize an omitted rule. Full-sequence excess NLL is primary, whereas the reported four-gram SARI variant is auxiliary. The response-stage BERTScore contrast is available only as a seed-level summary. Controlled Transformer, continuous-prior, and near-tie studies validate benchmark consequences rather than native-LM internal geometry.

8 Ethical Considerations

This paper is a foundational study of in-context prediction under support mismatch. The study uses nonsensitive text tasks, controlled numerical studies, and publicly documented model/benchmark protocols and is low-risk. A potential positive impact is better detection of cases in which additional context cannot repair an omitted capability. A negative impact would arise if the benchmark were misread as a complete account of production-LM behavior or used as a deployment rule in safety-critical settings, where support mismatch may interact with harms not represented in these controlled evaluations.

Acknowledgments and Disclosures

We thank Professor Zhuhua Hu (School of Information and Communication Engineering, Hainan University; International Joint Research Center of Hainan Province for Robotics and Artificial Intelligence; Hainan Seed Industry Laboratory) for helpful suggestions and guidance, and for providing computing resources.

This research was supported by the National Natural Science Foundation of China (62361024), the Hainan Seed Industry Laboratory (B23H10004, B25H10CK11), Hainan State Farms Investment Holding Group Science and Technology Innovation Project (HKKJ2025JBGS04), the Key Research and Development Project of Hainan Province (ZDYF2026SXLH042), and the Hainan University Postgraduate Innovative Talent Cultivation Project (HDJXAL2024Y0007, HDZBJC2024Y0008).

The authors declare no competing interests.

References

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? Investigations with linear models. ICLR, 2023.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4668–4679, 2020.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In Advances in Neural Information Processing Systems, volume 36, 2023.

Anirban Bhattacharya, Debdeep Pati, and Yun Yang. Bayesian fractional posteriors. Annals of Statistics, 47(1):39–66, 2019.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Advances in Neural Information Processing Systems, volume 33, pages 1877–1901, 2020.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. arXiv:1803.05457, 2018.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can GPT learn in-context? Language models secretly perform gradient descent as meta-optimizers. In Findings of the Association for Computational Linguistics: ACL 2023, pages 4005–4019, 2023.

Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. Mitigating label biases for in-context learning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, pages 14014–14031, 2023.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? A case study of simple function classes. In Advances in Neural Information Processing Systems, volume 35, pages 30583–30598, 2022.

Subhashis Ghosal and Aad van der Vaart. Fundamentals of Nonparametric Bayesian Inference. Cambridge University Press, 2017.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. arXiv:2009.03300, 2020.

Hong Jun Jeon, Jason D. Lee, Qi Lei, and Benjamin Van Roy. An information-theoretic analysis of in-context learning. In Proceedings of the 41st International Conference on Machine Learning, volume 235, pages 21522–21554. PMLR, 2024.

Zhongtao Jiang, Yuanzhe Zhang, Cao Liu, Jun Zhao, and Kang Liu. Generative calibration for in-context learning. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 2312–2333, 2023.

Bas J. K. Kleijn and Aad W. van der Vaart. The Bernstein–von Mises theorem under misspecification. Electronic Journal of Statistics, 6:354–381, 2012.

Madhur Panwar, Kabir Ahuja, and Navin Goyal. In-context learning through the Bayesian prism. In International Conference on Learning Representations, 2024.

Timo Schick and Hinrich Schütze. Exploiting cloze questions for few-shot text classification and natural language inference. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 255–269, 2021.

Zhihang Tan, Jingrui Hou, Ping Wang, Qibiao Hu, and Peng Zhu. Surprise calibration for better in-context learning. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 23045–23060, 2025.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In Proceedings of the 40th International Conference on Machine Learning, volume 202, pages 35151–35174. PMLR, 2023.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit Bayesian inference. In International Conference on Learning Representations, 2022.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing statistical machine translation for text simplification. Transactions of the Association for Computational Linguistics, 4:401–415, 2016.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. arXiv:1904.09675, 2020.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. Journal of Machine Learning Research, 25(49):1–55, 2024.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Proceedings of the 38th International Conference on Machine Learning, volume 139, pages 12697–12706. PMLR, 2021.

A Structural Corollaries

Corollary 12 (Prior forgetting)**.** *If π and $\tilde{\pi}$ share the same support with positive weights, $\mathbb{E}[\|w^{\pi}(D_k) - w^{\tilde{\pi}}(D_k)\|_1]$ decays exponentially at rate $\rho_{\min}$. The induced Bayes predictors forget their prior weights exponentially fast.*

Proof. For either prior, the odds argument in Theorem 3 bounds the expected mass outside i^* exponentially. Since $\|w^{\pi} - w^{\tilde{\pi}}\|_1 \le 2(1 - w^{\pi}_{i^*}) + 2(1 - w^{\tilde{\pi}}_{i^*})$, the same rate applies to their difference. The L^2 distance between predictors is at most $\max_j \|f_{\boldsymbol{\theta}_j}\|_{L^2}$ times this ℓ^1 distance. This proves forgetting in expected L^2 distance. □

Corollary 13 (Effective-rate competitor reduction)**.** *For any subset S containing i^*, let R^S_k be the risk of the posterior renormalized on S. Set $F = \max_j \|f_{\boldsymbol{\theta}_j} - f_{\boldsymbol{\theta}^*}\|_{L^2}$ and $D = \max_{j,l} \|f_{\boldsymbol{\theta}_j} - f_{\boldsymbol{\theta}_l}\|_{L^2}$. Then*

$$|R^{\pi}_k - R^S_k| \le 2FD \sum_{j \notin S} \left[\frac{w_j}{w_{i^*}} e^{-k u_j} + e^{-k I_j(u_j)} \right].$$

Thus omitting components with strictly larger admitted rates gives an exponentially faster remainder relative to the retained rate envelope. Matching an actual pair lower exponent additionally requires the separation in Corollary 14; a unique smallest mean KL gap alone is insufficient.

Proof. Let $q = \sum_{j \notin S} w_j(D_k)$ and g_S be the renormalized predictor. The full predictor differs from g_S by $\sum_{j \notin S} w_j(D_k)(f_{\boldsymbol{\theta}_j} - g_S)$, whose L^2 norm is at most Dq. Both predictors have distance at most F from $f_{\boldsymbol{\theta}^*}$. The difference of their squared norms is therefore at most $2FDq$. Take expectations and apply the componentwise odds bound. □

Corollary 14 (Lifting pair lower bounds)**.** *Suppose the embedded pair $S = \{i^*, j^{\dagger}\}$ satisfies $R^S_k - \varepsilon^2_{\mathrm{approx}} \ge c_\star k^{-\alpha} e^{-k\rho_\star}$. If the omitted components admit $\rho_{\mathrm{rem}} := \min_{j \notin S} \min\{u_j, I_j(u_j)\} > \rho_\star$, then for large k, $R^{\pi}_k - \varepsilon^2_{\mathrm{approx}} \ge (c_\star/2) k^{-\alpha} e^{-k\rho_\star}$. When S is the whole support, the conclusion is immediate.*

Proof. The preceding bound gives $|R^{\pi}_k - R^S_k| \le C_{\mathrm{rem}} e^{-k\rho_{\mathrm{rem}}}$. Its ratio to the pair lower bound tends to zero because $k^{\alpha} e^{-k(\rho_{\mathrm{rem}} - \rho_\star)} \to 0$. Choose k large enough that this ratio is at most $1/2$. □

Corollary 15 (Many-point aligned Gaussian lifting)**.** *If an embedded aligned Gaussian pair containing i^* has exponent $\lambda_\star$ as in Corollary 9, and the omitted components satisfy $\rho_{\mathrm{rem}} > \lambda_\star$, the full model inherits that pair lower bound and its comparison $\frac{1}{4}\Gamma_{j^{\dagger}} \le \lambda_\star \le \Gamma_{j^{\dagger}}$.*

Proof. Apply Corollary 14 to the aligned-pair lower bound; the KL comparison is the pair identity from Corollary 9. □

B Switching Diagnostic Guarantees

This appendix records the model-dependent switching diagnostic used only as an auxiliary controlled benchmark, not as a deployment rule.

Remark 16 (Model-dependent switch diagnostic)**.** *Let $\hat{R}^{\mathrm{adapt}}_k$ be an estimated risk curve for a support-unrestricted alternative (e.g., fine-tuning, retrieval-augmented adaptation, or OLS in linear-Gaussian settings), and let*

$$\hat{k}_\times := \hat{\rho}^{-1}_{\min} \log\Big(\frac{\hat{C}}{\hat{\varepsilon_{\mathrm{approx}}}^2} \Big)$$

be computed from estimated KL geometry and floor terms. Define the estimated frozen-prior envelope

$$\hat{U}^{\mathrm{Bayes}}_k := \hat{\varepsilon_{\mathrm{approx}}}^2 + \hat{C} e^{-k\hat{\rho}_{\min}}.$$

One benchmark-guided diagnostic instantiation is to compare, at each k, the estimated frozen-prior envelope $\hat{U}_k^{\mathrm{Bayes}}$ with $\hat{R}_k^{\mathrm{adapt}}$. A threshold approximation is

$$\pi_{\mathrm{switch}}(k) = \begin{cases} \textit{frozen-prior}, & k < \hat{k}_\times, \\ \textit{unrestricted adapt.}, & k \geq \hat{k}_\times. \end{cases}$$

This turns the transient-floor law into a benchmark-guided, model-dependent diagnostic in settings where $(\hat{\rho}_{\min}, \hat{C}, \hat{\varepsilon_{\mathrm{approx}}}^2)$ are estimable, and into a structural indicator otherwise.

Proposition 17 (Crossover sensitivity under geometry-estimation error)**.** *Let*

$$k_\times = \rho_{\min}^{-1} \log\Big(\frac{C}{\varepsilon_{\mathrm{approx}}^2}\Big),$$

$$\hat{k}_\times = \hat{\rho}_{\min}^{-1} \log\Big(\frac{\hat{C}}{\hat{\varepsilon_{\mathrm{approx}}}^2}\Big),$$

with $C, \varepsilon_{\mathrm{approx}}^2, \rho_{\min} > 0$, and suppose there exist $\delta_\rho, \delta_C, \delta_\varepsilon \in [0, 1)$ such that

$$(1-\delta_\rho)\rho_{\min} \leq \hat{\rho}_{\min} \leq (1+\delta_\rho)\rho_{\min},$$
$$(1-\delta_C)C \leq \hat{C} \leq (1+\delta_C)C,$$
$$(1-\delta_\varepsilon)\varepsilon_{\mathrm{approx}}^2 \leq \hat{\varepsilon_{\mathrm{approx}}}^2 \leq (1+\delta_\varepsilon)\varepsilon_{\mathrm{approx}}^2.$$

Write $L := \log(C/\varepsilon_{\mathrm{approx}}^2)$ and

$$B_+ := \log\Big(\frac{1+\delta_C}{1-\delta_\varepsilon}\Big),$$
$$B_- := \log\Big(\frac{1+\delta_\varepsilon}{1-\delta_C}\Big),$$
$$B := \max\{B_+, B_-\}.$$

Then

$$\big|\hat{k}_\times - k_\times\big| \leq \frac{\delta_\rho}{1-\delta_\rho}\,\frac{|L|}{\rho_{\min}} \;+\; \frac{B}{(1-\delta_\rho)\rho_{\min}}.$$

Moreover, for $k = \max\{1, \lceil \hat{k}_\times \rceil\}$, the decreasing envelope of Theorem 3 gives

$$R_k^\pi \leq \Big(1 + \exp(\rho_{\min}|\hat{k}_\times - k_\times|)\Big)\varepsilon_{\mathrm{approx}}^2.$$

In particular, when $\rho_{\min}|\hat{k}_\times - k_\times| \leq 1$,

$$\begin{aligned} R_k^\pi &\leq \big(2 + 2\rho_{\min}|\hat{k}_\times - k_\times|\big)\varepsilon_{\mathrm{approx}}^2 \\ &= 2\varepsilon_{\mathrm{approx}}^2 + O(\delta\,\varepsilon_{\mathrm{approx}}^2), \end{aligned}$$

with $\delta := \max\{\delta_\rho, \delta_C, \delta_\varepsilon\}$ and constants depending on L.

In the OLS benchmark of Proposition 11, one convenient instantiation is the combined threshold

$$\hat{k}_{\mathrm{switch}} := \max\Big\{\hat{k}_\times,\; d + 1 + \frac{d\sigma^2}{\hat{\varepsilon_{\mathrm{approx}}}^2}\Big\},$$

which enforces both transient decay and OLS-below-floor conditions before switching.

C Bounded-Likelihood-Ratio Specialization

Corollary 18 (Bounded-ratio specialization)**.** *Under Assumptions 2 and 1, Hoeffding's lemma gives $I_j(\Gamma_j/2) \geq \Gamma_j^2/(8H^2)$. Since $\|a\|_{L^2}, \|d_j\|_{L^2} \leq 2B$, Theorem 3 yields*

$$\big|R_k^\pi - \varepsilon_{\mathrm{approx}}^2\big| \leq 12B^2 \sum_{j \neq i^*} \left[\frac{w_j}{w_{i^*}} e^{-k\Gamma_j/2} + e^{-k\Gamma_j^2/(8H^2)}\right].$$

Proof. Since $Z_t^{(j)} \in [-H, H]$, $\log \mathbb{E}e^{-\lambda Z_t^{(j)}} \le -\lambda\Gamma_j + \lambda^2 H^2/2$. Substitution into the definition of I_j and maximization at $\lambda = \Gamma_j/(2H^2)$ gives the stated rate. Cauchy–Schwarz bounds $2|\langle a, d_j\rangle| + \|d_j\|_{L^2}^2 \le 12B^2$. □

Taking $H = 2B^2/\sigma^2$ recovers the corresponding numerical Hoeffding constant whenever that ratio bound is verified. This is a separate bounded-ratio specialization; a nondegenerate, untruncated Gaussian observation is not almost surely bounded. Gaussian location models instead use Appendix G.

