The Price of Choosing Examples: Context Overfitting in Adaptive In-Context Learning

Xiaoyu Li[*]
University of New South Wales
xiaoyu.li2@unsw.edu.au

Jiangxuan Long[*]
The University of Hong Kong
jxlong@connect.hku.hk

Abstract

In-context learning is usually analyzed as if the examples in the prompt were sampled before the learner is chosen. In deployed systems they are not: examples are retrieved, filtered, re-ranked, edited, and repeatedly tested after seeing the query and the data store. We formalize this gap by treating prompt construction as adaptive data analysis through a conditional stochastic-kernel calculus. Our main quantity is the *context capacity*, the mutual information or approximate max-information between the data store and the final prompt transcript. For any frozen in-context learner and any adaptive context pipeline with capacity κ, we prove a finite-sample context generalization bound of order $\sqrt{\kappa/k}$, where k is the number of selected examples whose losses form the reported score. We prove a matching lower bound, whose constant is exact when $\log k \ll \kappa \ll k$, and derive two PAC-Bayesian retrieval rules: a first-order Gibbs rule that charges data-dependent candidate gates their prior mass, and a self-normalized second-order rule. On Banking77 with 64 retrieval recipes and two model families, the certificates hold in all 24,000 resampled selection events, and from 64 (Qwen2.5-7B) or 128 (GPT-4.1-mini) reused validation queries the Gibbs rule certifies a risk below the zero-shot error in over 99% of validation draws. Finally, we instantiate the theory for Bayesian linear in-context regression, including an order-statistics analysis of query-aligned top-k Gaussian retrieval. The results separate context quality from context certification: longer or more relevant contexts improve risk only when the procedure used to choose them is stable enough for selected-context evidence to transfer.

1 Introduction

Large language models and tabular foundation models built on attention architectures [41] can adapt at inference time from examples supplied in the prompt [7, 19]. A growing theory explains when transformers can implement least squares, gradient descent, ridge regression, Bayesian inference, or related meta-algorithms in context [1, 5, 21, 28, 29, 42]. This line of work typically starts from a clean statistical abstraction: a prompt is a sample from a task distribution, possibly corrupted by a fixed retrieval noise model [20].

The clean abstraction is not how prompts are built in practice. Retrieval-augmented generation, few-shot prompting, data-centric prompt engineering, and agentic workflows select demonstrations after seeing a query and a large data store. The selected examples may have passed through embedding search, label filters, re-rankers, hand edits, repeated validation, and model-in-the-loop retries. The context is therefore an adaptive statistic of the same data on which it is judged. This creates a distinct failure mode that we call *context overfitting*: the prompt looks good on the examples or

[*]Equal contribution.

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

Analysis lens	Random object	Complexity term	What is missing
Uniform convergence [39, 40]	fixed hypothesis class	$\log \lvert\mathcal{H}\rvert$ or VC dimension	no query-dependent prompt transcript
PAC-Bayes [10, 32]	posterior over hypotheses	$\mathrm{KL}(Q\Vert\Pi)$	no explicit selected-context sample size
Adaptive data analysis [14, 36]	adaptive statistical queries	information per query	no prompt examples or frozen ICL map
Fixed-retrieval ICL theory [19, 20]	retrieved context distribution	retrieval noise / task mismatch	no reuse of validation data to choose retrieval
This paper	prompt-construction kernel $K_u(dt \mid S)$	$\mathrm{I}(T; S \mid U)$ or profiled KL	separates context quality from context certification

Table 1: Positioning relative to adjacent theory. The contribution is not a new concentration inequality in isolation, but a prompt-level statistical object whose complexity composes across retrieval, reranking, validation, and model-in-the-loop editing.

validation interactions used to select it, but its deployment risk is worse than the selected-context score suggests.

What is the statistical price of choosing prompt examples after seeing the data?

Our answer is deliberately architecture-agnostic. We model the transformer, language model, or tabular foundation model as a frozen map from a prompt transcript to a loss function. All adaptivity is charged to the prompt construction pipeline. The resulting theory does not explain why a transformer implements a good inner learner; instead, it explains when the empirical evidence used to choose or evaluate a context remains statistically valid. We do not claim that adaptive retrieval is bad for prediction. Good retrieval can reduce approximation error and improve the deployed predictor. The claim is narrower: selected-context performance is not automatically valid evidence of deployment risk unless the selection process is stable or the final prompt is evaluated on fresh data. The diagnosis, the $\sqrt{\kappa/k}$ price, covers every adaptive pipeline; certificates cover pipelines whose data-dependent choices are logged, finite, or randomized, and Theorem 4.4 and Remark 4.6 show that some such restriction is necessary.

Contributions. We contribute four pieces. First, we introduce *complete prompt transcripts* and *context capacity*, an information measure for the data-dependent choices used to construct a prompt. Second, we prove expected and high-probability selected-context validity bounds with leading rate $\sqrt{\mathrm{I}(T; S \mid U)/k}$, and show the rate is sharp, with an exact constant when $\log k \ll \kappa \ll k$ (Theorems 4.1, 4.2, and 4.4); the denominator is the selected context length k, not the store size n. Third, we extend the calculus to weighted contexts, clustered stores, latent tasks, multi-round prompt editing, and PAC-Bayesian randomized retrieval, whose certificate charges a data-dependent candidate gate its prior mass (Theorem 6.1) and has a variance-sensitive form (Theorem 6.2). Fourth, we instantiate the penalty in Bayesian linear in-context regression and test the certificates on a real retrieval pipeline and in synthetic and API-backed LLM stress tests. Appendix G names the source of each proof tool.

Section 3 defines transcripts and capacity, Section 4 proves the bounds and turns capacity into accounting rules, Section 5 tests them, Section 6 derives SCAR, and Section 7 treats linear regression. Appendix A collects notation and the assumptions in practice, and Appendix D the proofs.

2 Related work

Mechanisms of in-context learning. Garg et al. [19] showed that transformers trained from scratch can in-context learn simple function classes. Several works argue that trained attention layers can implement optimization or classical estimators in their activations [1, 21, 42]. Recent solvable models give sharp asymptotics for linear attention, task diversity, task alignment, and scaling laws [5, 28, 29]. Our results are complementary: they apply after the in-context learner is fixed and ask whether the chosen context statistically supports the claimed risk.

Retrieval and noisy contexts. RAG theory has begun to model retrieved documents as noisy or query-dependent in-context examples. Guo et al. [20] derive finite-sample risk bounds for retrieval-augmented generation in an in-context linear-regression model. Our model is more agnostic about the frozen predictor and focuses on adaptivity of the context pipeline itself: re-ranking, validation reuse, and model-in-the-loop search may leak information even if each retrieved item is individually relevant.

Adaptive data analysis and information-theoretic generalization. Adaptive data analysis studies how statistics remain valid when queries are selected after observing previous answers [2, 14, 15]. Mutual information bounds control selection bias in adaptive analysis [36] and learning algorithms [46]. PAC-Bayes provides another route to data-dependent, nonuniform generalization [10, 13, 32]. We translate these tools to prompt construction and isolate the dependence on context length.

Post-selection inference and the winner's curse. Reporting the best of many noisy estimates biases the report upward, the winner's curse [48]; Theorem 4.4 is a winner's-curse statement for prompts. Simultaneous inference pays a constant valid for every selection rule [3], the role κ plays here in nats, while selective inference conditions on the selection event [18, 27]. Reality checks [44], leaderboard reuse [4], and max-information corrections of p-values [35] treat related reuse problems. Our object is the selected score itself.

The distinction from applying existing bounds verbatim is the object being certified. If one treats the final prompt as just another data-dependent hypothesis, a mutual-information bound can charge $\mathrm{I}(T;S)$, but it does not by itself specify which empirical average is being certified, why the denominator is the selected context length k rather than the store size n, or how to account for selected indices, repeated examples, clustered documents, and multi-round prompt edits. The transcript-kernel formulation below makes those choices measurable and exposes the selected-context score as a ghost-sample statistic.

3 Setup

All measurable spaces are assumed to be standard Borel, so the stochastic kernels and regular conditional laws below exist. We use the word *transcript* in a strong statistical sense: it is not merely the visible prompt, but the full data-dependent object needed to replay the final predictor and audit how it was selected.

Let $U \in \mathcal{U}$ denote a query, task descriptor, or user request. Conditional on $U = u$, a data store $S = (Z_1, \dots, Z_n) \sim P_u^n$ is available to a prompt construction pipeline. A fresh deployment item is $Z_\star \sim P_u$, independent of S given u. The store elements may be input-label pairs, documents, demonstrations, or tool traces.

An adaptive context pipeline M observes (S, u) and outputs a transcript $T \in \mathcal{T}$. The transcript includes an ordered set of distinct selected indices $I(T) = (i_1(T), \dots, i_k(T)) \in [n]^k$ as well as any prompt text, instructions, re-ranking scores, and random seeds needed to define the frozen in-context predictor. If the visible prompt contains selected examples or labels, those values are part of T; this may increase capacity, but it keeps the statistical object replayable. For each fixed transcript t, the frozen predictor induces a loss function $\ell_t : \mathcal{Z} \to \mathbb{R}$. Write $L_u(t) := \mathbb{E}_{Z \sim P_u} \ell_t(Z)$, $\widehat{L}_S(t) := k^{-1} \sum_{j=1}^{k} \ell_t(Z_{i_j(t)})$, and $\mathrm{gap}_u(t, S) := L_u(t) - \widehat{L}_S(t)$. The target is selected-context validity rather than ordinary test accuracy: a holdout split has zero selection capacity only if it is never used to choose, edit, or stop the prompt-construction process. When $\ell_t \in [0, 1]$, Lemma D.2 gives $\sigma^2 = 1/4$ in the sub-Gaussian assumptions below.