D Proofs of Main Results

D.1 Proof of Theorem 3

Proof. Let $a = f_{\boldsymbol{\theta}_{i^*}} - f_{\boldsymbol{\theta}^*}$, $d_j = f_{\boldsymbol{\theta}_j} - f_{\boldsymbol{\theta}_{i^*}}$, and $b_k = \sum_{j \ne i^*} w_j(D_k) d_j$. Then

$$R_k^\pi - \varepsilon_{\text{approx}}^2 = \mathbb{E}\big[2\langle a, b_k\rangle + \|b_k\|_{L^2}^2\big].$$

The triangle inequality gives

$$|2\langle a, b_k\rangle| \le \sum_{j \ne i^*} 2w_j(D_k)|\langle a, d_j\rangle|.$$

Attach the unused mass $1 - \sum_{j \ne i^*} w_j(D_k)$ to the zero function. Convexity then yields

$$\|b_k\|_{L^2}^2 \le \sum_{j \ne i^*} w_j(D_k)\|d_j\|_{L^2}^2.$$

Consequently,

$$\big|R_k^\pi - \varepsilon_{\text{approx}}^2\big| \le \sum_{j \ne i^*} \big(2|\langle a, d_j\rangle| + \|d_j\|_{L^2}^2\big)\,\mathbb{E}[w_j(D_k)].$$

Fix $j \ne i^*$ and let $G_{j,k} = \{\sum_{t=1}^k Z_t^{(j)} \ge k u_j\}$. On $G_{j,k}$,

$$\frac{w_j(D_k)}{w_{i^*}(D_k)} = \frac{w_j}{w_{i^*}} \exp\left(-\sum_{t=1}^k Z_t^{(j)}\right) \le \frac{w_j}{w_{i^*}} e^{-k u_j}.$$

For any $\lambda \ge 0$ in the finite negative-mgf domain,

$$\Pr(G_{j,k}^c) \le \exp\Big(k\lambda u_j + k \log \mathbb{E}e^{-\lambda Z_t^{(j)}}\Big).$$

Optimizing over λ gives $\Pr(G_{j,k}^c) \le e^{-kI_j(u_j)}$. Since $w_j(D_k) \le 1$,

$$\mathbb{E}[w_j(D_k)] \le \frac{w_j}{w_{i^*}} e^{-k u_j} + e^{-kI_j(u_j)}.$$

Substitution and summation prove (1). The finite negative moment on a nonzero interval and integrability of $Z_t^{(j)}$ imply the right derivative $\frac{d}{d\lambda} \log \mathbb{E}e^{-\lambda Z_t^{(j)}}\big|_{0+} = -\Gamma_j$. Hence $-\lambda u_j - \log \mathbb{E}e^{-\lambda Z_t^{(j)}} = \lambda(\Gamma_j - u_j) + o(\lambda) > 0$ for small positive λ, so $I_j(u_j) > 0$. Finiteness of the L^2 coefficients and finite support give the stated finite C and positive $\rho_{\min}$. □

D.2 Proof of Proposition 17

Proof. Define

$$L := \log\Big(\frac{C}{\varepsilon_{\text{approx}}^2}\Big), \qquad \hat{L} := \log\Big(\frac{\hat{C}}{\hat{\varepsilon_{\text{approx}}}^2}\Big),$$

so $k_\times = L/\rho_{\min}$ and $\hat{k}_\times = \hat{L}/\hat{\rho}_{\min}$. Then

$$\begin{aligned}\big|\hat{k}_\times - k_\times\big| &= \left|\frac{\hat{L}}{\hat{\rho}_{\min}} - \frac{L}{\rho_{\min}}\right| \\ &\le \frac{|\hat{L} - L|}{\hat{\rho}_{\min}} + |L|\left|\frac{1}{\hat{\rho}_{\min}} - \frac{1}{\rho_{\min}}\right|.\end{aligned}$$

From $(1-\delta_\rho)\rho_{\min} \le \hat{\rho}_{\min} \le (1+\delta_\rho)\rho_{\min}$:

$$\begin{aligned}\frac{1}{\hat{\rho}_{\min}} &\le \frac{1}{(1-\delta_\rho)\rho_{\min}},\\ \left|\frac{1}{\hat{\rho}_{\min}} - \frac{1}{\rho_{\min}}\right| &= \frac{|\hat{\rho}_{\min}-\rho_{\min}|}{\hat{\rho}_{\min}\rho_{\min}}\\ &\le \frac{\delta_\rho}{(1-\delta_\rho)\rho_{\min}}.\end{aligned}$$

From the multiplicative bounds on $(\hat{C}, \hat{\varepsilon_{\text{approx}}}^2)$:

$$\frac{1-\delta_C}{1+\delta_\varepsilon} \le \frac{\hat{C}/\hat{\varepsilon_{\text{approx}}}^2}{C/\varepsilon_{\text{approx}}^2} \le \frac{1+\delta_C}{1-\delta_\varepsilon}.$$

Hence

$$\begin{aligned}|\hat{L}-L| &= \left|\log\left(\frac{\hat{C}/\hat{\varepsilon_{\text{approx}}}^2}{C/\varepsilon_{\text{approx}}^2}\right)\right|\\ &\le \max\left\{\log\Big(\frac{1+\delta_C}{1-\delta_\varepsilon}\Big), \log\Big(\frac{1+\delta_\varepsilon}{1-\delta_C}\Big)\right\}\\ &= B.\end{aligned}$$

Substituting proves

$$\big|\hat{k}_\times - k_\times\big| \le \frac{\delta_\rho}{1-\delta_\rho}\,\frac{|L|}{\rho_{\min}} \;+\; \frac{B}{(1-\delta_\rho)\rho_{\min}}.$$

For the risk bound, apply Theorem 3 at $k = \max\{1, \lceil \hat{k}_\times \rceil\} \ge \hat{k}_\times$:

$$\begin{aligned}R_k^\pi &\le \varepsilon_{\text{approx}}^2 + Ce^{-\rho_{\min}k}\\ &\le \varepsilon_{\text{approx}}^2 + Ce^{-\rho_{\min}\hat{k}_\times}\\ &= \varepsilon_{\text{approx}}^2 + \varepsilon_{\text{approx}}^2 e^{-\rho_{\min}(\hat{k}_\times - k_\times)}\\ &\le \Big(1 + e^{\rho_{\min}|\hat{k}_\times - k_\times|}\Big)\varepsilon_{\text{approx}}^2.\end{aligned}$$

If $\rho_{\min}|\hat{k}_\times - k_\times| \le 1$, then $e^x \le 1 + 2x$ for $x \in [0,1]$, so

$$R_k^\pi \le \big(2 + 2\rho_{\min}|\hat{k}_\times - k_\times|\big)\varepsilon_{\text{approx}}^2.$$

Using the already proved bound on $|\hat{k}_\times - k_\times|$ yields the stated $2\varepsilon_{\text{approx}}^2 + O(\delta\,\varepsilon_{\text{approx}}^2)$ form. □

E Proof of Proposition 5

Proof. Let $R_k := \log(w_1(D_k)/w_2(D_k)) = \sum_{t=1}^k \log(p_{\mu_1}(y_t)/p_{\mu_2}(y_t))$. Under $y_t \sim \mathcal{N}(0,\sigma^2)$:

$$R_k \sim \mathcal{N}\left(k\frac{\mu_2^2-\mu_1^2}{2\sigma^2},\; k\frac{\Delta^2}{\sigma^2}\right).$$

Since $\hat{\mu}_k^\pi = \mu_1 + \Delta w_2(D_k)$ and $w_2 = 1/(1+e^{R_k})$:

$$R_k^\pi = \mu_1^2 + 2\mu_1\Delta\,\mathbb{E}[w_2] + \Delta^2\mathbb{E}[w_2^2].$$

On $\{R_k \le 0\}$: $w_2 \in [1/2, 1]$. On $\{R_k > 0\}$: $\frac{1}{2}e^{-R_k} \le w_2 \le e^{-R_k}$. Therefore $\frac{1}{2}(A_k + B_k) \le \mathbb{E}[w_2] \le A_k + B_k$, where $A_k = \Pr(R_k \le 0)$ and $B_k = \mathbb{E}[e^{-R_k}\mathbf{1}\{R_k > 0\}]$.

Standardizing: $A_k = \Phi(-\sqrt{k}(\mu_1+\mu_2)/(2\sigma))$.

For B_k: completing the square in the Gaussian-Laplace integral yields

$$B_k = \exp(-k\mu_1\Delta/\sigma^2)\,\Phi(\sqrt{k}(3\mu_1-\mu_2)/(2\sigma)).$$

Using $w_2^2 \le w_2$, the sandwich follows. The three asymptotic cases come from Mills' ratio:

$\mu_2 < 3\mu_1$**:** The argument of Φ in B_k diverges to $+\infty$, so $B_k \asymp e^{-k\mu_1\Delta/\sigma^2}$. Since $(\mu_1 + \mu_2)^2/(8\sigma^2) > \mu_1\Delta/\sigma^2$ in this regime, B_k dominates A_k.

$\mu_2 = 3\mu_1$**:** $B_k = \frac{1}{2}e^{-2k\mu_1^2/\sigma^2}$; $A_k \asymp k^{-1/2}e^{-2k\mu_1^2/\sigma^2}$, so B_k dominates.

$\mu_2 > 3\mu_1$**:** Mills' ratio on B_k's Φ factor gives $B_k \asymp k^{-1/2}e^{-k(\mu_1+\mu_2)^2/(8\sigma^2)}$, matching A_k's order. □

F Proofs of Corollaries 8 and 9

Proof of Corollary 8. From Proposition 5, $R_k^\pi - \mu_1^2 \ge \mu_1\Delta(A_k + B_k)$. On compact subcritical sets, $B_k \ge \frac{1}{2}e^{-k\lambda}$. On the critical line, $B_k = \frac{1}{2}e^{-k\lambda}$. On compact supercritical sets, $A_k \ge ck^{-1/2}e^{-k\lambda}$ by Mills' ratio. The KL comparison $\frac{1}{4}\Gamma_2 \le \lambda \le \Gamma_2$ follows from elementary algebra in each regime. □

Proof of Corollary 9. The log-odds are Gaussian with $\mu_1\Delta \to a = \mu_1^\top\Delta$ and $\Delta^2 \to \|\Delta\|^2$. The geometry parameter $\eta = 2a/\|\Delta\|^2$ replaces the scalar $\mu_2/(3\mu_1)$ threshold. The same two-term analysis applies; compactness and $a > 0$ give uniform prefactors. □

G Gaussian CGF Optimization (Proof of Corollary 10)

Proof. In the Gaussian location model, $X_t = \Delta^\top y_t/\sigma^2 \sim \mathcal{N}(0, \|\Delta\|^2/\sigma^2)$, so $\Lambda(\lambda) = \lambda^2\|\Delta\|^2/(2\sigma^2)$ and $I(u) = \sigma^2(\Gamma_2 - u)^2/(2\|\Delta\|^2)$. Setting $u = I(u)$ and solving for $u \in (0, \Gamma_2)$:

$$\begin{aligned}\rho_{\mathrm{G}} &= \Gamma_2 + \|\Delta\|^2/\sigma^2 \\ &\quad - \sqrt{(\|\Delta\|^2/\sigma^2)(\|\Delta\|^2/\sigma^2 + 2\Gamma_2)} \\ &= \frac{\|\Delta\|^2}{2\sigma^2}(\eta + 3 - 2\sqrt{\eta+2}).\end{aligned}$$

Case $\eta \ge 1$**:** $\lambda = a/\sigma^2 = \eta\|\Delta\|^2/(2\sigma^2)$. Upper bound $\rho_{\mathrm{G}} \le \lambda$: equivalent to $3 \le 2\sqrt{\eta+2}$, true for $\eta \ge 1$. Lower bound $\rho_{\mathrm{G}} \ge \lambda/2$: equivalent to $(\eta-2)^2 \ge 0$.

Case $0 < \eta < 1$**:** $\lambda = (\eta+1)^2\|\Delta\|^2/(8\sigma^2)$. Set $t = \sqrt{\eta+2}$. Upper: $(t-1)^3(t+3) \ge 0$ for $t > 1$. Lower: $-(t-1)^2(t^2+2t-7) \ge 0$ for $t \in (\sqrt{2}, \sqrt{3})$.

Tightness at $\eta = 2$: $\rho_{\mathrm{G}} = \|\Delta\|^2/(2\sigma^2) = \lambda/2$. □

H Explicit Laplace CGF Illustration

This section records the concrete non-Gaussian CGF example referenced in Section 4.

Proposition 19 (Laplace two-point CGF)**.** *In the scalar two-point Laplace location model with $0 < \mu_1 < \mu_2$, scale $b > 0$, and*

$$\begin{aligned}Z &:= \log\frac{p_{\mu_2}(Y)}{p_{\mu_1}(Y)} \\ &= \frac{|Y-\mu_1| - |Y-\mu_2|}{b}, \qquad Y \sim \mathrm{Laplace}(0, b),\end{aligned}$$

let $d := \mu_2 - \mu_1$. Then

$$\begin{aligned}\mathbb{E}[e^{\lambda Z}] = e^{-\lambda d/b}\Big(1 - \tfrac{1}{2}e^{-\mu_1/b}\Big) + e^{\lambda d/b}\tfrac{1}{2}e^{-\mu_2/b} \\ + \frac{e^{(\lambda d-\mu_2)/b} - e^{-(\mu_1+\lambda d)/b}}{2(2\lambda-1)},\end{aligned}$$

for $\lambda \ne \frac{1}{2}$, with continuous extension at $\lambda = \frac{1}{2}$.

Proof. Write

$$\mathbb{E}[e^{\lambda Z}] = \frac{1}{2b}\int_{\mathbb{R}} \exp\left(\lambda \frac{|y-\mu_1| - |y-\mu_2|}{b}\right) \cdot e^{-|y|/b}\, dy.$$

Split the integral over $(-\infty, \mu_1]$, $[\mu_1, \mu_2]$, and $[\mu_2, \infty)$. On the outer intervals, $|y-\mu_1|-|y-\mu_2| = \mp d$ is constant; on $[\mu_1, \mu_2]$, it is affine: $2y - (\mu_1 + \mu_2)$. Evaluating the three elementary exponential integrals and simplifying yields the stated expression. The apparent pole at $\lambda = \frac{1}{2}$ is removable because the numerator vanishes at the same rate. □

I Proof of Proposition 11

Proof. Condition on X: $\hat{\boldsymbol{\theta}}_{\mathrm{OLS}} - \boldsymbol{\theta}^* = (X^\top X)^{-1} X^\top \xi$ with $\mathrm{Cov} = \sigma^2 (X^\top X)^{-1}$. For independent $x \sim \mathcal{N}(0, I_d)$, the prediction risk is $\sigma^2 \mathrm{tr}((X^\top X)^{-1})$. Since $X^\top X \sim \mathrm{Wishart}(I_d, k)$, the standard exact first-moment identity for inverse Wishart matrices, valid for $k > d+1$, gives $\mathbb{E}[(X^\top X)^{-1}] = I_d/(k-d-1)$. Therefore $R_k^{\mathrm{OLS}} = d\sigma^2/(k-d-1)$. □

Remark 20 (When beating the floor does not beat Bayes)**.** *The comparison in Proposition 11 is with the limiting frozen-support* floor*, not automatically with the finite-k Bayes risk. In the aligned Gaussian pair regime with $\langle f_{\theta_{i^*}} - f_{\theta^*}, f_{\theta_j} - f_{\theta_{i^*}} \rangle \geq 0$ for all competitors, $R_k^\pi \geq \varepsilon^2_{\mathrm{approx}}$, so $R_k^{\mathrm{OLS}} < \varepsilon^2_{\mathrm{approx}}$ does suffice. Without alignment, the general theorem gives the sufficient test*

$$R_k^{\mathrm{OLS}} < \varepsilon^2_{\mathrm{approx}} - Ce^{-k\rho_{\min}} \quad \Longrightarrow \quad R_k^{\mathrm{OLS}} < R_k^\pi.$$

The margin cannot generally be removed. For example, consider the same *$d = 1$ isotropic linear-Gaussian regression family with $\theta^* = 0$, $\sigma^2 = 1$, equal prior masses on $\theta_1 = 0.4$ and $\theta_2 = -0.5$. The unique KL-nearest component is θ_1 with gap $\Gamma_2 = (0.5^2 - 0.4^2)/2 = 0.045$ and floor 0.16. At $k = 9$, $R_k^{\mathrm{OLS}} = 1/7 = 0.142857\ldots < 0.16$, while a direct Gaussian–chi-square quadrature gives $R_k^\pi \approx 0.096451$, strictly below the OLS risk. Here $S = \sum_{t=1}^9 x_t^2 \sim \chi^2_9$; conditional on S, the likelihood log-odds L of component 1 over 2 is Gaussian with mean $0.045S$ and variance $0.81S$. Thus the numerical check is exactly the two-dimensional integral*

$$R_9^\pi = \mathbb{E}_{S,L}\left[\left(0.4 - \frac{0.9}{1+e^L}\right)^2\right].$$

The accompanying CPU script independently checks the value at multiple quadrature orders. This counterexample concerns finite-context ordering; for fixed separated support, OLS eventually outperforms the positive Bayes floor as k increases.

Proposition 21 (Fixed-subspace adaptation)**.** *Let $x_t \sim \mathcal{N}(0, I_d)$ and $y_t = x_t^\top \boldsymbol{\theta}^* + \xi_t$, with independent $\xi_t \sim \mathcal{N}(0, \sigma^2)$. Fix a baseline $\boldsymbol{\theta}_{i^*}$ and $U \in \mathbb{R}^{d\times r}$, $1 \leq r \leq d$, with $U^\top U = I_r$, both independent of the context. Fit $\hat{\beta}$ by OLS of $y_t - x_t^\top \boldsymbol{\theta}_{i^*}$ on $U^\top x_t$, and predict with $\boldsymbol{\theta}_{i^*} + U\hat{\beta}$. Write $\delta = \boldsymbol{\theta}^* - \boldsymbol{\theta}_{i^*}$,*

$$a_U^2 = \|UU^\top \delta\|^2, \qquad b_U^2 = \|(I_d - UU^\top)\delta\|^2.$$

For $k > r+1$, the squared prediction risk against the true regression function is

$$R_k^{U\text{-}OLS} = b_U^2 + \frac{r(\sigma^2 + b_U^2)}{k-r-1}.$$

It is below the frozen-component floor $\varepsilon^2_{\mathrm{approx}} = a_U^2 + b_U^2$ exactly when $a_U^2 > 0$ and

$$k > r + 1 + \frac{r(\sigma^2 + b_U^2)}{a_U^2}.$$

Proof. The projected design $z_t = U^\top x_t$ is standard Gaussian in $\mathbb{R}^r$. The residual $x_t^\top (I_d - UU^\top)\delta + \xi_t$ is independent of z_t and has variance $b_U^2 + \sigma^2$. The OLS coefficient error therefore has mean squared norm $r(b_U^2 + \sigma^2)/(k-r-1)$ by the same inverse-Wishart identity. On a fresh isotropic test point, its prediction error adds to the orthogonal bias b_U^2. Comparing this risk with $a_U^2 + b_U^2$ gives the threshold. □

This fixed-subspace calculation is the fixed-subspace linear adaptation surrogate. It requires the stated fixed U and Gaussian design; data-dependent subspace learning and nonlinear LoRA or prefix tuning require separate analysis.

J Risk Decomposition

Proof of the Bayes imitation bound in Section 3. Write $\hat{f} - f_{\boldsymbol{\theta}^*} = (\hat{f} - f^{\pi}_{\mathrm{Bayes}}) + (f^{\pi}_{\mathrm{Bayes}} - f_{\boldsymbol{\theta}^*})$. By $(a+b)^2 \le 2a^2 + 2b^2$:

$$\begin{aligned} R_k(\hat{f}) &\le 2\mathbb{E}[(\hat{f} - f^{\pi}_{\mathrm{Bayes}})^2] \\ &\quad + 2\mathbb{E}[(f^{\pi}_{\mathrm{Bayes}} - f_{\boldsymbol{\theta}^*})^2] \\ &= 2\mathrm{BG}_k + 2R^{\pi}_k. \end{aligned}$$

When $\hat{f} = f^{\pi}_{\mathrm{Bayes}}$, $\mathrm{BG}_k = 0$. □

K Transformer Representability

Theorem 22 (Fixed-length continuous-task approximation)**.** *Fix $k \ge 1$, $M < \infty$, $B, \sigma > 0$, positive prior weights, and continuous scalar functions $f_{\boldsymbol{\theta}_j}$ on $[-B,B]^d$. Suppose the supported likelihoods have a common covariate density that is positive on this box and Gaussian conditional noise of variance σ^2. For every $\delta > 0$, a causal softmax-attention network with positional encodings, affine readout, and finitely many pointwise ReLU FFN blocks satisfies*

$$\sup_{\substack{x_1,\dots,x_{k+1}\in[-B,B]^d \\ y_1,\dots,y_k\in[-B,B]}} |T(D_k, x_{k+1}) - f^{\pi}_{\mathrm{Bayes}}(D_k, x_{k+1})| \le \delta.$$

The architecture uses $k+1$ routing heads and finite depth and width depending on the fixed domain, k, the functions, and δ. Here the architecture consists of residual attention and FFN blocks without layer normalization. This is compact-domain existence, with no asserted logarithmic-depth or uniform-in-k complexity bound.

Proof. **1. Continuous target.** Flatten the context and query into $z \in K = [-B,B]^n$, $n = k(d+1)+d$. The shared covariate density cancels from posterior odds, so the target is

$$F(z) = \frac{\sum_{j=1}^{M} w_j e^{-\frac{1}{2\sigma^2}\sum_{t=1}^{k}(y_t - f_{\boldsymbol{\theta}_j}(x_t))^2} f_{\boldsymbol{\theta}_j}(x_{k+1})}{\sum_{j=1}^{M} w_j e^{-\frac{1}{2\sigma^2}\sum_{t=1}^{k}(y_t - f_{\boldsymbol{\theta}_j}(x_t))^2}}.$$

If $C_f = \max_j \|f_{\boldsymbol{\theta}_j}\|_\infty$, the denominator is at least $e^{-k(B+C_f)^2/(2\sigma^2)} > 0$. Hence F is bounded and uniformly continuous on K.

2. Routing to the query token. Encode each (x_t, y_t) as one token, followed by $(x_{k+1}, 0)$; add positional one-hot vectors e_t and a constant coordinate. At the final query token, head t assigns score A to position t and zero to every other position, using only these constant and positional coordinates. With the causal mask the query can see all $k+1$ tokens. The target position has weight $e^A/(e^A+k)$, so every copied data coordinate has error at most $2Bke^{-A}$. Concatenate the $k+1$ head outputs in initially blank residual channels and discard the dummy query response. The resulting $\tilde{z}$ lies in K. Choose finite A large enough that uniform continuity gives $|F(\tilde{z}) - F(z)| \le \delta/2$.

3. A finite ReLU approximation at the query. On a rectangular grid of sufficiently small mesh, multilinear interpolation has the form

$$I(z) = \sum_{v\in\mathcal{V}} F(v) \prod_{a=1}^{n} h_{v,a}(z_a),$$

where the one-dimensional piecewise-linear hat functions $h_{v,a}$ form a partition of unity and lie in $[0,1]$. Uniform continuity makes $\|I - F\|_\infty \le \delta/4$. Each hat is exactly a finite affine combination of ReLUs, since a continuous piecewise-linear function can be written as its initial affine segment plus the slope changes times $(z_a - b)_+$ at its finitely many knots.

Products are approximated by finite ReLU blocks as follows. Linear interpolation Q of r^2 on $[-2,2]$ with mesh h has error at most $h^2/4$. The identity $st = ((s+t)^2 - (s-t)^2)/4$ therefore gives multiplication error at most $h^2/8$ on $[0,1]^2$. Clip the approximate output to $[0,1]$ with ReLUs; this cannot increase error. Composing $n-1$ such multiplications gives product error at most $(n-1)h^2/8$, by induction. Choose h so that $|\mathcal{V}| \max_v |F(v)|(n-1)h^2/8 \le \delta/4$ (the zero-coefficient case is immediate). The finite parallel sum then yields a ReLU network N with $\|N - F\|_\infty \le \delta/2$.

4. FFN realization and total error. Embed the successive affine–ReLU computations of N into finitely many pointwise FFN blocks, using extra channels for their intermediate values. Signed coordinates can be carried by $u = \mathrm{ReLU}(u) - \mathrm{ReLU}(-u)$; an FFN can cancel the old residual and write the next intermediate state. Set subsequent attention outputs to zero, so they do not alter the routed query state. An affine readout returns $N(\tilde{z})$. Thus $|N(\tilde{z}) - F(z)| \le \delta/2 + \delta/2$, as required. □

The domain bound here specifies the approximation problem, not an almost-sure bound on a Gaussian random observation. The result does not supply a training algorithm or certify the learned eight-layer architecture. The earlier informal depth/width estimate is replaced by this explicitly scoped existence statement; discontinuous task functions and growing context lengths require separate assumptions.

L Scalar Gaussian Corollaries

Corollary 23 (Unequal weights)**.** *For fixed $\alpha \in (0,1)$, replacing equal weights by $\alpha\delta_{\mu_1} + (1-\alpha)\delta_{\mu_2}$ changes prefactors only; the large-k exponent is unchanged.*

Proof. For fixed $\alpha \in (0,1)$, writing $r = (1-\alpha)/\alpha$ gives posterior mass $re^{-R_k}/(1+re^{-R_k})$. Its ratio to the equal-weight mass $e^{-R_k}/(1+e^{-R_k})$ lies between $\min\{1,r\}$ and $\max\{1,r\}$. Its first and second moments, and hence the positive excess risk, differ only by fixed positive factors. □

Corollary 24 (Finite supports)**.** *For $0 < \mu_1 < \cdots < \mu_M$ with positive weights, the excess $\mathbb{E}[(\hat{\mu}_k^\pi)^2] - \mu_1^2 \asymp \psi_{2,k}$ is controlled by the nearest competitor μ_2.*

Proof. Write $\Delta_j = \mu_j - \mu_1$, $R_{j,k} = k\Delta_j((\mu_1+\mu_j)/2 - \bar{Y})/\sigma^2$, and $r_j = w_j/w_1$. Then $w_j(D_k) \le \min\{1, r_j e^{-R_{j,k}}\}$, so $\mathbb{E}w_j(D_k) \le A_{j,k} + r_j B_{j,k}$ with the two Gaussian integrals from Proposition 5. If $b = \sum_{j\ge 2} \Delta_j w_j(D_k)$, convexity gives $b^2 \le \sum_{j\ge 2} \Delta_j^2 w_j(D_k)$. Thus the excess is bounded above by $\sum_{j\ge 2}(2\mu_1\Delta_j + \Delta_j^2)(A_{j,k} + r_j B_{j,k})$. The scalar exponent $\lambda(\mu_1,\mu_j)$ strictly increases with μ_j, so $j=2$ controls this upper bound, including its polynomial prefactor.

For a lower bound, on $R_{2,k} > 0$ we have $\bar{Y} < (\mu_1+\mu_2)/2$, hence $R_{j,k} > 0$ for every $j \ge 2$. The posterior denominator $\sum_j w_j e^{-R_{j,k}}$ (with $R_{1,k} = 0$) is then at most one, giving $w_2(D_k) \ge w_2 e^{-R_{2,k}}$. Therefore the excess is at least $2\mu_1\Delta_2 w_2 B_{2,k}$. The explicit Gaussian integral $B_{2,k}$ has order $\psi_{2,k}$ in all three regimes, proving both sides. □

Corollary 25 (Exponent law)**.** $\lim_{k\to\infty} -k^{-1}\log(\mathbb{E}[(\hat{\mu}_k^\pi)^2] - \mu_1^2) = \lambda(\mu_1,\mu_2)$ *for all finite positive supports. The critical boundary and piecewise law persist.*

Proof. Take logarithms of the preceding two-sided bound and divide by $-k$; both constant prefactors and $\log k/k$ vanish. □

M Additional Experiments

M.1 Evaluation Design and Reporting Conventions

The experiments combine (i) an exact frozen-support Bayes model, with defined task family, likelihood and posterior, (ii) a controlled learned Transformer, (iii) an explicitly specified conditional-expert calculation, and (iv) matched language-task interventions. These components address different estimands, with exact posterior claims confined to the defined Bayesian models. Full-sequence excess NLL is not normalized by target length and is not compared in absolute value across tokenizers.

Classification contrasts are percentage points. Positive Wrong–True and Restricted–True contrasts favor True restoration.

M.2 Controlled Learned-Model Protocol

The archived implementation package contains the controlled training, architecture, sweep, aggregation and plotting scripts. Its central aggregate configuration records $d = 5$, $M = 4$, Gaussian noise variance $\sigma^2 = 0.25$, floor 0.05, $50{,}000$ training steps, batch size 64, and learning rate 3×10^{-4}. The fixed softmax model has eight layers, $d_{\text{model}} = 128$, FFN width 256, and training lengths $k = 5$–60. At unseen $k = 80$, its saved risk is 0.05088066 versus approximately 0.05000 for exact Bayes. The bridge sweeps floors $0.02, 0.05, 0.10$; the family archive contains 48 OOD settings.

The supplied source uses four attention heads by default, AdamW with weight decay 10^{-4}, $2{,}000$ warmup steps followed by cosine decay, and base seed 42 by default; its sweep launcher lists $42, 43, 44$. It samples fixed support vectors from $\mathcal{N}(0, I/d)$ with uniform weights and covariates from $\mathcal{N}(0, I)$. Context labels have Gaussian noise variance σ^2; the test parameter offsets one support vector by $\sqrt{\varepsilon^2_{\text{approx}}}$ orthogonally to the support span. Training targets are noiseless query responses. The ridge call uses $\lambda = \sigma^2$, hence 0.25 in the central configuration.

The central aggregate and source defaults have different scopes: aggregate values determine the plotted central condition, whereas defaults specify the implementation and executable settings. In the bridge figures, uncertainty bands summarize variation over anchor tasks, not independently retrained network seeds. The appendix reports the evaluation estimand and averaging unit separately from the optimization settings.

M.3 E1: Open-Sequence Text-Simplification Study

Protocol. The open-sequence protocol evaluates complete sentence transformations rather than candidate-set classifications. It does not mask output tokens or impose a grammar constraint. The first instruction variant is evaluated at $k \in \{0, 1, 2, 4, 8, 16, 32\}$; variants v2/v3 and the Llama comparison have $k = 32$ records only. Each variant uses four paired context draws. Source texts and target references are held fixed across Demonstrations-only, Restricted, Wrong size-matched, and True-restored conditions. The primary outcome is full-sequence oracle-relative excess NLL, without length normalization. At $k = 32$, each variant–condition has four seeds, 359 test sources per seed, and six target samples per source (8,616 sequence scores). Within a seed we average six targets per source before aggregating sources; the reported 95% paired-bootstrap intervals resample four seed means within a prompt variant, or the 12 variant–seed means for pooled Qwen contrasts. Individual sequences are therefore not treated as independent confidence- interval replicates. Rows, matching IDs, and paired contrasts are supplied.