Running example. Try M prompt templates on the same k validation examples and keep the template with the lowest selected-context loss. The transcript contains the chosen template id and selected indices, so its capacity is at most $\log M$ when the candidate set is fixed. Even if every template has the same population risk, the selected validation score is optimistically biased at scale $\sqrt{\log M / k}$; this is the basic effect formalized and sharpened below.

Definition 3.1 (Complete and admissible transcript)**.** *A transcript T is complete if, conditional on $(T, U) = (t, u)$, every downstream data-dependent object used by the frozen predictor or its evaluation is a measurable function of (t, u). This includes selected indices, prompt text, retrieved*

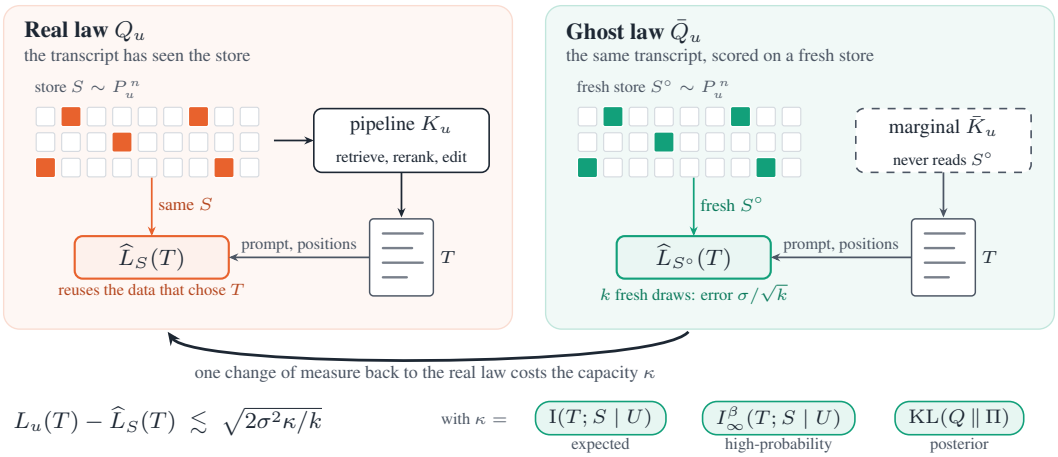

Figure 1: Coupled (real) and decoupled (ghost) laws of store and transcript. The ghost law scores the same transcript on a fresh store, so its selected-context score averages k fresh losses; one change of measure back to the real law costs the context capacity κ.

examples, random seeds, model critiques, and edit histories; validation scores used only to select need not be recorded, and recording them adds their release cost (Table 2). A pipeline is admissible if it outputs a complete transcript, the selected indices $i_j(t)$ are measurable, and $\ell_t(z)$ is jointly measurable in (t, z) and integrable under the corresponding deployment law.

Definition 3.2 (Context-valid certificate)**.** *For a class of admissible pipelines, a random variable $B(T, S, U, \delta)$ is a one-sided context-valid certificate at level δ if*

$$\mathbb{P}\Big(L_U(T) \le \widehat{L}_S(T) + B(T, S, U, \delta)\Big) \ge 1 - \delta.$$

Theorems 4.2, 6.1, and 6.2 construct such certificates under different capacity accounting schemes; the last two do so for a transcript drawn from a posterior Q, with both sides averaged over Q.

Definition 3.3 (Selector kernel and decoupled law)**.** *For each query u, write the prompt construction pipeline as a stochastic kernel $K_u(dt \mid s)$, $s \in \mathcal{Z}^n$, $t \in \mathcal{T}$. The coupled law is $Q_u(ds, dt) = P_u^n(ds)K_u(dt \mid s)$, and the decoupled or "ghost" law is*

$$\bar{Q}_u(ds, dt) = P_u^n(ds)\bar{K}_u(dt), \qquad \bar{K}_u(dt) := \int K_u(dt \mid s)\, P_u^n(ds).$$

When $Q_u \ll \bar{Q}_u$, define the pointwise information density

$$\imath_u(t; s) := \log \frac{dQ_u}{d\bar{Q}_u}(s, t) = \log \frac{dK_u(\cdot \mid s)}{d\bar{K}_u}(t).$$

Under the ghost law the transcript is independent of the store it is scored on, so a fixed transcript's selected score (Figure 1) is an average of k fresh losses. Every bound below is proved under the ghost law and transferred back to the coupled law at the price of the following capacity.

Definition 3.4 (Context capacity)**.** *The expected context capacity of a pipeline M is $\mathsf{C}(M) := \mathrm{I}(T; S \mid U)$, where $T = M(S, U)$ includes all randomness and all prompt choices. The high-probability context capacity is the smallest κ such that the joint law of (T, S) and the product law of (T, S'), with $S' \sim P_U^n$ conditionally independent of T given U, satisfy the approximate max-information transfer inequality*

$$\mathbb{P}((T, S) \in E \mid U = u) \le e^\kappa \mathbb{P}((T, S') \in E \mid U = u) + \beta$$

for P_U-almost every u and every measurable event $E \subseteq \mathcal{T} \times \mathcal{Z}^n$. We write this as $I^\beta_\infty(T; S \mid U) \le \kappa$.

Intuitively, the capacity counts the nats about the store that the final transcript can carry: for a choice among M fixed recipes, $\mathrm{I}(T; S \mid U) \le \log M$ and $I^\beta_\infty(T; S \mid U) \le \log(M/\beta)$ (Lemma D.7).

The capacity can be small for randomized, noisy, compressed, or regularized retrievers. It can be as large as $\log \binom{n}{k}$ for deterministic index selection and can be infinite if the transcript reveals arbitrary real-valued data without noise or finite precision. This is not a defect of the definition: such a prompt can indeed memorize the store.

4 Main bounds

4.1 Expected context generalization

Theorem 4.1 (Expected adaptive context bound)**.** *Assume that for every u and every fixed transcript t, the centered loss $\ell_t(Z) - L_u(t)$, $Z \sim P_u$, is σ^2-sub-Gaussian. Then every adaptive context pipeline with k distinct selected examples satisfies*

$$|\mathbb{E}\left[\mathrm{gap}_U(T,S)\right]| \le \sqrt{\frac{2\sigma^2}{k}\,\mathrm{I}(T;S \mid U)}.$$

For losses in $[0,1]$, this becomes

$$|\mathbb{E}\left[\mathrm{gap}_U(T,S)\right]| \le \sqrt{\frac{\mathrm{I}(T;S \mid U)}{2k}}.$$

Proof idea. For a fixed transcript the selected indices are constants, so under a fresh store the selected score averages k independent losses and its gap is σ^2/k-sub-Gaussian. The Donsker–Varadhan inequality between the coupled and decoupled laws then charges the dependence of T on S exactly $\mathrm{I}(T;S \mid U)$ (Appendix D.3). This selected-subsample step, with denominator k rather than the store size n, is the only nonstandard step.

Theorem 4.1 says that selected examples count as k statistical samples only after discounting the information used to choose them. With Bernstein curvature the price is an information-tilted Legendre transform of the ghost-context cumulant (Theorem D.19 in Appendix D.4), the context form of the cumulant bounds of Jiao et al. [24] and Bu et al. [8]. A deterministic nearest-neighbor retriever may still be excellent for prediction, but its selected-context score is not automatically a valid estimate of deployment risk.

4.2 High-probability transfer

Theorem 4.2 (High-probability context transfer)**.** *Under the assumptions of Theorem 4.1, suppose $I_\infty^\beta(T;S \mid U) \le \kappa$. Then with probability at least $1-\beta-\delta$,*

$$\mathrm{gap}_U(T,S) \le \sqrt{\frac{2\sigma^2}{k}\left(\kappa + \log\frac{1}{\delta}\right)}.$$

The same bound holds for $-\mathrm{gap}_U(T,S)$.

This statement is useful when a prompt is reused as a certificate: the empirical score of the selected context transfers to a fresh draw from the same task distribution as long as the transcript has bounded approximate max-information. The proof (Appendix D.5) moves the ghost-law tail $\exp\{-k\epsilon^2/(2\sigma^2)\}$ of a fixed transcript to the coupled law at the price e^κ and slack β. Differentially private prompt search, finite-description-length search, and Gibbs retrieval all give routes to controlling κ; for a finite menu, Theorem 6.1 with a point-mass posterior gives the same rate without the slack β (Section 4.6).

4.3 Weighted contexts and latent tasks

Real contexts are rarely simple unweighted lists. Demonstrations can be repeated, truncated, attended to with unequal weights, or ensembled across retrieved passages, and RAG stores contain near-duplicate passages, multiple chunks from the same document, or clustered demonstrations from the same source. The same proof gives an effective-context-length version (Appendix D.6).