Table 2: **Text-simplification contrasts at $k = 32$.** Paired full-sequence excess-NLL contrasts.

Variant	Wrong–True	Restricted–True	Wrong–Restricted	W–T positive
v1	$1.27\,[1.00, 1.55]$	$0.59\,[0.07, 1.02]$	$0.68\,[-0.00, 1.20]$	4/4
v2	$1.30\,[1.14, 1.46]$	$0.60\,[0.34, 0.86]$	$0.70\,[0.60, 0.80]$	4/4
v3	$1.10\,[0.94, 1.26]$	$0.40\,[0.14, 0.66]$	$0.70\,[0.60, 0.80]$	4/4
Pooled	$\mathbf{1.22\,[1.09, 1.36]}$	$\mathbf{0.53\,[0.31, 0.75]}$	$\mathbf{0.69\,[0.47, 0.88]}$	**12/12**

Table 3: **Sample-index-0 text-quality contrasts at $k = 32$.** The metric is a fully specified four-gram set-based SARI variant on the enclosed ten-reference text corpus; paired intervals resample context seeds.

Contrast	Estimate [95% CI]
True–Wrong n-gram SARI	$0.80\,[0.58, 1.04]$
True–Restricted n-gram SARI	$1.24\,[1.05, 1.46]$

Across variants, the text-quality numbers report the first sampled continuation per source and context draw. The reported n-gram SARI is recomputed from the accompanying reference texts and decoded strings and is secondary; the theory-facing estimand is full-sequence excess NLL.

Auditable conditional-expert arm. The conditional-expert analysis uses an explicitly specified finite family of sequence-likelihood models. At $k = 32$, mean excess NLL is 39.6166 for Restricted, 39.6166 for Wrong size-matched support, and 0.0000 for True restoration. Restricted–True is $39.6166\,[37.9108, 41.6281]$ and Wrong–True is $39.6166\,[37.9122, 41.6153]$. Its posterior is defined over the specified expert family and does not identify internal states of a pretrained language model.

Second-model confirmation. For the Llama-3.1-8B-Instruct protocol at $k = 32$, excess NLL is 22.3 (Restricted), 23.2 (Wrong), and 21.6 (True), giving Wrong–True $1.60\,[1.44, 1.76]$ and Restricted–True $0.70\,[0.44, 0.96]$. Absolute NLL values are not compared across Qwen and Llama tokenizers.

M.4 E2: Seed-Level Subject-Structured Questions

The four-option subject-structured questions have 12 paired seeds at $k = 2, 4, 8$. Confidence intervals cluster by seed and, for the five-subject panel, by subject. The panel comprises abstract algebra, college computer science, high-school physics, professional law, and clinical knowledge. Every corresponding question, answer key, demonstration ID, candidate score, and matched support condition is included.

Table 4: **Twelve-seed classification results.** Errors and contrasts are percentage points.

Task / k	Restricted	Wrong	True	Wrong–True	Restricted–True
Five-subject / 2	42.2	42.0	39.9	$2.1\,[1.6, 2.7]$	$2.3\,[1.8, 3.0]$
Five-subject / 4	39.8	39.6	35.3	$4.3\,[3.4, 5.2]$	$4.5\,[3.9, 5.3]$
Five-subject / 8	38.7	38.5	31.1	$7.4\,[6.4, 8.6]$	$7.6\,[6.5, 9.0]$
Science / 2	40.6	40.5	38.2	$2.3\,[1.5, 3.0]$	$2.4\,[1.8, 3.2]$
Science / 4	37.9	38.0	33.4	$4.6\,[3.7, 5.5]$	$4.5\,[3.5, 5.6]$
Science / 8	37.0	36.8	29.5	$7.3\,[6.2, 8.7]$	$7.5\,[6.3, 8.9]$

M.5 E3: Ambiguous-Context Distributional Diagnostics

We restrict to evaluation contexts satisfying $\max_c p_{\mathrm{Bayes}}(c \mid D_k, x) < 0.8$, using 8 seeds and 200 contexts per seed and k. Nulls are a temperature-fitted count softmax, Bayes-MAP one-hot prediction, and uniform prediction. In the supplied local preparation archive, all 7,200 stored calibration/evaluation contexts meet this threshold, with zero fallback selections. Preparation context IDs match the corresponding scored rows supplied in the supplement.

Table 5: **Distributional diagnostics on deliberately non-collapsed contexts.**

k	LM KL	LM Brier	LM–count ΔKL	LM–MAP ΔBrier	log-odds slope / R^2
2	0.089	0.052	$-0.017\,[-0.026, -0.008]$	$-0.475\,[-0.480, -0.469]$	$0.76\,[0.72, 0.80]$ / 0.54
4	0.078	0.045	$-0.078\,[-0.093, -0.046]$	$-0.470\,[-0.473, -0.466]$	$0.81\,[0.78, 0.84]$ / 0.61
8	0.071	0.037	$-0.212\,[-0.234, -0.166]$	$-0.206\,[-0.219, -0.192]$	$0.85\,[0.83, 0.87]$ / 0.66

These comparisons are more discriminating than overall median TV after posterior collapse. They reveal structure beyond the specified nulls but do not identify native Bayesian computation.

M.6 E4: Nonce Labels and Calibration

We use fixed nonce verbalizers, four Latin-square mappings, 12 paired seeds, and raw, contextual, and batch calibration. At $k = 8$, Wrong–True error is approximately 7.1 percentage points on both the five-subject and science panels under all three calibrations. The nonce-minus-semantic Restricted–True interaction is small but not identically zero: at $k = 8$ on the five-subject panel it is $-0.49\,[-1.04, 0.07]$ percentage points under raw calibration, $-0.49\,[-1.02, 0.02]$ under contextual calibration, and $-0.49\,[-1.00, 0.02]$ under batch calibration. The corresponding science-panel intervals also include zero. At $k = 2$ in the five-subject panel the negative interaction remains small but the three paired intervals exclude zero. Thus these controls constrain rather than eliminate label-mapping effects. The supplementary paired-interaction table is recomputed from matched E2 and E4 rows under a shared seed–subject–question hierarchy.

M.7 E2–E4: Demonstration-Disjoint Sensitivity

We additionally performed a post hoc subset analysis to check that matched support contrasts were not driven exclusively by repeated question stems. We kept a single query for each normalized stem only if that stem was absent from the complete demonstration bank, and removed scored units whose selected demonstrations repeated a stem within the prompt. Conditions were removed as complete matched triplets, never individually. The exact-stem filter folds case and whitespace; the stricter template filter additionally replaces numerical literals before matching. These are fixed reanalyses of the accompanying scored prompts, not new model executions.

Table 6: **Disjoint-subset $k = 8$ Wrong–True error contrast.** Error differences and paired bootstrap confidence intervals are in percentage points. E4 is shown under raw calibration; contextual and batch estimates and all three tested values of k appear in the supplementary CSV. The units and available seeds reflect filtering.

Study	Panel	Exclusion	Paired units	Seeds	Δerror, 95% CI
E2	Five-subject	Exact stem	1,866	12	$7.38\,[5.67, 9.27]$
E2	Five-subject	Numeric template	474	10	$8.37\,[5.05, 11.97]$
E2	Science	Exact stem	1,404	12	$6.70\,[5.13, 8.33]$
E2	Science	Numeric template	413	7	$7.02\,[4.12, 10.41]$
E4	Five-subject	Exact stem	1,134	12	$7.51\,[5.08, 10.31]$
E4	Five-subject	Numeric template	384	10	$8.53\,[4.77, 12.74]$
E4	Science	Exact stem	360	12	$5.83\,[1.67, 10.28]$
E4	Science	Numeric template	189	7	$3.70\,[-1.06, 8.47]$

All eight point estimates remain positive; the exact-stem subset has confidence intervals excluding zero for these four $k = 8$ comparisons. The more restrictive template filter leaves fewer clusters (as few as seven seeds in the science panel), and the E4 science interval includes zero. Confidence intervals resample paired seed, subject where applicable, and question units (2,500 replicates with a fixed analysis seed). The complete selection, point estimates, cluster counts, and confidence intervals are recomputed by `code/panel_sensitivity.py`.

M.8 E5: Controlled Near-Tie Sweep

Here *empirical grid half-life* means the first tested k for which the average posterior weight on the wrong expert in the equal-prior target-in-support pair is at most 0.25, down from 0.50 at $k = 0$. It is a discrete-grid crossing, not a continuously fitted decay rate.

Table 7: **Controlled gap–half-life relationship.**

Disagreement fraction	Exact per-sample KL gap	Posterior half-life
0.5000	0.8789	4
0.2500	0.4394	8
0.1250	0.2197	16
0.0625	0.1099	32

At $k = 128$ in the target-absent condition, nearest-wrong posterior mass is $0.9860\,[0.9707, 0.9965]$ and excess NLL is $0.1108\,[0.1099, 0.1123]$, matching the exact nearest-wrong plateau 0.1099. This is an exact controlled benchmark and not a native-LM internal KL measurement.

M.9 Auxiliary Closed-Set Diagnostics

The complementary closed-set suite uses balanced three-way candidate-label tasks, 12 seeds, and 320 held-out queries per seed. Because omitting one class creates a mechanical error contribution equal to that class prevalence, these results are now treated as auxiliary support-ablation and robustness checks rather than standalone evidence of an LM-internal support floor.

Candidate cardinality, lexical length, and label frequency are matched across the relevant controls. Nonce mappings remove fixed label semantics and position, while calibration reweights existing

Table 8: **Closed-set classification results at $k = 8$.** Lower is better.

Model	Task	Restricted	Full	Expanded	Calibrated
gpt2	Topic	0.412	0.231	0.154	0.359
gpt2	Sentiment	0.512	0.285	0.303	0.366
Qwen2.5-3B	Topic	0.333	0.018	0.018	0.043
Qwen2.5-3B	Sentiment	0.330	0.011	0.010	0.183
Qwen2.5-7B	Topic	0.330	0.015	0.022	0.029
Qwen2.5-7B	Sentiment	0.330	0.004	0.005	0.079
Gemma4-E4B-it (chat)	Topic	0.435	0.242	0.245	0.267
Gemma4-E4B-it (chat)	Sentiment	0.332	0.026	0.056	0.058

labels without restoring an omitted one. Retrieval is a controlled proxy that adds query-conditioned aliases from a fixed balanced bank rather than a learned retriever.

M.10 Response-Stage Summary and Final Reporting

The author-provided response-stage snapshot is retained separately from the per-source analyses reported above. Its saved units are prompt–seed metric means, seed–subject error counts, and E3 seed-level diagnostics. Table 9 records the correspondence between the two versions. Matching source IDs alone does not establish equality of decoded predictions or context snapshots.

Table 9: **Response-stage summaries and final per-source analyses.** Intervals are 95% intervals under each version's stated method. NLL contrasts are Wrong–True or Restricted–True; quality contrasts are True–Wrong. Classification contrasts are percentage points. The SARI implementations and E3 context snapshots differ between versions.

Quantity	Response-stage summary	Final per-source analysis
Qwen NLL, Wrong–True	$1.23\ [0.91, 1.55]$	$1.22\ [1.09, 1.36]$
Qwen NLL, Restricted–True	$0.53\ [0.22, 0.84]$	$0.53\ [0.31, 0.75]$
Llama NLL, Wrong–True	$1.60\ [0.70, 2.50]$	$1.60\ [1.44, 1.76]$
SARI, True–Wrong	$1.37\ [0.82, 1.92]$	$0.80\ [0.58, 1.04]$
BERTScore F1, True–Wrong	$0.003\ [0.001, 0.005]$	Not evaluated
Five-subject error, Wrong–True	$7.4\ [6.2, 8.6]$	$7.4\ [6.4, 8.6]$
Five-subject error, Restricted–True	$7.6\ [6.4, 8.8]$	$7.6\ [6.5, 9.0]$
Science error, Wrong–True	$7.3\ [5.9, 8.7]$	$7.3\ [6.2, 8.7]$
Science error, Restricted–True	$7.5\ [6.1, 8.9]$	$7.5\ [6.3, 8.9]$
E3 LM-to-Bayes KL, $k = 8$	0.068	0.071
E3 LM–count ΔKL, $k = 2$	$-0.131\ [-0.158, -0.104]$	$-0.017\ [-0.026, -0.008]$
E3 LM–count ΔKL, $k = 4$	$-0.143\ [-0.170, -0.116]$	$-0.078\ [-0.093, -0.046]$
E3 LM–count ΔKL, $k = 8$	$-0.151\ [-0.179, -0.123]$	$-0.212\ [-0.234, -0.166]$

Statistical units. The response-stage replay uses normal-reference intervals $\hat{\delta} \pm 1.959964\,\widehat{\mathrm{SE}}$, rather than percentile bootstrap intervals. For G fixed prompt strata, its mean is the average of stratum means, with variance $G^{-2}\sum_g s_g^2/n_g$. Qwen has three strata and four context seeds per stratum; Llama has four paired seed means. E2 uses the variance of twelve seed contrasts divided by twelve, with an additional within-seed subject component for the five-subject panel. Test queries are fixed. These summary intervals and the final row-based bootstrap intervals answer different conditional uncertainty questions.

Quality definitions and analysis files. The response-stage SARI uses EASSE with 13a lowercase normalization, ten references, and deletion F1. Its BERTScore uses RoBERTa-large, layer 17, without IDF or baseline rescaling, taking ten-reference maxima. The supplement supplies the paired seed summaries, their analysis program, and a text-quality evaluator that exports sentence scores and corpus SARI separately. These auxiliary values retain their score version; they are not pooled with the final per-source analysis. E3 likewise retains its own context version and supplied seed-level diagnostics. Dataset aliases refer to the enclosed structured panels rather than official benchmark test splits. Configurations, comparison tables, and programs are in `response_summary/`. The paper's primary empirical conclusions use the final per-source analysis.

M.11 Reproducibility and Compute

The controlled-model supplement contains training scripts and aggregate training records. The accompanying supplement provides complete paired E1–E4 score tables, input/question banks, quality metric definitions, and matching rules, alongside E5 numerical rows. Training workloads differ across experiments; the reported tables are paired at their specified unit of variation rather than grouped by a shared hardware-session identifier. Source-level SARI is recomputed from the supplied text and references. The separate response-stage BERTScore summary and its execution scope are documented in Appendix M.10. Large pretrained weight files and the canonical public benchmark corpora are not bundled.

External assets and terms. ASSET uses CC BY-NC 4.0; `cais/mmlu`, MIT; and `allenai/ai2_arc` (ARC-Challenge), CC BY-SA 4.0. Qwen2.5-7B base/Instruct and Gemma-4-E4B-it use Apache 2.0; Qwen2.5-3B base/Instruct use the Qwen Research License; Llama-3.1-8B-Instruct uses the Llama 3.1 Community License and acceptable-use terms; GPT-2 uses MIT. The supplement provides upstream release links and licensing information in `ASSETS_AND_LICENSES.md`. The software license does not supersede third-party terms.

N Additional Diagnostic Figures

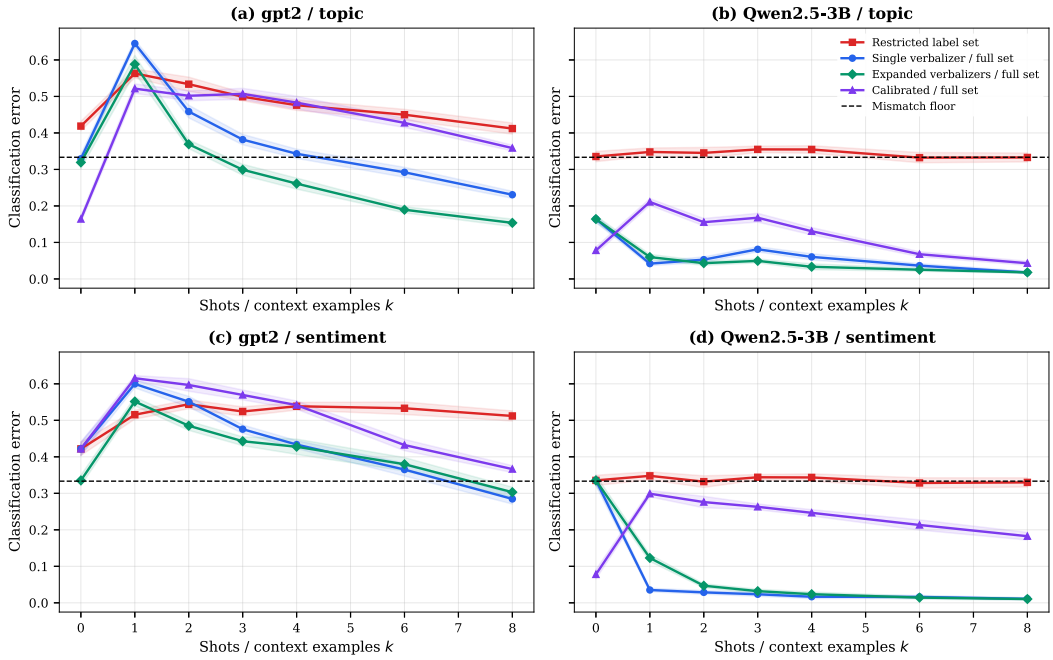


Figure 2: **Closed-set classification diagnostics.** Topic and sentiment classification on gpt2 and Qwen2.5-3B under restricted, full, expanded-verbalizer, and calibrated candidate sets. These results are retained as auxiliary ablations because deleting the true class creates a mechanical error contribution.

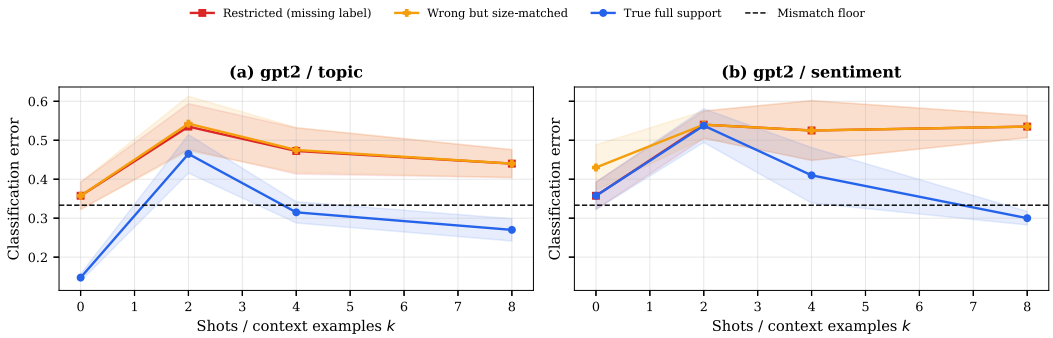


Figure 3: **Targeted size-matched support control on the legacy LM.** Restricted, wrong-but-size-matched, and true-full candidate sets use the same prompt family and query pool. Matching cardinality alone does not reproduce true restoration.

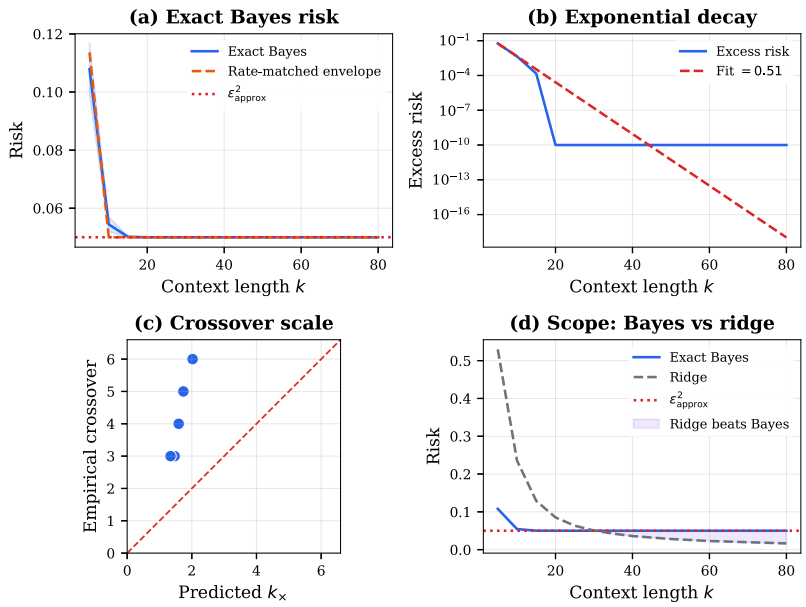


Figure 4: **Self-correction and adaptation diagnostics.** (a) Risk approaches the controlled floor $\varepsilon^2_{\text{approx}} = 0.05$. (b) The stored excess decreases over the observed transient before reaching numerical saturation. (c) Observed crossover summaries are compared with envelope thresholds $k_\times$ on the specified evaluation grid. (d) Ridge eventually beats the frozen-prior predictor, illustrating the scope boundary.

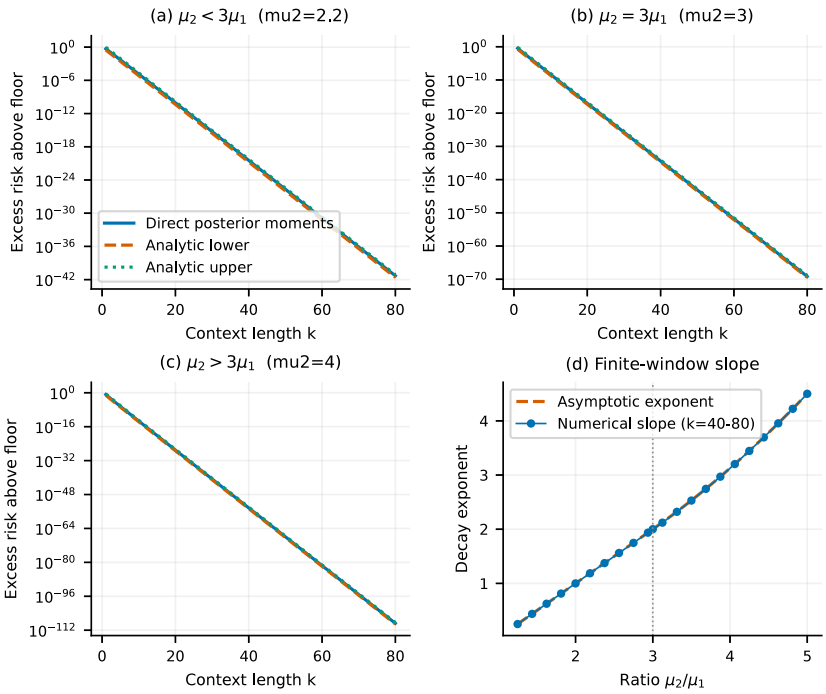


Figure 5: **Numerically stable Gaussian illustration.** Equal weights, $\mu_1 = \sigma = 1$. (a)–(c) Direct quadrature of $2\mu_1\Delta\mathbb{E}[w_2] + \Delta^2\mathbb{E}[w_2^2]$ for $\mu_2 = 2.2, 3, 4$ lies between the analytic envelopes of Proposition 5. No subtraction of nearly equal risk and floor is used. (d) The theoretical exponent is compared with slopes over $k = 40$–80, after removing the $k^{-1/2}$ factor in the supercritical regime. All curves use the stated Gaussian parameters and deterministic direct quadrature.

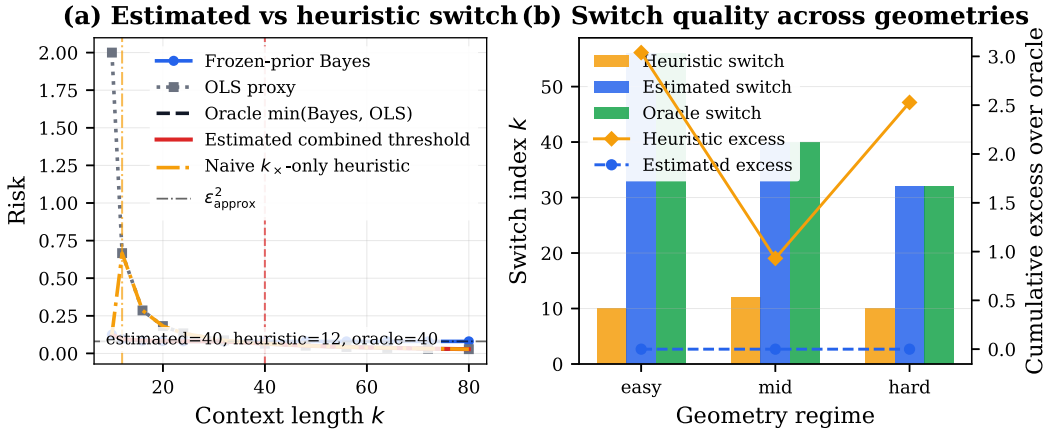


Figure 6: **Benchmark-guided switching diagnostic (oracle vs estimated vs heuristic).** (a) For a representative geometry, the estimated combined-threshold criterion $\hat{k}_{\text{switch}} = \max\{\hat{k}_\times, \hat{k}_{\text{OLS}}\}$ stays close to the oracle envelope, while a naive $k_\times$-only heuristic switches too early. (b) Across easy/mid/hard geometries, the estimated switch matches the oracle switch on the evaluation grid, whereas the heuristic switch incurs large early-switching errors and substantial excess over oracle.

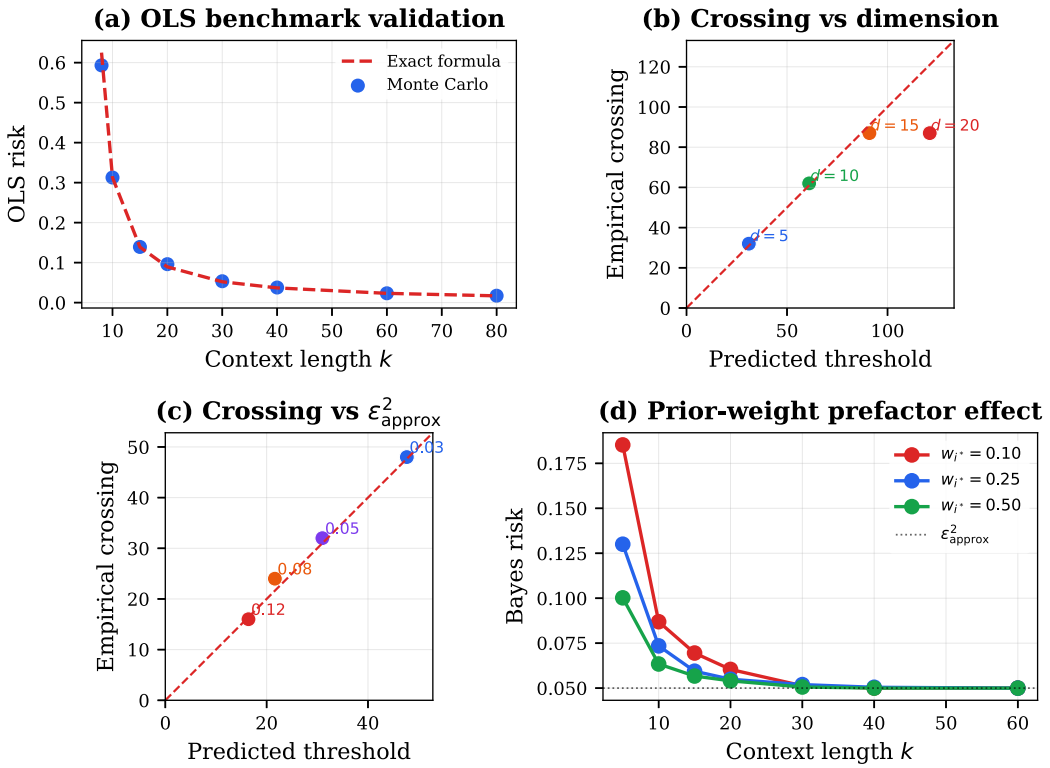


Figure 7: **Robustness suite.** (a) OLS validation. (b) Crossing vs dimension. (c) Crossing vs floor. (d) Weight effect.

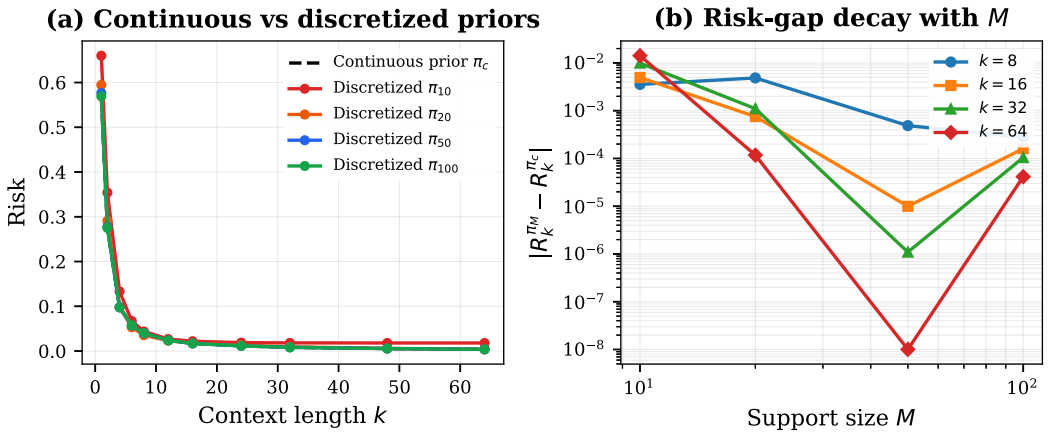


Figure 8: **Continuous-prior discretization diagnostic.** (a) In a scalar Gaussian-prior linear model, Bayes risk under M-point discretized priors tracks the continuous-prior Bayes risk increasingly well as M grows. (b) At fixed context lengths, the absolute risk gap $|R_k^{\pi_M} - R_k^{\pi_c}|$ decreases with M. This is numerical evidence for the finite-discrete prior as a coarse-grained approximation in a controlled setting.