Remark 4.3 (Effective context length and task identification). *If the score weights the selected examples by nonnegative $w_j(t)$ with $\sum_j w_j(t) = 1$ and $\sum_j w_j(t)^2 \le 1/k_{\text{eff}}$, Theorem 4.1 holds with k_{eff} in place of k (Proposition D.20): attention concentration or repeated demonstrations reduce the effective evidence even when the displayed context contains many tokens. If the store is partitioned into public blocks that are independent given U, each selected example has law P_u, losses lie in $[0,1]$, and the selected block masses satisfy $\sum_g \pi_g(t)^2 \le 1/k_{\text{blk}}$, its $[0,1]$ form holds with k_{blk} (Proposition D.21): retrieving many chunks from one document certifies like one heavily weighted block, not like many independent examples. If a latent task W generates the store, then $\mathrm{I}(T;S \mid U) = \mathrm{I}(T;W \mid U) + \mathrm{I}(T;S \mid U,W)$, and the gap to the task's own risk is charged only the second term (Proposition D.22). The term $\mathrm{I}(T;W \mid U)$ is task-identification capacity; the second term is the within-task overfitting budget.*

4.4 Sharpness

Theorem 4.4 (Lower bound). *There is a universal constant $c > 0$ such that for every k and every $1 \le \kappa \le k/16$, there exist a distribution P, a finite prompt class with losses in $[0,1]$, and an adaptive context pipeline M with $\mathrm{I}(T;S) \le \kappa$ for which*

$$\mathbb{E}\left[L(T) - \widehat{L}_S(T)\right] \ge c\sqrt{\frac{\kappa}{k}}.$$

If moreover $k \to \infty$ with $\log k = o(\kappa)$ and $\kappa = o(k)$, the same construction has $\mathbb{E}[L(T) - \widehat{L}_S(T)] \ge (1 - o(1))\sqrt{\kappa/(2k)}$, and its gap exceeds $(1 - o(1))\sqrt{\kappa/(2k)}$ with probability tending to one.

Thus the dependence in Theorem 4.1 is minimax optimal up to constants, and in the regime of the second statement even its constant $\sqrt{\kappa/(2k)}$ for losses in $[0,1]$ is exact (Appendix D.7). The lower-bound selector simply searches over about $\exp(\kappa)$ candidate prompts and chooses the one with the most favorable empirical fluctuation, the winner's curse. This is the statistical core of context overfitting. The construction uses a store of size $n = k$; appending store elements that the selector never reads extends it to every $n \ge k$ without changing the capacity or the gap.

4.5 Adaptive composition

Prompt construction is often multi-round: retrieve, re-rank, test, ask the model to critique, edit the instruction, and retrieve again. The capacity tracks this composition exactly.

Proposition 4.5 (Capacity composition). *If $T = (T_1, \dots, T_R)$ is produced in R adaptive rounds, then*

$$\mathrm{I}(T;S \mid U) = \sum_{r=1}^{R} \mathrm{I}(T_r;S \mid U, T_{<r}).$$

Consequently Theorem 4.1 holds with the sum of the per-round capacities.

The proposition (Appendix D.8) gives a bookkeeping rule for prompt engineering. Repeatedly trying prompts on the same validation interactions is safe only when each round spends little conditional information.

The same identity profiles the sub-gamma transport (Corollary D.23): an apparently harmless sequence of small validation probes behaves like a single prompt selector whose complexity is the whole capacity profile, not only the last retrieved context. A score vector released with Gaussian noise of precision ρ spends at most $\rho\Delta^2/2$ nats when it moves by at most Δ between stores, regardless of how many deterministic filters are applied afterward (Proposition D.24).

4.6 Operational accounting

The abstract capacity can be upper bounded from logs produced by standard retrieval systems. Table 2 gives a menu of sufficient accounting rules. The entries are intentionally pessimistic: they are meant to be auditable without reverse-engineering the full retriever. The key audit principle is simple: every data-dependent prompt-construction choice must either be logged as capacity, randomized through a stable posterior, or evaluated on fresh data.

Prompt-construction step	Capacity charge	Required log
Choose one of M prompt templates	$\le \log M$; $\le \log(M/\beta)$ for Theorem 4.2 (Lemma D.7)	template id and candidate set size
Try M context recipes, keep best validation score	$\le \log M$ ($\log(M/\beta)$ for Theorem 4.2) plus score-release cost	recipe ids and validation reuse count
Gibbs/SCAR sampling from prior Π	$\mathbb{E}_S \mathrm{KL}(Q_S \Vert \Pi)$, including candidate-gate mass; bounds I only	posterior weights, sampling temperature, and $\Pi(\mathcal{A})$
Gaussian-noisy score vector	$\le \rho\Delta^2/2$ by Proposition D.24; bounds I only	noise precision and an almost-sure bound Δ on $\Vert a(S) - a(S^\circ)\Vert_2$
R interactive prompt edits	$\sum_{r=1}^{R} C_r$ by Proposition 4.5	per-round transcript and candidate pool
Fresh held-out evaluation after prompt selection	0 for selection certificate	split identifier and no reuse guarantee

Table 2: Practical capacity accounting rules. They turn hidden prompt engineering into auditable statistical metadata.

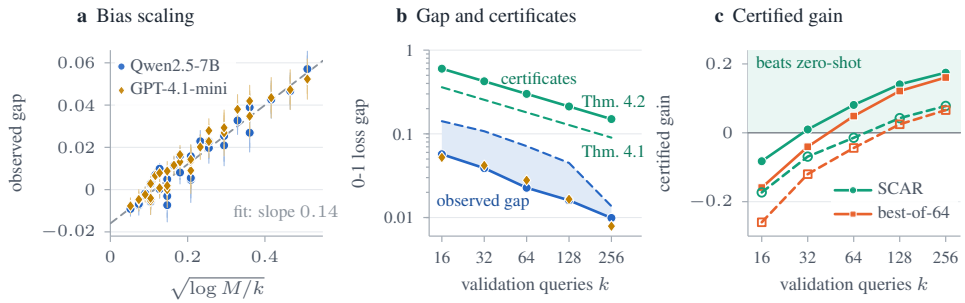

Figure 2: Banking77 recipe selection. (a) Mean gap of the best-of-M recipe in 30 (M, k) cells, with 95% intervals. (b) $M = 64$: mean and 95th-percentile gap (blue; amber: GPT-4.1-mini mean) against Theorems 4.1 (dashed) and 4.2 (solid). (c) Zero-shot error minus certified risk for SCAR and for the best-of-64 winner certified by Theorem 6.1; solid: Qwen, dashed: GPT-4.1-mini.

For example, best-of-M selection that reuses k validation examples with losses in $[0, 1]$ spends at most $\log M$ nats, or $\log(M/\beta)$ in Theorem 4.2. At $M = k = 64$, Theorem 4.1 bounds the expected gap by $\sqrt{\log M/(2k)} = 0.180$, and Theorem 4.2 with $\beta = \delta = 0.025$ certifies the gap at 0.300. Theorem 6.1, applied to a point mass on the winner under the uniform prior ($\mathrm{KL} = \log M$) at the temperature $\lambda^\star$ of Section 6, certifies it at $\sqrt{(\log M + \log(1/\delta))/(2k)} = 0.236$ for $\delta = 0.05$.

The capacity of an unmodified deterministic pipeline can be as large as $\log \binom{n}{k}$, and a surcharge that ignores it is nearly vacuous.

Remark 4.6 (Cherry-picking)**.** *Let $n \ge 2k$, fix one prompt whose loss is Bernoulli(p) under P, and let the pipeline select k store items on which that loss is 0. If $n(1-p) \ge 2k$, then with probability at least $1 - e^{-k/4}$ the selected score is 0 while the risk is p. A surcharge $b(n, k, \delta)$ that is valid at a level $\delta < 1 - e^{-k/4}$ for every $p \le 1 - 2k/n$ is therefore at least $1 - 2k/n$ (Appendix D.11).*

5 Experiments

A real retrieval pipeline. On Banking77 intent classification [9], frozen Qwen2.5-7B-Instruct [47] and GPT-4.1-mini choose among M retrieval-based prompt recipes on k reused validation queries (Appendix B.1). The selection gap follows the $\sqrt{\log M/k}$ scale (Figure 2a), and the certificate of Theorem 4.2 held in all 12,000 selection events per model; on this benign library it is conservative (Figure 2b). SCAR's certificate (Theorem 6.1) is the tightest at every k: at $k = M = 64$ it is 0.292 for Qwen, against 0.324 for the best-of-64 winner certified by Theorem 6.1 with a point mass and 0.388 by Theorem 4.2. From $k = 64$ (Qwen) and $k = 128$ (GPT-4.1-mini) it lies below the zero-shot error in over 99% of validation draws (Figure 2c and Appendix B.1), at a holdout risk within 0.01 of the best-of-64 winner's.

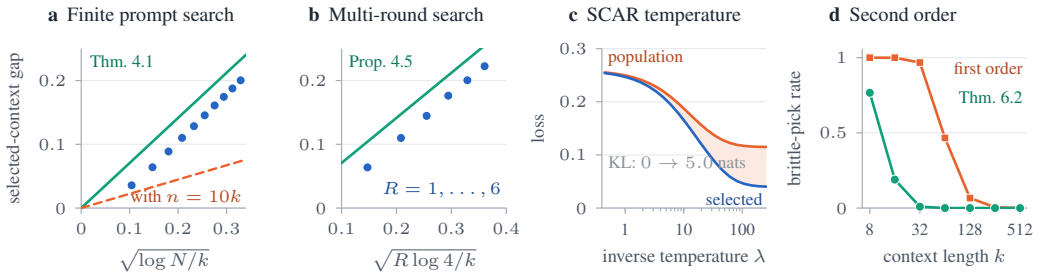

Figure 3: Synthetic context-overfitting diagnostics. (a) Trying N candidate prompts on the same $k = 64$ examples yields the lower-bound scaling; orange: Theorem 4.1 with the store size $n = 10k$ in place of k. (b) An R-round workflow with $M = 4$ edits per round overfits according to the composed capacity $R \log M$. (c) Increasing SCAR's inverse temperature improves selected empirical fit but increases context overfitting (shaded). (d) Rate of brittle selections. Trials per point: 4,000 in (a), 2,500 in (b, c), 3,000 in (d).

Lower-bound scaling. The lower bound predicts that even a finite prompt search can produce a selected-context gap of order $\sqrt{\log N/k}$ when N candidate prompts are tried on the same k examples. Figure 3(a) simulates exactly the construction in Theorem 4.4: each candidate prompt corresponds to one Rademacher coordinate, and the pipeline selects the coordinate with the lowest empirical loss. The observed gap is nearly linear in the predicted scale, which is the behavior our theory charges to context capacity; a bound normalized by the store size would miss it.

Multi-round capacity accumulation. Proposition 4.5 predicts that a branching prompt-editing workflow pays for the entire sequence of choices. Figure 3(b) simulates R rounds in which each round opens $M = 4$ possible prompt edits, yielding M^R terminal prompts. Selecting the best terminal prompt on the same $k = 64$ examples produces a gap that tracks $\sqrt{R \log M/k}$, the composed capacity scale.

SCAR temperature. Figure 3(c) simulates a library of 256 prompt recipes with heterogeneous population risks. A small inverse temperature behaves like a stable prior retriever; a large inverse temperature approximates hard empirical prompt search. As predicted by Theorem 6.1, aggressive temperature lowers selected empirical loss while increasing both the context gap and the KL capacity (Figure 7 in Appendix C).

Second-order certification. Figure 3(d) tests Theorem 6.2 in a brittle prompt library. Most candidate prompts are high-variance: they sometimes look perfect on a short selected context but have poor population risk. The first-order empirical selector repeatedly chooses these brittle prompts. The second-order objective pays for empirical variance and therefore selects stable prompts much more often, especially at small and moderate context lengths: at $k = 32$ it picks a brittle prompt in 0.9% of trials instead of 97%. Its certificate is the tighter one from $k = 64$ (0.446 against 0.509); below that the logarithmic term dominates (Figure 8).

API-backed protocols. The supplementary material also contains OpenAI embedding and real-LLM prompt-search protocols. They instantiate finite-view selection, validation-vs-holdout prompt choice, SCAR temperature sampling, and validation-leakage stress tests on fictional-label tasks; details are in Appendix B. The semantic prompt-selection task is intentionally easy, while the nonce-label prompt-search task isolates the overfitting mechanism with an actual LLM.

6 Stability-constrained adaptive retrieval

The bounds above suggest replacing hard empirical prompt search by a regularized randomized rule. Let Π_u be a data-independent prior over an ambient transcript space. In applications Π_u may encode a base retriever, a distribution over index sets, or a library of prompt recipes. Given a score $\widehat{L}_S(t)$,

define the SCAR posterior $Q_\lambda(dt \mid S,u) \propto \Pi_u(dt)\exp\{-\lambda \widehat{L}_S(t)\}$. Equivalently, Q_λ minimizes $Q \mapsto \mathbb{E}_{t\sim Q}\widehat{L}_S(t) + \lambda^{-1}\mathrm{KL}(Q\|\Pi_u)$. The KL term is the information budget spent by retrieval.

Data-dependent candidate gates are not free: if Q is supported on $\mathcal{A}(S,u)$, the certificate still uses $\mathrm{KL}(Q\|\Pi_u)$, which includes the gate charge $-\log \Pi_u(\mathcal{A}(S,u))$ whenever the conditional prior is used internally.

Algorithm 1: Stability-constrained adaptive retrieval (SCAR).
Input: query u, store S, ambient prior Π_u, optional gate $\mathcal{A}$, selected-context score $\widehat{L}_S(t)$, inverse temperature λ.
Steps:

1. Generate an optional gate $\mathcal{A}(S,u)$; log its size, prior mass, and all data-dependent filters.
2. Define Π_u before reading selected-context scores, using only public metadata or a separate split.
3. Form $Q_\lambda(t \mid S,u) \propto \Pi_u(t)\exp\{-\lambda\widehat{L}_S(t)\}$, restricted to $\mathcal{A}(S,u)$ if a gate is used.
4. Sample a transcript $T \sim Q_\lambda(\cdot \mid S,u)$ afresh for each query, or use the posterior mean prediction when the loss is convex in the prediction.
5. Report $\mathbb{E}_{Q_\lambda}\widehat{L}_S$ together with $\mathrm{KL}(Q_\lambda\|\Pi_u)$, λ, and the effective context length k.

Certificate: Theorem 6.1 converts the logged KL into a risk surcharge; Theorem 6.2 replaces the range surcharge by an empirical-variance surcharge.

Theorem 6.1 (PAC-Bayesian oracle inequality for SCAR)**.** *Assume the sub-Gaussian condition of Theorem 4.1, and let Π_u be fixed before observing the store, except for public metadata or a separate split independent of S. Fix $\lambda > 0$ and $\delta \in (0,1)$. With probability at least $1-\delta$ over $S \sim P_u^n$, simultaneously for all data-dependent posteriors Q over $\mathcal{T}$ that are absolutely continuous with respect to Π_u,*

$$\mathbb{E}_{t\sim Q}L_u(t) \le \mathbb{E}_{t\sim Q}\widehat{L}_S(t) + \frac{\mathrm{KL}(Q\|\Pi_u)+\log(1/\delta)}{\lambda} + \frac{\lambda\sigma^2}{2k}.$$

Moreover, if $\int e^{-\lambda\widehat{L}_S(t)}\,\Pi_u(dt) < \infty$, the unconstrained SCAR posterior obeys the oracle comparison

$$\mathbb{E}_{t\sim Q_\lambda}L_u(t) \le \inf_{Q\ll\Pi_u}\left\{\mathbb{E}_{t\sim Q}\widehat{L}_S(t) + \frac{\mathrm{KL}(Q\|\Pi_u)}{\lambda}\right\} + \frac{\log(1/\delta)}{\lambda} + \frac{\lambda\sigma^2}{2k}.$$

If an optional candidate gate $\mathcal{A}(S,u)$ with $\Pi_u(\mathcal{A}) > 0$ is imposed, the gated posterior $Q_\lambda^{\mathcal{A}} \propto \Pi_u \mathbf{1}_{\mathcal{A}} e^{-\lambda\widehat{L}_S}$ obeys the same comparison over comparators supported on that gate, with all KL terms still measured relative to the ambient prior Π_u.

The theorem (Appendix D.12) has the usual bias-variance interpretation. Large λ aggressively follows empirical prompt performance and risks context overfitting. Small λ stays close to the prior retriever. For a comparator whose KL is at most a budget κ fixed before the store is seen, the surcharge is minimized at $\lambda^\star = \sigma^{-1}\sqrt{2k(\kappa+\log(1/\delta))}$, where it equals $\sqrt{2\sigma^2(\kappa+\log(1/\delta))/k}$, the rate of Theorem 4.2; for a single recipe under a uniform prior on M recipes, $\kappa = \log M$. For losses in $[0,1]$ and the uniform prior on M recipes, SCAR at this temperature has Gibbs risk within $\sqrt{2(\log M+\log(1/\delta))/k}$ of the best recipe's risk with probability at least $1-2\delta$, while every rule that sees only the k reused examples has worst-case expected excess risk at least $0.07\sqrt{\log M/k}$ when $4 \le M \le e^{2k/3}$ (Proposition D.25): certification costs no rate.

Theorem 6.2 (Self-normalized PAC-Bayes for heteroskedastic contexts)**.** *Assume $\ell_t \in [0,1]$ for every transcript. For fixed t, write $X_j(t) := \ell_t(Z_{i_j(t)})$ in the order of the selected indices, let $\widehat{\mu}_{j-1}(t) := (\frac{1}{2} + \sum_{i<j} X_i(t))/j$, and define the predictable empirical variance by*

$$\widehat{V}_S(t) := \frac{1}{k}\sum_{j=1}^{k}\left(X_j(t) - \widehat{\mu}_{j-1}(t)\right)^2.$$

Let Π_u be fixed before observing the store, as in Theorem 6.1, let $\Gamma \subset (0,1)$ be a finite grid fixed in advance, and let $\psi(\gamma) := -\log(1-\gamma) - \gamma$. Then with probability at least $1-\delta$ over $S \sim P_u^n$,

simultaneously for all data-dependent posteriors $Q \ll \Pi_u$ and all $\gamma \in \Gamma$,

$$\mathbb{E}_Q L_u(t) \le \mathbb{E}_Q \widehat{L}_S(t) + \frac{\psi(\gamma)}{\gamma}\,\mathbb{E}_Q \widehat{V}_S(t) + \frac{\mathrm{KL}(Q\|\Pi_u) + \log(|\Gamma|/\delta)}{\gamma k}.$$

Consequently the second-order retrieval posterior obtained by minimizing the right-hand side can certify low-risk prompts whose selected losses have small empirical variance even when the worst-case range bound is pessimistic.