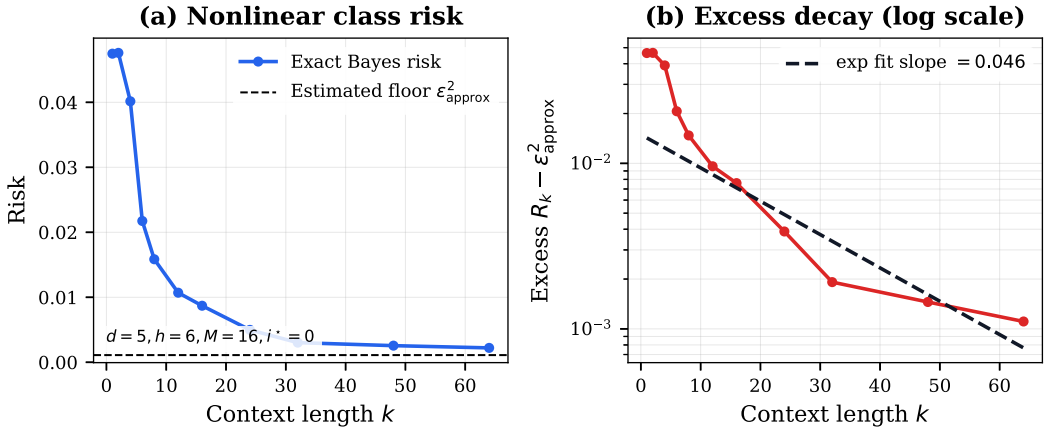


Figure 9: **Nonlinear class diagnostic.** In a two-layer ReLU function class with frozen finite prior support over network parameters, exact-Bayes risk again shows transient-to-floor behavior: panel (a) risk approaches an irreducible floor set by support mismatch; panel (b) the excess above that floor decays approximately exponentially over the observed context range.

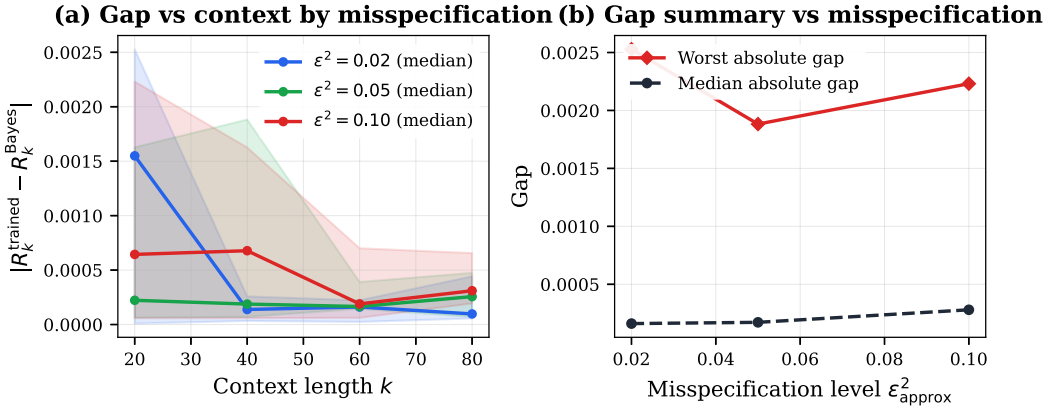


Figure 10: **Misspecification-strength bridge diagnostic.** Using the trained-family OOD sweep, panel (a) shows trained-vs-exact-Bayes absolute gaps versus context length for multiple $\varepsilon^2_{\text{approx}}$, with bands over anchor tasks; panel (b) summarizes median and worst-case absolute gaps versus $\varepsilon^2_{\text{approx}}$. The bridge remains tight across the tested misspecification levels.

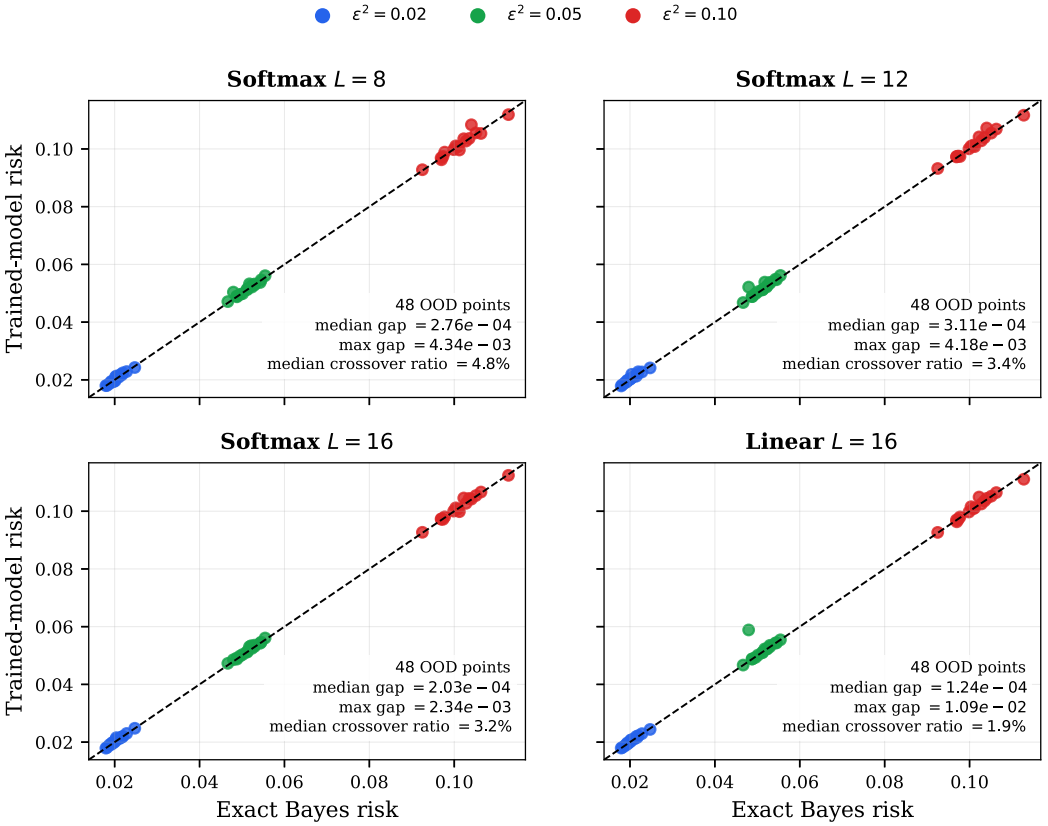


Figure 11: **Stronger-model bridge diagnostic.** On the same 48-point misspecified OOD family used for the learned bridge, Bayes-risk vs trained-risk scatter is shown for softmax $L{=}8, 12, 16$ and linear-attention $L{=}16$. The deeper softmax models remain tightly clustered around the Bayes identity line, while linear attention remains close on median but shows a heavier worst-case tail. This strengthens the learned-model bridge without changing the benchmark scope.

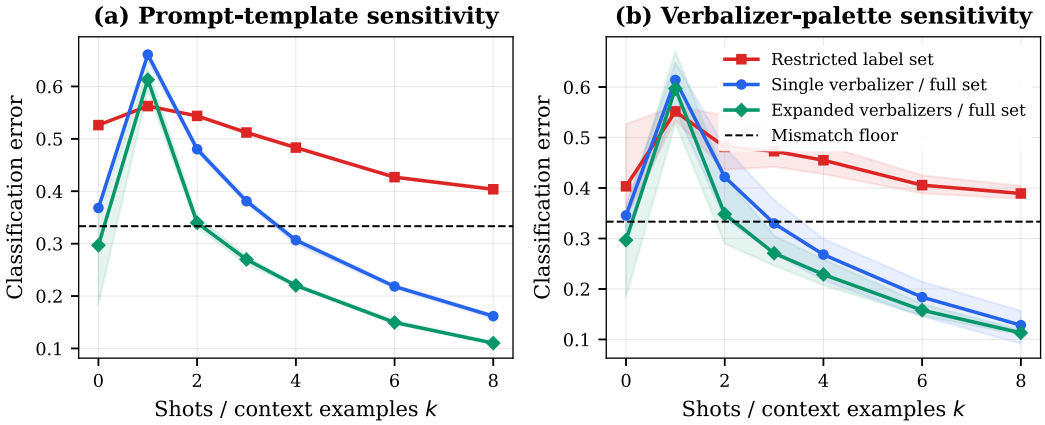


Figure 12: **Prompt and verbalizer sensitivity of the direct gpt2 baseline.** Panel (a) aggregates gpt2 restricted/full/expanded curves across three prompt templates, shading the min–max envelope over templates. Panel (b) aggregates the same curves across three single/expanded verbalizer palettes. Across the tested prompts and palettes, the qualitative ordering restricted $>$ single-verbalizer full $>$ expanded-verbalizer full is unchanged at moderate and large k, indicating that the Figure 2 pattern is not tied to one wording or one verbalizer palette.

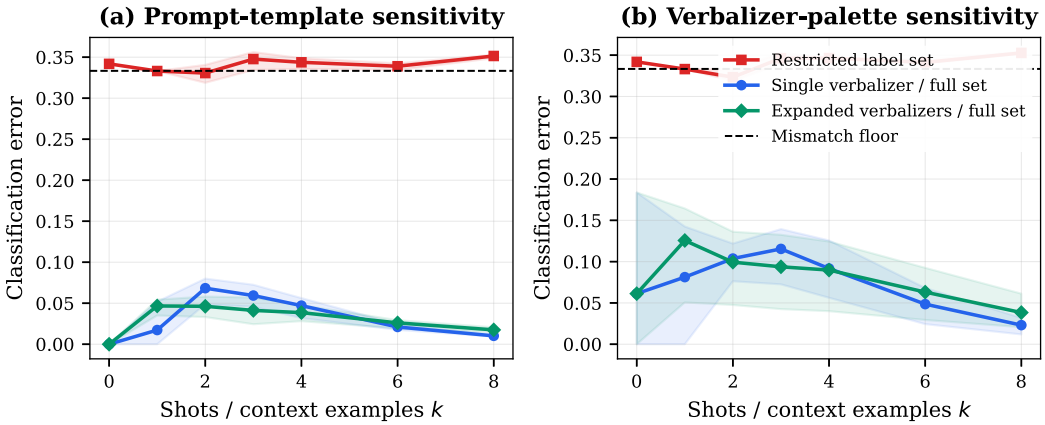


Figure 13: **Prompt and verbalizer sensitivity of the direct Qwen2.5-3B probe.** Panel (a) aggregates Qwen2.5-3B restricted/full/expanded curves across three prompt templates, shading the min–max envelope over templates. Panel (b) aggregates the same curves across three single/expanded verbalizer palettes. Even on the more modern decoder-only model, the qualitative ordering restricted $>$ full remains stable across the tested prompt and verbalizer variants.

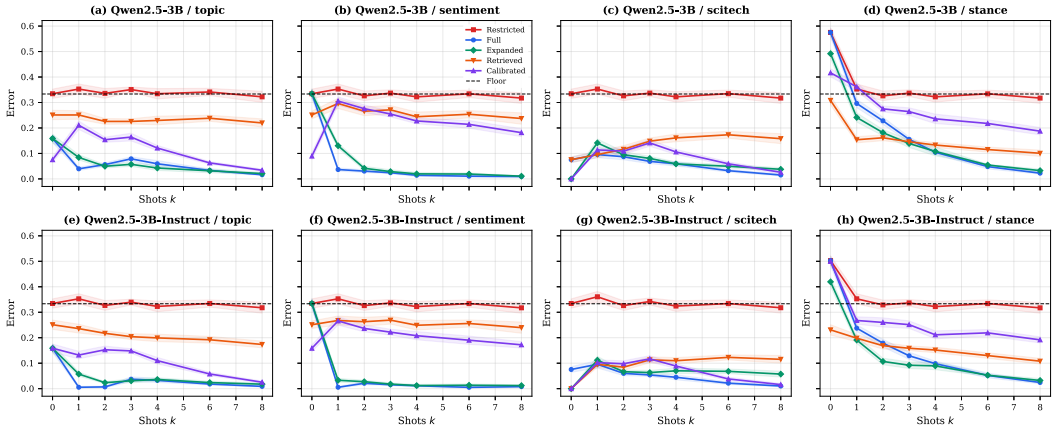


Figure 14: **Extended direct-LM suite with instruction tuning, additional tasks, and retrieval-augmented candidate pools.** Rows compare the base Qwen2.5-3B model with the instruction-tuned Qwen2.5-3B-Instruct variant. Columns span four task families, including a deliberately near-tie science-versus-technology task and a stance task with a mixed-label class. Each panel compares restricted decoding, a single-verbalizer full candidate set, static verbalizer expansion, a retrieval-augmented candidate pool, and contextual calibration. Across tasks and both models, retrieval lowers error relative to restricted decoding but remains weaker and less stable than true full-set support; the near-tie task also shows visibly slower early improvement than the separated topic task.

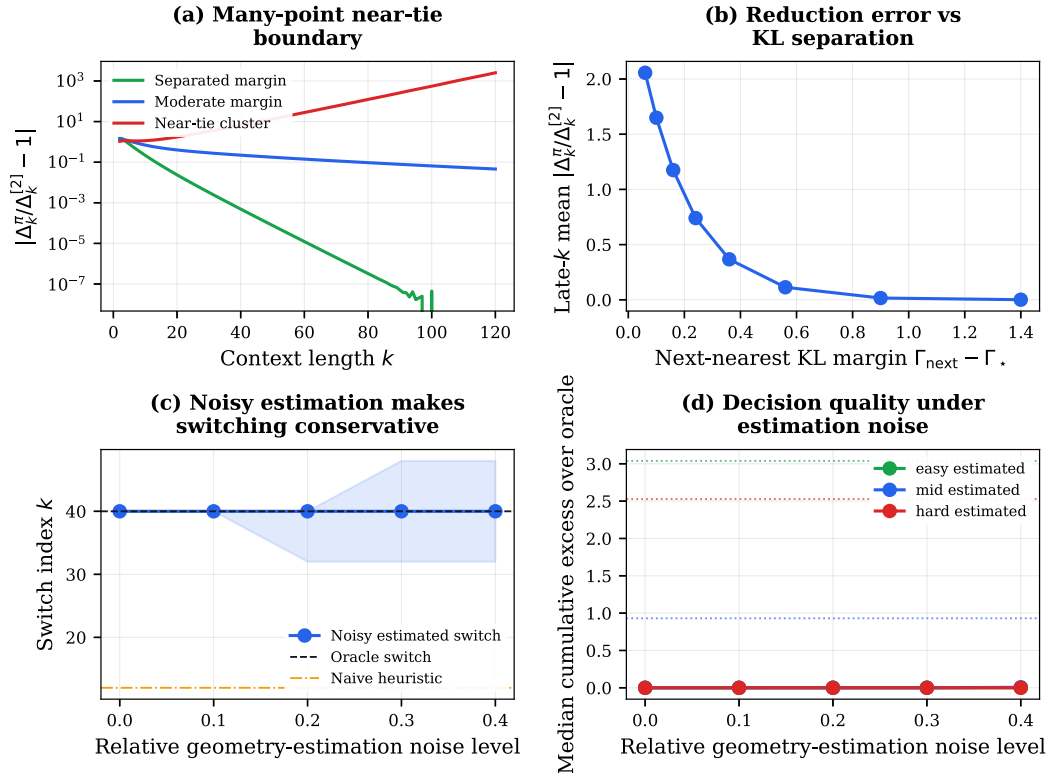


Figure 15: **Boundary diagnostics.** Panels (a,b) compare nearest-competitor reduction across the tested separated and near-tie geometries; a near-tie competitor cluster degrades the comparison. The lower-bound lifting theorem additionally requires separation of the omitted-component rates from the retained pair exponent. Panels (c,d) stress-test the switching diagnostic under multiplicative geometry-estimation noise: the estimated switch becomes progressively more conservative, yet its excess over oracle remains far below the naive heuristic across the tested regimes.