Since $\psi(\gamma)/\gamma = \gamma/2 + O(\gamma^2)$, for large k a grid point near the optimal γ makes the surcharge, to leading order, at most 1.07 times $\sqrt{2\,\mathbb{E}_Q \widehat{V}_S(t)\,(\mathrm{KL}(Q\|\Pi_u) + \log(|\Gamma|/\delta))/k}$, which beats the range proxy $\sigma^2 = 1/4$ when the losses vary little; the dyadic grid $\Gamma_k = \{0.9 \cdot 2^{-i} : 0 \le i \le \lceil \log_2 k \rceil\}$ costs only $\log|\Gamma_k| = O(\log\log k)$ (Appendix D.14).

Remark 6.3 (Why the second-order term matters)**.** *Two prompt recipes can have the same empirical average and the same KL complexity while behaving very differently on the selected context: one may fail on a few brittle examples and another may be uniformly competent. Theorem 6.2 distinguishes them through $\mathbb{E}_Q \widehat{V}_S(t)$, which is invisible in the first-order SCAR objective.*

7 Application: Bayesian linear in-context regression

This section separates two effects that are often conflated. Query-aligned retrieval can improve prediction by increasing posterior precision in the relevant direction, while adaptive reuse of the same store can still invalidate the selected-context score as a certificate. Let $\theta \sim N(0, \tau^2 I_d)$, $x_i \sim N(0, I_d)$, $y_i = x_i^\top \theta + \xi_i$, and $\xi_i \sim N(0, \sigma_\xi^2)$, and consider a query $x_\star$. Given selected contexts (X_I, y_I), the Bayes predictor is the posterior mean, equivalent to ridge regression:

$$\widehat{y}_\star = x_\star^\top \left(X_I^\top X_I + \frac{\sigma_\xi^2}{\tau^2} I_d \right)^{-1} X_I^\top y_I.$$

Proposition 7.1 (Conditional Bayes risk)**.** *For any fixed selected design X_I,*

$$\mathbb{E}\left[(y_\star - \widehat{y}_\star)^2 \mid X_I, x_\star\right] = \sigma_\xi^2 + x_\star^\top \left(\tau^{-2} I_d + \sigma_\xi^{-2} X_I^\top X_I\right)^{-1} x_\star.$$

If $u = x_\star/\|x_\star\|$ and $k^{-1} X_I^\top X_I \succeq \alpha u u^\top + \beta(I_d - u u^\top)$ for $\alpha, \beta \ge 0$, then

$$\mathbb{E}\left[(y_\star - \widehat{y}_\star)^2 \mid X_I, x_\star\right] \le \sigma_\xi^2 + \frac{\|x_\star\|^2}{\tau^{-2} + k\alpha/\sigma_\xi^2}.$$

Appendix D.15 proves the proposition and writes this risk exactly through the Schur complement of the query direction (Proposition D.28); the last term of that complement is the nuisance-coordinate leakage penalty induced by selecting examples that are aligned with the query direction but not orthogonal to irrelevant directions. For top-k Gaussian retrieval by $(u^\top x_i)^2$, the selected directional mass is at least the upper χ_1^2 quantile $q_{n,k} = F^{-1}_{\chi_1^2}(1 - 2k/n)$ with probability at least $1 - e^{-k/4}$ (Corollary D.29).

These results formalize why query-aligned retrieval can help: examples with large design mass in the query direction reduce posterior uncertainty, while the Schur-complement term measures how much of that gain is lost to nuisance coordinates. The main bounds explain a different issue. If the retrieval rule, the similarity metric, or the number of retrieved examples is selected on the same store or validation interactions, the observed prompt score must still pay the capacity term. Good retrieval reduces approximation error; stable retrieval prevents statistical overclaiming.

8 Conclusion

In-context learning is not only about the model inside the prompt; it is also about the process that chose the prompt. We showed that adaptive context selection has a sharp information-theoretic price of order $\sqrt{\kappa/k}$. This price composes across prompt-engineering rounds, is unavoidable in the worst case, and leads naturally to stability-regularized retrieval. The theory suggests a simple reporting norm for RAG and few-shot learning: along with context length and retrieval accuracy, report how much information the context construction pipeline was allowed to spend.

Acknowledgments

We thank the reviewers and the area chair for their helpful comments. This work received no external funding. The authors declare no competing interests.

References

[1] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? Investigations with linear models. In *International Conference on Learning Representations*, 2023.

[2] Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*, pages 1046–1059, 2016. doi: 10.1145/2897518.2897566.

[3] Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013. doi: 10.1214/12-AOS1077.

[4] Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1006–1014, 2015.

[5] Blake Bordelon, Mary I. Letey, and Cengiz Pehlevan. Theory of scaling laws for in-context regression: Depth, width, context and time. In *International Conference on Learning Representations*, 2026.

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

[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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.

[8] Yuheng Bu, Shaofeng Zou, and Venugopal V. Veeravalli. Tightening mutual information-based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 1(1):121–130, 2020. doi: 10.1109/JSAIT.2020.2991139.

[9] Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.nlp4convai-1.5.

[10] Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *Institute of Mathematical Statistics Lecture Notes – Monograph Series*. Institute of Mathematical Statistics, 2007. doi: 10.1214/074921707000000391.

[11] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, second edition, 2006. doi: 10.1002/047174882X.

[12] Monroe D. Donsker and S. R. Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, I. *Communications on Pure and Applied Mathematics*, 28 (1):1–47, 1975. doi: 10.1002/cpa.3160280102.

[13] Benjamin Dupuis, Paul Viallard, George Deligiannidis, and Umut Simsekli. Uniform generalization bounds on data-dependent hypothesis sets via PAC-Bayesian theory on random sets. *Journal of Machine Learning Research*, 25(409):1–55, 2024.

[14] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[15] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248): 636–638, 2015. doi: 10.1126/science.aaa9375.

[16] Xiequan Fan, Ion Grama, and Quansheng Liu. Exponential inequalities for martingales with applications. *Electronic Journal of Probability*, 20(1):1–22, 2015. doi: 10.1214/EJP.v20-3496.

[17] Vitaly Feldman and Thomas Steinke. Calibrating noise to variance in adaptive data analysis. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 535–544, 2018.

[18] William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

[19] 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. doi: 10.52202/068431-2217.

[20] Yang Guo, Yutian Tao, Yifei Ming, Robert D. Nowak, and Yingyu Liang. Retrieval-augmented generation as noisy in-context learning: A unified theory and risk bounds. *arXiv preprint arXiv:2506.03100*, 2025.

[21] Jianliang He, Xintian Pan, Siyu Chen, and Zhuoran Yang. In-context linear regression demystified: Training dynamics and mechanistic interpretability of multi-head softmax attention. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 22686–22742, 2025.

[22] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

[23] Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

[24] Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Dependence measures bounding the exploration bias for general measurements. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1475–1479, 2017. doi: 10.1109/ISIT.2017.8006774.

[25] Sharu Theresa Jose and Osvaldo Simeone. Information-theoretic generalization bounds for meta-learning and applications. *Entropy*, 23(1):126, 2021. doi: 10.3390/e23010126.

[26] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626. ACM, 2023. doi: 10.1145/3600006.3613165.

[27] Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016. doi: 10.1214/15-AOS1371.

[28] Mary I. Letey, Jacob A. Zavatone-Veth, Yue M. Lu, and Cengiz Pehlevan. Pretrain–test task alignment governs generalization in in-context learning. In *International Conference on Learning Representations*, 2026.

[29] Yue M. Lu, Mary I. Letey, Jacob A. Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *Proceedings of the National Academy of Sciences*, 122(28):e2502599122, 2025. doi: 10.1073/pnas.2502599122.

[30] Jiří Matoušek and Jan Vondrák. The probabilistic method: Lecture notes. Lecture notes, Department of Applied Mathematics, Charles University, Prague, 2008. URL https://www.cs.cmu.edu/~15850/handouts/matousek-vondrak-prob-ln.pdf. Revision of March 2008.

[31] Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

[32] David A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pages 164–170, 1999. doi: 10.1145/307400.307435.

[33] Zakaria Mhammedi, Peter Grünwald, and Benjamin Guedj. PAC-Bayes un-expected Bernstein inequality. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[34] Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2025. doi: 10.1017/9781108966351.

[35] Ryan Rogers, Aaron Roth, Adam Smith, and Om Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 487–494, 2016. doi: 10.1109/FOCS.2016.59.

[36] Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1232–1240, 2016.

[37] Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3437–3452, 2020.

[38] Ilya O. Tolstikhin and Yevgeny Seldin. PAC-Bayes-empirical-Bernstein inequality. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

[39] Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984. doi: 10.1145/1968.1972.

[40] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971. doi: 10.1137/1116025.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[42] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao 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 of *Proceedings of Machine Learning Research*, pages 35151–35174, 2023.

[43] Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009.

[44] Halbert White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000. doi: 10.1111/1468-0262.00152.

[45] Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 641–649. ACM, 2024. doi: 10.1145/3626772.3657878.

[46] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[47] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

[48] Sebastian Zöllner and Jonathan K. Pritchard. Overcoming the winner's curse: Estimating penetrance parameters from case-control data. *The American Journal of Human Genetics*, 80 (4):605–615, 2007. doi: 10.1086/512821.