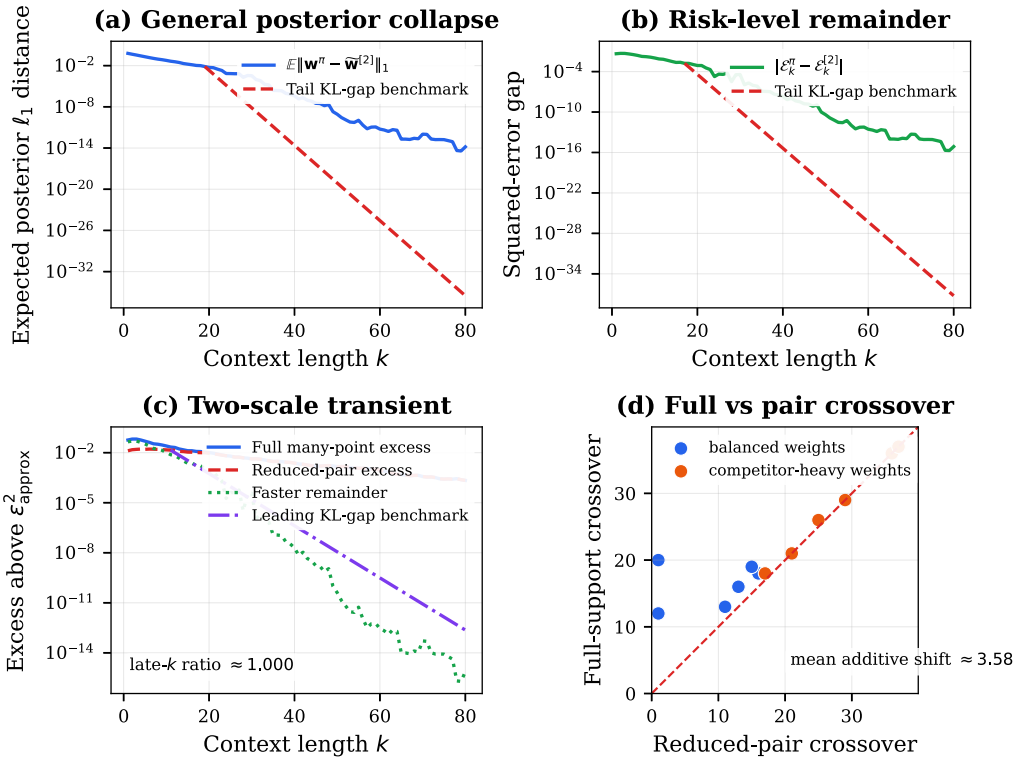


Figure 16: **Nearest-competitor reduction.** (a) Posterior collapse. (b) Risk remainder. (c) Two-scale decomposition. (d) Crossover tracking.

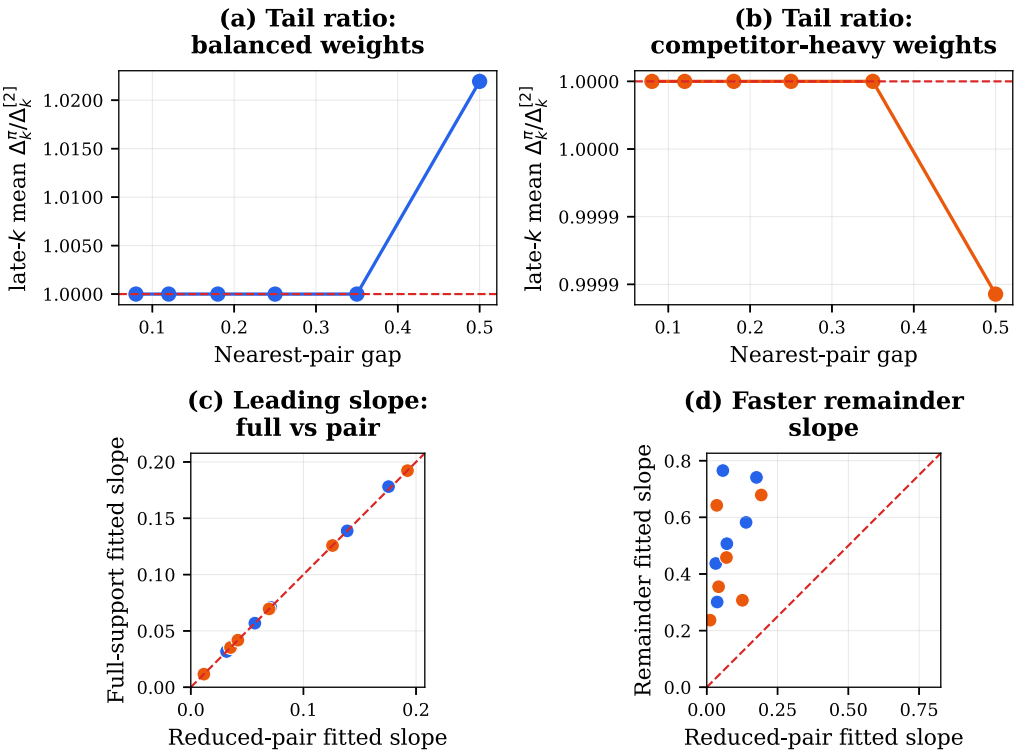


Figure 17: **Family-level reduction.** (a,b) Tail ratios. (c) Leading slopes. (d) Scale separation.

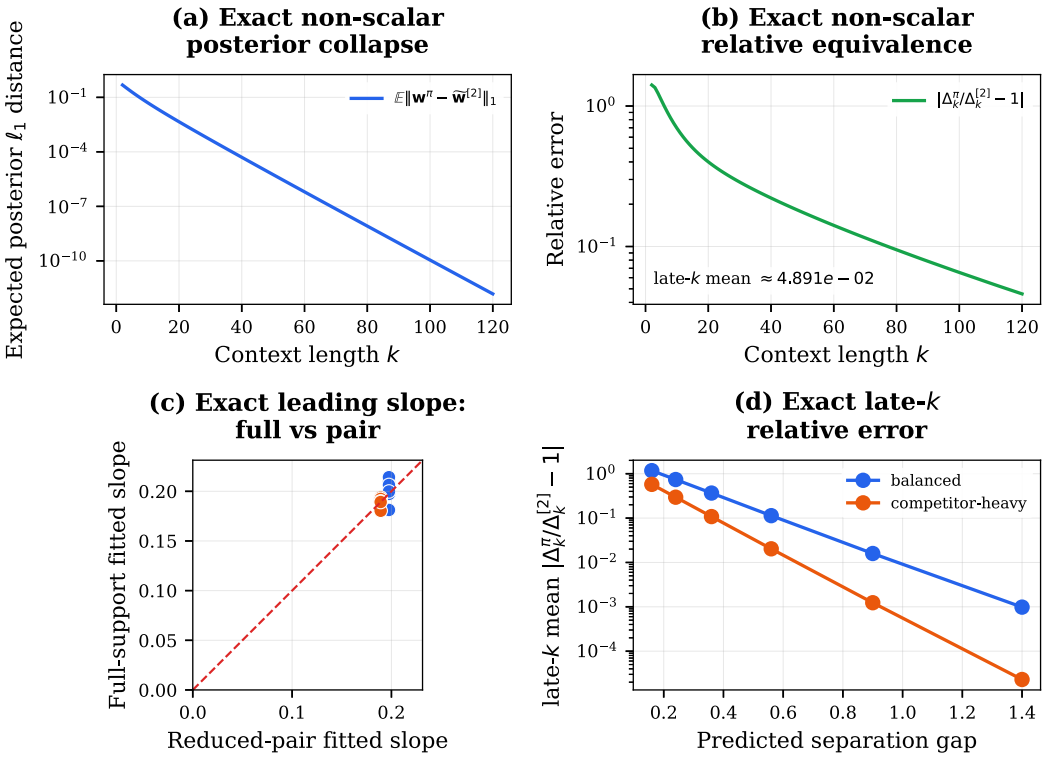


Figure 18: **Exact-quadrature reduction.** (a) Posterior collapse. (b) Relative equivalence. (c) Slopes. (d) Separation.

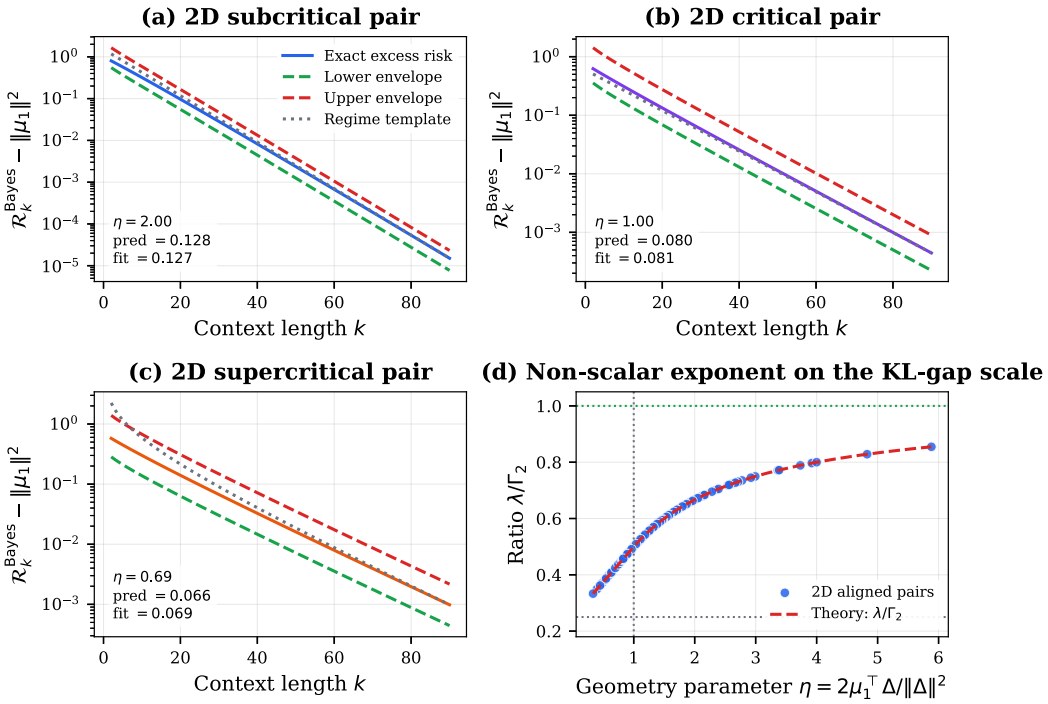


Figure 19: **Non-scalar lower bound.** (a)–(c) Three regimes. (d) λ/Γ_2 vs η.

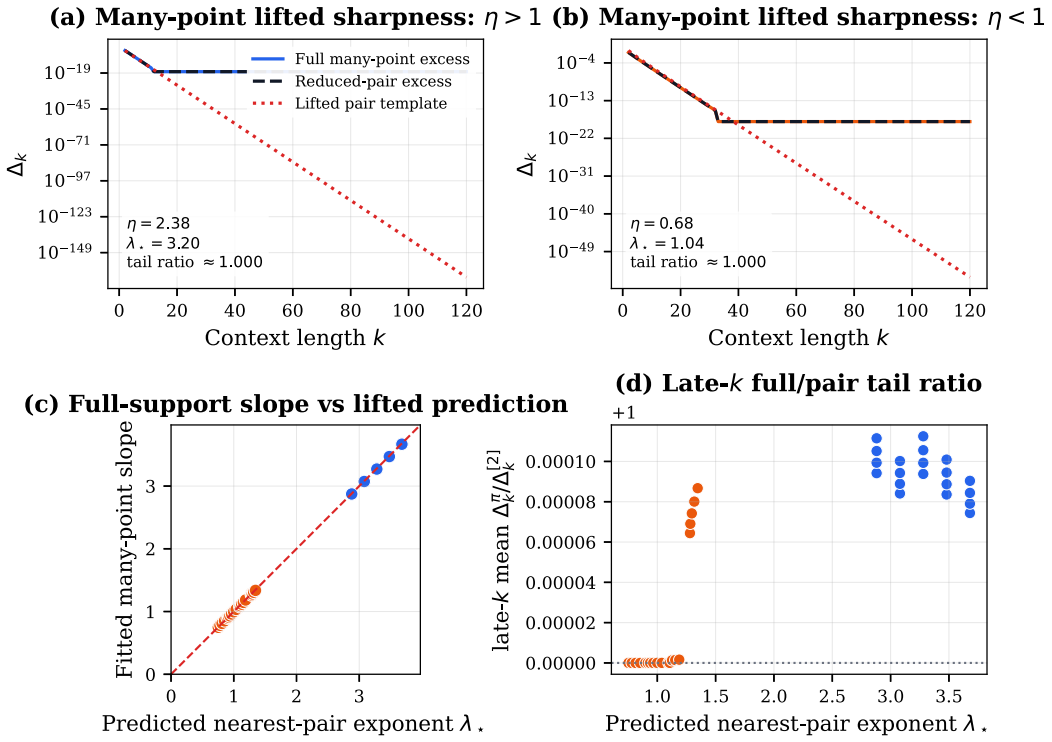


Figure 20: **Many-point diagnostics.** (a,b) Pair/full-model comparison. (c) Excess-risk slopes. (d) Tail ratio. The theoretical pair lower-bound transfer requires the rate separation stated in Corollary 14; a unique nearest mean KL gap alone does not establish it.