A Operational Diagnostics and Assumptions

A.1 Notation

Symbol	Meaning
U, u	query, task descriptor, or user request
$S = (Z_1, \dots, Z_n) \sim P_u^n$; $Z_\star \sim P_u$	data store given $U = u$; fresh deployment item
T, $t \in \mathcal{T}$; $K_u(dt \mid s)$	complete transcript (Definition 3.1); pipeline kernel (Definition 3.3)
Q_u, $\bar{Q}_u = P_u^n \otimes \bar{K}_u$	coupled law and decoupled (ghost) law of store and transcript
$i_1(t), \dots, i_k(t)$; k	distinct selected indices; context length
ℓ_t, $L_u(t)$, $\widehat{L}_S(t)$	loss of the frozen predictor, its risk, and its selected-context score
$\mathrm{gap}_u(t, S) = L_u(t) - \widehat{L}_S(t)$	selected-context gap; σ^2 is its sub-Gaussian proxy ($1/4$ for losses in $[0, 1]$)
$\mathrm{I}(T; S \mid U)$; $I_\infty^\beta(T; S \mid U) \le \kappa$	expected and high-probability context capacity (Definition 3.4)
k_{eff}, k_{blk}	effective context length for weighted contexts and clustered stores
Π_u, Q, Q_λ, λ	prior fixed before the store, posterior, SCAR posterior, inverse temperature
$\mathcal{A}(S, u)$, $\Pi_u(\mathcal{A})$	data-dependent candidate gate and its prior mass
$\widehat{V}_S(t)$, Γ, $\psi(\gamma)$	predictable variance, grid, and $-\log(1-\gamma) - \gamma$ (Theorem 6.2)

Table 3: Notation.

A.2 Implications

Context length is not effective sample size. Increasing k helps only when capacity grows sublinearly in k. A pipeline that searches over exponentially many prompts can spend $\Theta(k)$ bits and erase the apparent benefit of a long context.

Report the retrieval budget. For few-shot and RAG evaluations, the number of attempted prompts, re-ranking candidates, validation passes, and data-dependent edits should be treated as part of the statistical method. Theorems 4.1 and 4.2 convert these choices into a risk surcharge.

Use randomized prompt selection when possible. Soft retrieval, temperature sampling, DP noise, finite-description-length prompt recipes, and PAC-Bayesian priors all reduce context capacity. SCAR gives one implementation that can be layered on top of existing retrievers.

Separate prediction quality from certification quality. A deterministic nearest-neighbor context can be the best context for a particular query. Our results do not argue against it. They argue that selected-context empirical performance is not a certificate unless the selection process is stable or evaluated on fresh data.

Assumption	Why it is used	When it holds in deployment	How to weaken or check it
Conditional i.i.d. store $S \sim P_u^n$	gives a ghost store with clean selected-example concentration	exchangeable items within a task; the strongest idealization for scraped corpora	use Proposition D.21, blocking, clustered bootstrap, or an effective k_{eff}
Frozen in-context learner after transcript T	separates prompt adaptivity from model training	a version-pinned model during the search window	include any model-in-the-loop edits in T
Logged transcript contains all data-dependent choices	makes capacity auditable	retrieval stacks that log candidates, scores, filters, and seeds	store candidate ids, validation scores, filters, and random seeds
Bounded or sub-Gaussian losses	yields finite cumulant envelope	accuracy, exact match, task success ($\sigma^2 = 1/4$)	use robust losses, clipping, or sub-gamma Theorem D.19
Prior retriever Π_u dominates generated candidates	PAC-Bayes KL charges data-dependent gates	a base retriever or recipe library fixed before any score is read	condition the prior only on public metadata or a separate split; log $\Pi_u(\mathcal{A})$

Table 4: Assumption audit for applying the theory to real prompt pipelines.

A.3 Assumption audit

The conditional i.i.d. store is the strongest of these assumptions: scraped corpora contain near-duplicates and cross-document dependence. Proposition D.21 covers dependence inside public blocks, and a split that no stage reads remains valid in every case.

A.4 Limitations

The results are intentionally agnostic to the internals of the in-context learner. They do not characterize when transformers learn an optimal inner algorithm, nor do they replace task-specific analyses for linear attention or softmax attention. Proposition D.21 handles public independent clusters, but stronger corpus dependence would require sharper mixing or graph-dependent concentration. Finally, mutual information can be difficult to compute for continuous prompt transcripts; practical use requires finite-precision logging, randomized retrieval, or PAC-Bayesian upper bounds such as Theorem 6.1. The synthetic and API-backed experiments are diagnostic, and the Banking77 study (Appendix B.1) covers one task and one recipe library; there the certificates are valid but conservative by roughly an order of magnitude relative to the mean gap, the price of validity against the worst case of Theorem 4.4.

A.5 Broader impact

The paper studies statistical validity of prompt construction rather than capabilities of any particular model. The positive use case is more honest evaluation of RAG, few-shot learning, and model-in-the-loop prompt engineering: reporting capacity budgets can reduce accidental overclaiming when validation examples are repeatedly reused. A possible misuse is performative certification, where a system reports a small budget while omitting unlogged prompt-engineering rounds. This is why the theory emphasizes transcript logging and split hygiene. The real-LLM stress tests use fictional labels and synthetic task text, and the Banking77 study uses a public benchmark, so no sensitive user data is processed.

B API-Backed Protocols

Embedding retrieval. The supplementary code includes an OpenAI-embeddings retrieval protocol that embeds a small labeled text retrieval benchmark, tries multiple prompt views, chooses the view with best validation relevance, and reports the validation-holdout relevance gap. This protocol is intended as a lightweight real-embedding analogue of the finite-view selection bound: trying M prompt views spends roughly $\log M$ nats before any examples enter the final context.

Real LLM prompt search. We also provide experiments that call an instruction-following LLM through the OpenAI Responses API rather than using embeddings as a proxy. The semantic fictional-label benchmark is deliberately easy: in the reported run, all prompt recipes reach perfect validation and holdout accuracy, showing that recipe search need not overfit when the task is stable. The stronger nonce-label stress test searches over context recipes that leak $0, 2, 4$, or 8 validation examples into the prompt. Validation-selected recipes become increasingly confident on the reused validation split while holdout accuracy falls, and the same recorded outputs yield a SCAR curve showing how concentrating posterior mass on validation-favored recipes spends KL capacity.

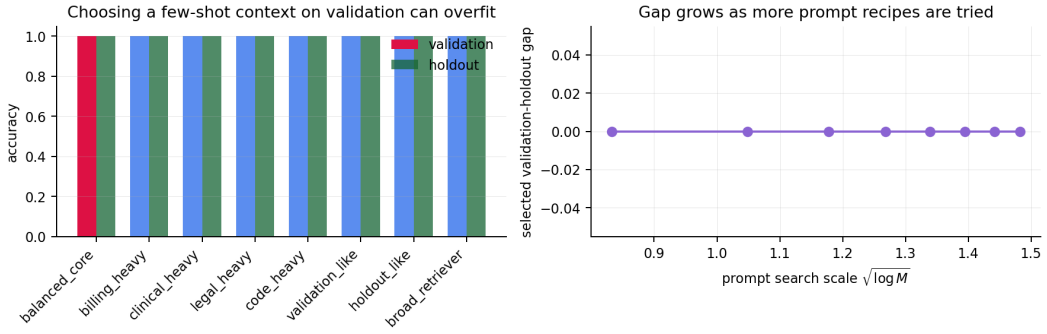

Figure 4: Real LLM semantic prompt-recipe selection using the OpenAI Responses API. The reported run is an intentionally easy control: all recipes solve the fictional-label task on both validation and holdout.

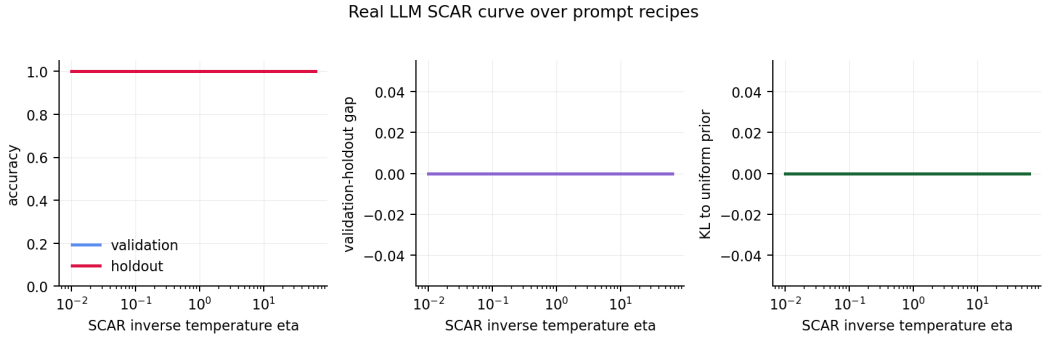

Figure 5: Real LLM SCAR curve over prompt recipes on the easy semantic task, a null control. Every recipe is perfect on validation and holdout, so the SCAR posterior stays at the uniform prior: the KL capacity and the context gap are zero at every inverse temperature. Figure 6 shows the overfitting case.

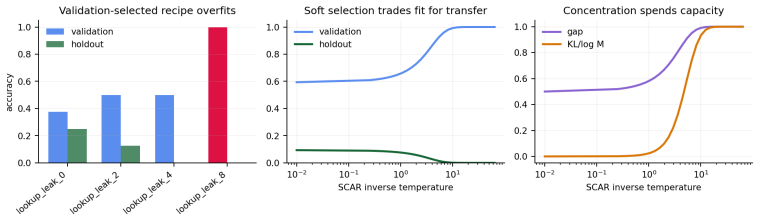

Figure 6: Real LLM nonce-label prompt-search stress test. Left: validation-selected context recipes that reuse more validation examples have inflated validation accuracy and poor holdout accuracy. Middle/right: the corresponding SCAR curve trades validation fit against transfer while KL concentration approaches the finite recipe-search budget.

B.1 Banking77 retrieval pipeline

Protocol. Banking77 [9] (CC BY 4.0) labels online-banking queries with 77 intents. With a fixed seed we draw a store of 2,000 labeled training queries, and a validation pool of 256 and a holdout of 1,024 test queries, disjoint from each other. Queries are embedded with bge-small-en-v1.5 [45] (MIT license). The library holds $M = 64$ recipes fixed before any score is read: 4 instruction templates $\times$ 4 retrieval rules (random, top-m, label-diverse top-m, maximal marginal relevance) $\times$ 4 demonstration configurations ($m \in \{4, 8\}$, most similar first or last). The frozen learners are Qwen2.5-7B-Instruct [47] (Apache 2.0) served with vLLM [26] (Apache 2.0), and GPT-4.1-mini through the OpenAI API, both at temperature 0; GPT-4.1-mini uses the first 512 holdout queries. For each $M \in \{2, \dots, 64\}$ and $k \in \{16, \dots, 256\}$ we draw 400 validation subsets of size k (the whole pool at $k = 256$), each with a random sub-library of M of the 64 recipes, select the recipe with the lowest validation error, and record its holdout error minus its validation error. The loss is 0–1, so $\sigma^2 = 1/4$. The best-of-64 winner is certified by Theorem 4.2 with $\kappa = \log(M/\beta)$ (Lemma D.7) and $\beta = \delta = 0.025$; SCAR uses the uniform prior and $\lambda^\star = \sqrt{8k(\log 64 + \log 20)}$, fixed in advance, and is certified by Theorem 6.1 at $\delta = 0.05$, averaged over 50 validation subsets. Because Theorem 6.1 holds simultaneously for all posteriors, the same $\lambda^\star$ also certifies the best-of-64 winner through a point mass, at the surcharge $\sqrt{(\log 64 + \log 20)/(2k)}$; this is the paired comparison with SCAR.

	Qwen2.5-7B-Instruct (zero-shot 0.373)					GPT-4.1-mini (zero-shot 0.282)				
k	16	32	64	128	256	16	32	64	128	256
mean gap, best-of-64	0.057	0.039	0.023	0.016	0.010	0.052	0.042	0.028	0.017	0.008
95th percentile gap	0.142	0.107	0.071	0.045	0.014	0.148	0.109	0.074	0.045	0.010
bound on mean gap, Thm. 4.1	0.361	0.255	0.180	0.127	0.090	0.361	0.255	0.180	0.127	0.090
certificate, Thm. 4.2	0.600	0.425	0.300	0.212	0.150	0.600	0.425	0.300	0.212	0.150
certified risk, SCAR	0.455	0.363	0.292	0.232	0.198	0.455	0.351	0.296	0.239	0.204
certified risk, best-of-64, Thm. 6.1	0.532	0.413	0.324	0.251	0.212	0.542	0.402	0.326	0.257	0.216
certified risk, best-of-64, Thm. 4.2	0.659	0.503	0.388	0.296	0.244	0.669	0.492	0.390	0.302	0.248
logged $\mathrm{KL}(Q\|\Pi)$, SCAR	1.08	1.23	1.34	1.59	1.63	0.86	0.99	1.23	1.47	1.78

Table 5: Banking77 with $M = 64$ recipes. Top: realized gap of the best-of-64 recipe over 400 validation subsets, against the bound of Theorem 4.1 and the certificate of Theorem 4.2. Bottom: certified risk (validation error plus surcharge) of SCAR and of the best-of-64 winner, whose capacity is $\log 64 = 4.16$ nats; these rows average 50 validation subsets, with standard errors at most 0.01 (0.06 for the KL).

Results. Across the 30 (M, k) cells the mean gap grows with $\sqrt{\log M/k}$, and the realized gap stayed below the certificate of Theorem 4.2 in all 30×400 selection events for each model (Table 5). The certificates are conservative on this benign library, by roughly an order of magnitude relative to the mean gap; Theorem 4.4 shows that no method improves the rate in the worst case. SCAR spends 0.9 to 1.8 logged nats instead of 4.16 and gives the tighter certificate at every k, while the two selectors' holdout errors agree within 0.01. The point-mass certificate of the winner also held in every event; the largest gap reaches 95% of it. SCAR's certificate is 0.01 to 0.09 below this paired certificate, as it must be: $Q_{\lambda^\star}$ minimizes $\mathbb{E}_Q \widehat{L}_S + \mathrm{KL}(Q\|\Pi)/\lambda^\star$, and the point mass on the winner is one such Q. Over 4,000 further validation draws per k, SCAR's certificate lies below the zero-shot holdout error in 99.7% of draws at $k = 64$ for Qwen and 99.2% at $k = 128$ for GPT-4.1-mini, against 96.1% and 89.2% for the paired best-of-64 certificate and 38.8% and 8.6% for the certificate of Theorem 4.2. A three-round variant that chooses the template, the retrieval rule, and the configuration in turn, each among four options on the same $k = 64$ queries, has mean gaps 0.004, 0.005, and 0.015 for Qwen (0.011, 0.011, 0.019 for GPT-4.1-mini), below the composed bounds $\sqrt{R \log 4/(2k)} = 0.104, 0.147, 0.180$ of Theorem 4.1 and Proposition 4.5. The zero-shot errors are themselves holdout estimates, with standard errors 0.015 and 0.020.

Compute. Embedding took under a minute on one T4 GPU; the 87,040 Qwen2.5-7B-Instruct calls ran as nine parallel H100 jobs in 11 minutes, and the 52,224 GPT-4.1-mini calls took 18 minutes. All analyses run on a CPU from cached model outputs, which are released with the scripts.

C Additional Synthetic Figures

Figure 7 shows the full SCAR temperature ablation summarized in Figure 3(c), and Figure 8 the second-order experiment of Figure 3(d), with the certificates of Theorem 6.2.

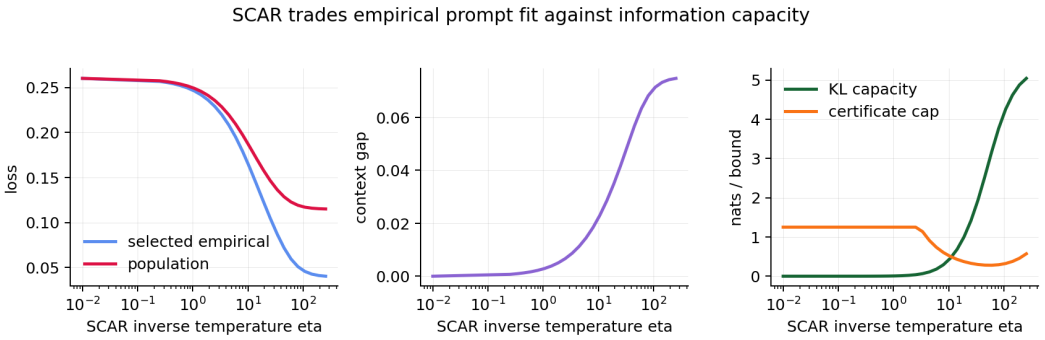

Figure 7: SCAR temperature ablation over 2500 Monte Carlo trials with context length $k = 64$. Increasing inverse temperature improves selected empirical fit but spends more KL capacity and increases context overfitting.

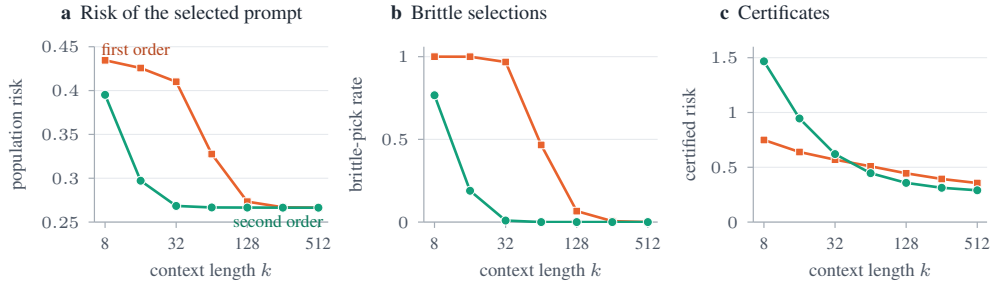

Figure 8: Variance-sensitive prompt certification over 3000 trials. The second-order objective reduces the probability of selecting brittle high-variance prompts and lowers the population risk of the selected prompt; its certificate (Theorem 6.2) is the tighter one from $k = 64$.

D Proofs

D.1 Auxiliary inequalities

Lemma D.1 (Variational KL inequality). *Let $Q \ll P$ and let f be measurable with $\mathbb{E}_P e^f < \infty$. Then*

$$\mathbb{E}_Q f \le \mathrm{KL}(Q\|P) + \log \mathbb{E}_P e^f.$$

This is the standard Donsker–Varadhan variational inequality; see, e.g., Donsker and Varadhan [12] or the PAC-Bayesian form in Catoni [10].