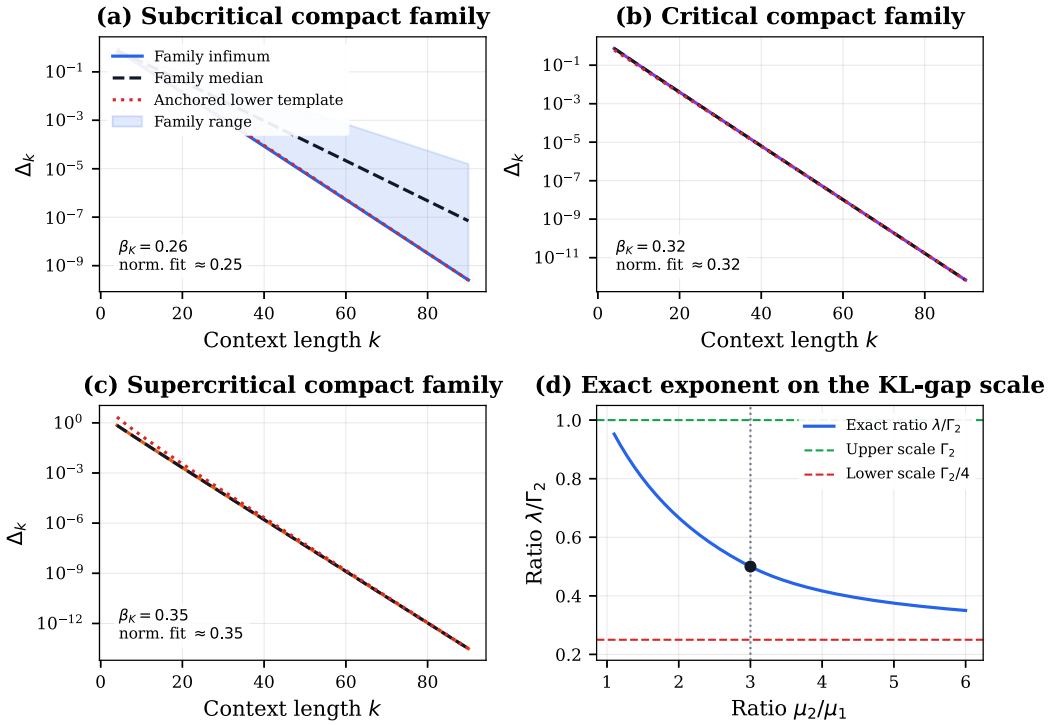


Figure 21: **Scalar lower bound.** (a)–(c) Three regimes. (d) KL ratio.

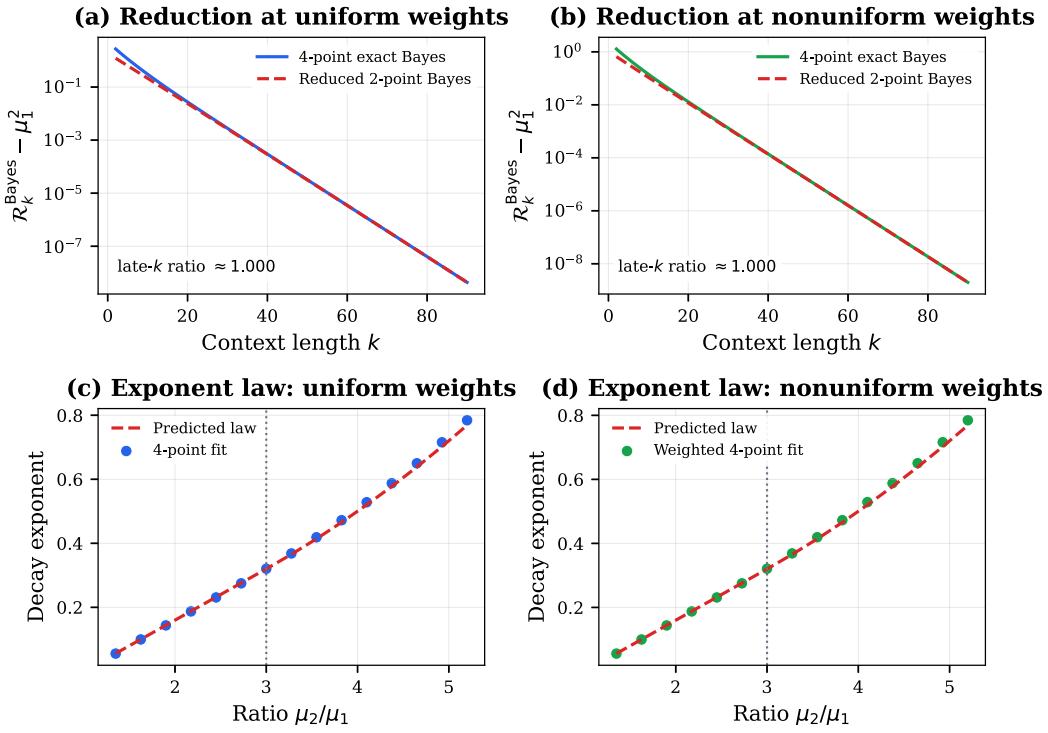


Figure 22: **Finite-support reduction.** (a,b) Collapse. (c,d) Exponents.

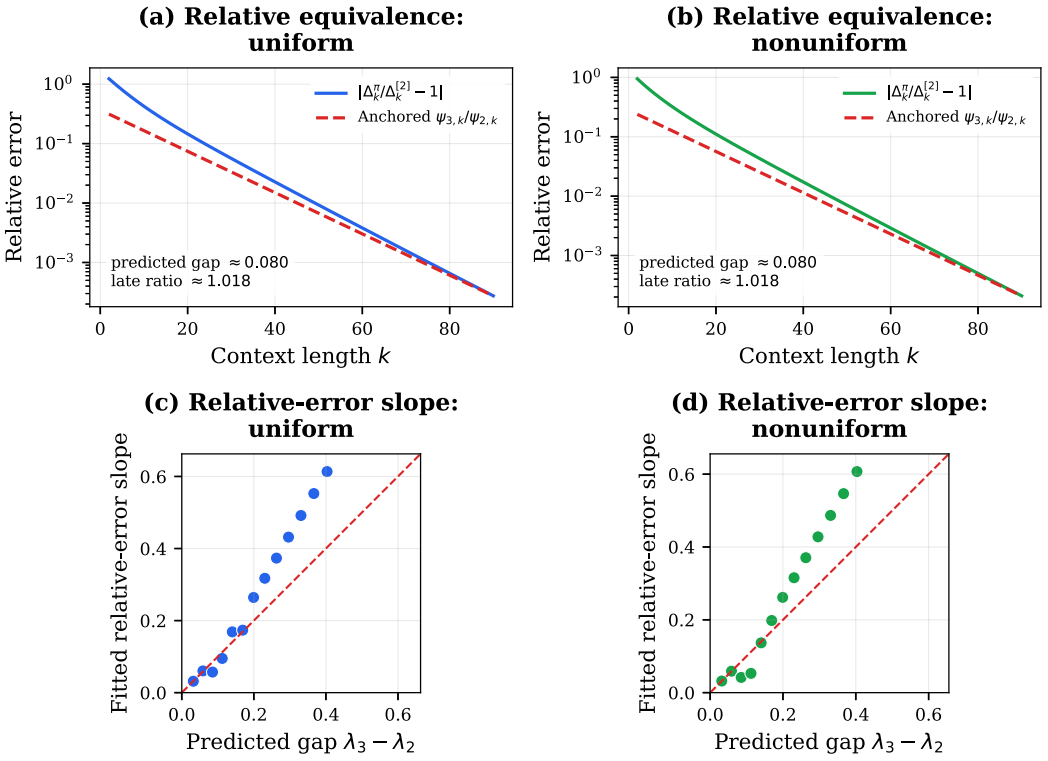


Figure 23: **Relative equivalence.** (a,b) Error. (c,d) Slopes.

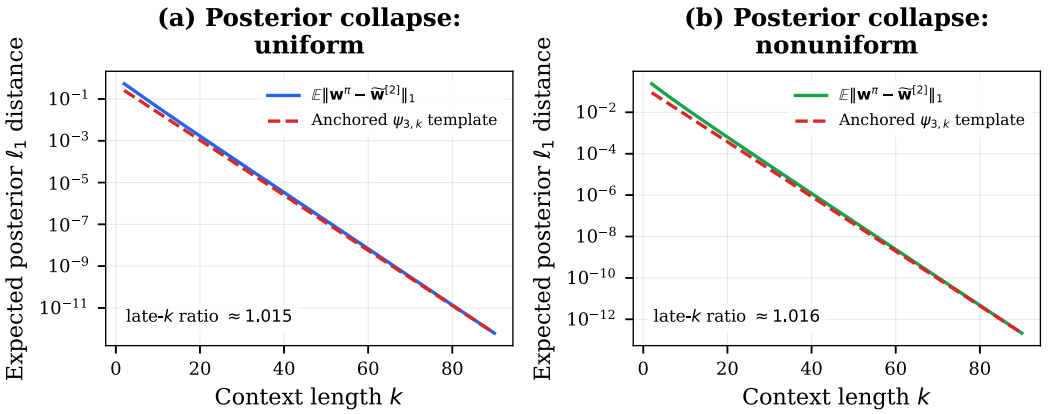


Figure 24: **Posterior collapse.** (a,b) ℓ_1 decay.

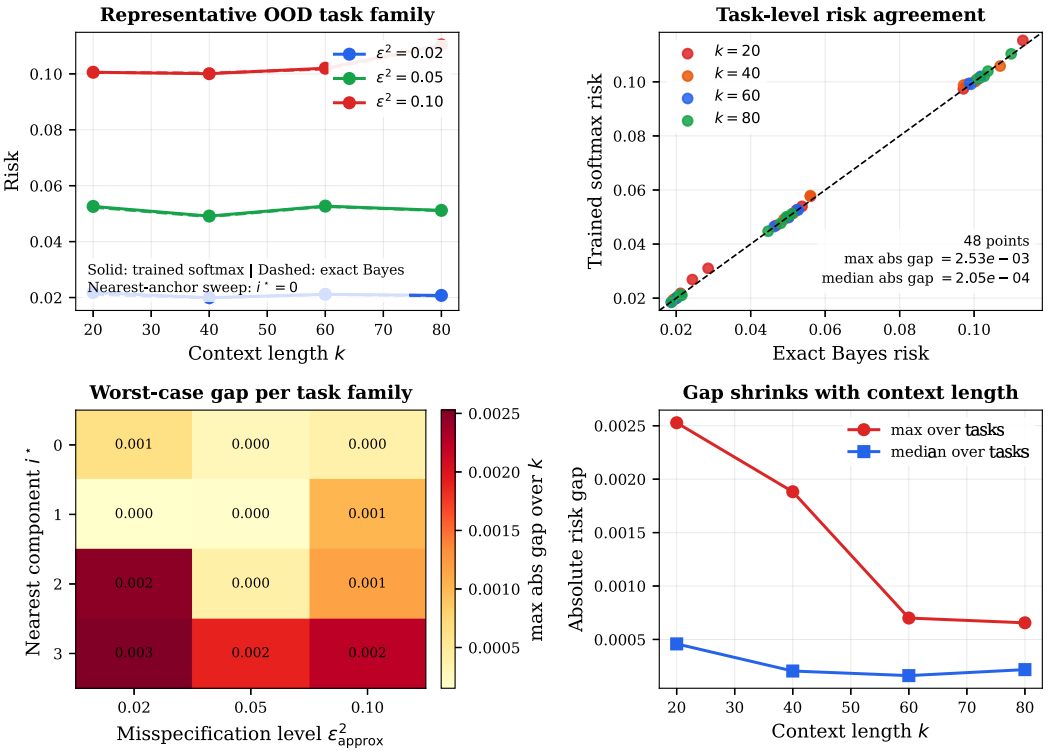


Figure 25: **OOD bridge.** (a) Curves. (b) Scatter. (c) Worst gap. (d) Gap vs k.

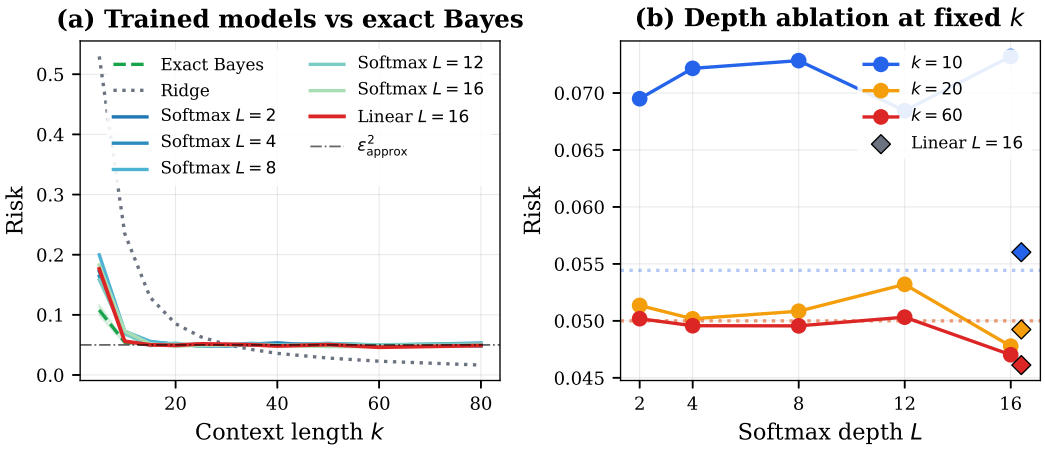


Figure 26: **Architecture.** (a) Model comparison. (b) Depth.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Introduction state the exact benchmark claim, the four evidence levels, the principal empirical effects, and the non-identification boundary.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 7 discusses finite-support assumptions, ties, large support, compositional escape, prompt/model coverage, and the distinction between behavioral diagnostics and internal mechanisms.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Sections 3–4 state the assumptions. Appendices provide the predictor-risk, Gaussian, scalar many-point, OLS, and fixed-subspace proofs. Appendix K includes a complete compact-domain, fixed-length continuous-task construction; no unsupported quantitative architecture bound is retained.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [No]

Justification: Section 5 and Appendix M specify the interventions, estimands and statistical analysis. The supplement permits recalculation from stored predictions. End-to-end reproduction of all reported model experiments is not fully documented; some historical training overrides remain unavailable.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The supplement provides evaluation inputs, prediction records, configurations, analysis code and saved-record reproduction commands. Coverage of fresh model inference and historical training is partial and documented in the release.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [No]

   Justification: Evaluation settings and recorded central training parameters appear in Section 5, Appendix M, and the released configurations. Saved run values are distinguished from source defaults; exact historical overrides are incomplete for some trained runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Primary E1 and its text-quality contrast use paired seed-level bootstrap intervals; E2–E4 use their stated paired hierarchical intervals, additionally clustering the

five-subject panel by subject. Response-stage summaries use the normal-reference calculation specified in Appendix M.10. The variability units and pairing are stated in Section 5 and Appendix M.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Appendix M records the eight-RTX-4090 compute server. The supplement preserves available environment and legacy run records, but per-experiment wall-clock totals and a full project compute audit, including preliminary runs, were not retained.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses public datasets and publicly released models for a low-risk methodological study and follows the NeurIPS Code of Ethics; see Section 8.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 8 discusses the potential reliability benefit of detecting support-sensitive failures and the risks of interpretive overreach or inappropriate use in high-stakes settings.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk model or dataset. The accompanying package contains research code and aggregate artifacts for controlled diagnostics.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Appendix M cites model and benchmark releases. The included evaluation panels and analysis routines are distributed with the supplementary material; third-party model weights and benchmark downloads are not redistributed.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The supplement documents the author-owned reproduction code, its MIT license, inputs, dependencies, stage commands, protocol configurations and known completeness limits. It does not release a new pretrained model or externally sourced benchmark dataset. The paper PDF contains all written theoretical appendices.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The study does not use crowdsourcing or human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The study does not involve human subjects and therefore does not require IRB approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Qwen, Llama, GPT-2, and Gemma models are experimental systems in the direct-LM diagnostics. Their roles, prompts, scoring or generation procedures, and interpretation boundaries are described in Section 5 and Appendix M.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.