Lemma D.2 (Hoeffding's bounded-difference mgf). *If $X \in [a, b]$ almost surely, then for every $\lambda \in \mathbb{R}$,*

$$\log \mathbb{E} \exp\{\lambda(X - \mathbb{E}X)\} \le \frac{\lambda^2 (b-a)^2}{8}.$$

This is Hoeffding's lemma for bounded random variables [6, 22].

Lemma D.3 (Weighted sums of sub-Gaussian variables). *Let $X_1, \dots, X_m$ be independent, mean-zero variables satisfying*

$$\log \mathbb{E} e^{\lambda X_j} \le \frac{\nu_j \lambda^2}{2} \qquad \text{for all } \lambda \in \mathbb{R}.$$

Then $\sum_j w_j X_j$ is mean zero and sub-Gaussian with variance proxy $\sum_j w_j^2 \nu_j$.

This is the standard tensorization rule for independent sub-Gaussian variables [6].

Lemma D.4 (Sub-Gaussian tail)**.** *If X is mean zero and ν-sub-Gaussian, then for every $\epsilon > 0$,*

$$\mathbb{P}(X > \epsilon) \le \exp\left\{-\frac{\epsilon^2}{2\nu}\right\}.$$

This is Chernoff's method specialized to sub-Gaussian cumulants [6].

Lemma D.5 (Bernstein conjugate)**.** *For $v, b, x \ge 0$,*

$$\inf_{\lambda \in (0,1/b)} \left\{ \frac{x}{\lambda} + \frac{\lambda v}{2(1 - b\lambda)} \right\} = \sqrt{2vx} + bx,$$

with the convention $(0, 1/0) = (0, \infty)$.

This is the usual sub-gamma conjugate [6, Chapter 2]. For $x > 0$ the objective is convex and is minimized where $\lambda/(1 - b\lambda) = \sqrt{2x/v}$, that is, at $\lambda = \sqrt{2x}/(\sqrt{v} + b\sqrt{2x})$, where it equals $\sqrt{2vx} + bx$; at $x = 0$ the infimum is 0.

Lemma D.6 (Conditional chain rule and data processing)**.** *For random variables on standard Borel spaces,*

$$\mathrm{I}(A, B; S \mid U) = \mathrm{I}(A; S \mid U) + \mathrm{I}(B; S \mid U, A).$$

If R is generated from (T, U) and is conditionally independent of S given (T, U), then

$$\mathrm{I}(R; S \mid U) \le \mathrm{I}(T; S \mid U).$$

These are the standard mutual-information chain rule and data-processing inequality [11, Chapter 2].

Lemma D.7 (Finite menus)**.** *If for P_U-almost every u the transcript lies almost surely in a set $\mathcal{T}_u$ that does not depend on S, with $|\mathcal{T}_u| \le M$, then $\mathrm{I}(T; S \mid U) \le \log M$ and $I_\infty^\beta(T; S \mid U) \le \log(M/\beta)$ for every $\beta \in (0, 1]$.*

Proof. First, $\mathrm{I}(T; S \mid U) \le H(T \mid U) \le \log M$. Fix u and an event E, write $p(t) = \mathbb{P}(T = t \mid U = u)$ and $E_t = \{s : (t, s) \in E\}$, and let $\mathcal{B} = \{t \in \mathcal{T}_u : p(t) \le \beta/M\}$, so that $\mathbb{P}(T \in \mathcal{B} \mid U = u) \le \beta$. For $t \notin \mathcal{B}$, $\mathbb{P}(T = t, S \in E_t \mid U = u) \le \mathbb{P}(S \in E_t \mid U = u) \le (M/\beta)\, p(t)\, \mathbb{P}(S' \in E_t \mid U = u)$. Summing over t gives $\mathbb{P}((T, S) \in E \mid U = u) \le \beta + (M/\beta)\, \mathbb{P}((T, S') \in E \mid U = u)$. □

This is the description-length bound of Dwork et al. [14, Thm. 6], stated there in bits.

Lemma D.8 (Exponential Markov)**.** *If $Y \ge 0$ and $\mathbb{E}Y \le 1$, then*

$$\mathbb{P}\{Y > e^x\} \le e^{-x}.$$

This is Markov's inequality in exponential form [6, Chapter 2].

Lemma D.9 (Gaussian release comparison)**.** *Let S° be independent of S, with any law, and let*

$$Y = a(S) + G, \qquad G \sim N(0, \rho^{-1} I_m)$$

independently of S. Then

$$\mathrm{I}(Y; S) \le \frac{\rho}{2} \mathbb{E}\|a(S) - a(S^\circ)\|_2^2.$$

Proof. Let P_s denote the law $N(a(s), \rho^{-1} I_m)$ and let $R = \mathbb{E}_{S^\circ} P_{S^\circ}$. For every law R, $\mathrm{I}(Y; S) \le \mathbb{E}_S \mathrm{KL}(P_S \| R)$ [34, Thm. 4.1 and Cor. 4.2]. For fixed s with density p_s,

$$\mathrm{KL}(P_s \| R) = \int p_s(y) \log \frac{p_s(y)}{\mathbb{E}_{S^\circ} p_{S^\circ}(y)}\, dy \le \mathbb{E}_{S^\circ} \int p_s(y) \log \frac{p_s(y)}{p_{S^\circ}(y)}\, dy,$$

because $-\log$ is convex. Averaging over S gives

$$\mathrm{I}(Y; S) \le \mathbb{E}_{S, S^\circ} \mathrm{KL}(P_S \| P_{S^\circ}).$$

Expanding the two Gaussian densities with the common covariance $\rho^{-1} I_m$ yields

$$\mathrm{KL}(P_S \| P_{S^\circ}) = \frac{\rho}{2} \|a(S) - a(S^\circ)\|_2^2.$$

□

Lemma D.10 (Binomial lower-tail Chernoff)**.** *If $B \sim \mathrm{Binomial}(n,p)$ with mean $\mu = np$, then for every $\delta \in [0,1]$,*

$$\mathbb{P}(B \le (1-\delta)\mu) \le \exp\{-\mu\delta^2/2\}.$$

This is the standard multiplicative Chernoff lower-tail bound [6, 22].

Lemma D.11 (Maximal Rademacher fluctuation)**.** *There is a numerical constant $c_0 > 0$ such that, for independent Rademacher variables ϵ_{mj} and*

$$X_j = \frac{1}{k}\sum_{m=1}^{k} \epsilon_{mj}, \qquad j = 1, \dots, N,$$

whenever $k \ge 16$, $N \ge 2$, and $\log N \le k/16$,

$$\mathbb{E}\max_{j \le N} X_j \ge c_0 \sqrt{\frac{\log N}{k}}.$$

Proof. Set $a = \sqrt{\log N/(64k)}$. The assumption gives $a \le 1/32$. We first bound one average from below with the reverse Chernoff bound for the symmetric binomial [30, Prop. 7.3.2]: if n' is even and $B \sim \mathrm{Bin}(n', 1/2)$, then $\mathbb{P}(B \ge n'/2 + r) \ge \frac{1}{15} e^{-16r^2/n'}$ for every integer $r \in [0, n'/8]$. Let $k' = 2\lfloor k/2 \rfloor \ge k - 1 \ge 15$, so $k' \ge 16$, and let S' be the sum of the first k' signs of X_1, so that $kX_1 \ge S' - 1$. The integer $r = \lceil (ka+1)/2 \rceil$ satisfies $2r \ge ka + 1$ and $r \le ka/2 + 3/2 \le k/64 + 3/2 \le k'/8$. Applying the bound to $(S' + k')/2 \sim \mathrm{Bin}(k', 1/2)$, and using $k/(k-1) \le 16/15$ and $a \le 1/32$,

$$\begin{aligned}\mathbb{P}(X_1 \ge a) \ge \mathbb{P}(S' \ge 2r) &\ge \frac{1}{15} e^{-16r^2/k'} \ge \frac{1}{15}\exp\Big\{-\frac{4k^2a^2 + 24ka + 36}{k-1}\Big\} \\ &\ge c_1 e^{-64ka^2/15} = c_1 N^{-1/15} \ge c_1 N^{-1/4},\end{aligned}$$

with $c_1 = e^{-3.2}/15$. Let $M' := \max_{2 \le j \le N} X_j$. Since the X_j are independent and $(N-1)N^{-1/4} \ge N^{3/4}/2 \ge 2^{-1/4}$ for $N \ge 2$,

$$\mathbb{P}\{M' \ge a\} \ge 1 - \exp\{-c_1(N-1)N^{-1/4}\} \ge 1 - \exp\{-2^{-1/4}c_1\} := c_2 > 0.$$

The maximum can be negative, so a probability bound alone does not bound its mean. Instead write $\max_{j \le N} X_j = X_1 + (M' - X_1)_+$. Since $\mathbb{E}X_1 = 0$, X_1 is symmetric, and X_1 is independent of M',

$$\mathbb{E}\max_{j \le N} X_j = \mathbb{E}(M' - X_1)_+ \ge a\,\mathbb{P}\{X_1 \le 0\}\,\mathbb{P}\{M' \ge a\} \ge \frac{c_2 a}{2}.$$

Since $a = \frac{1}{8}\sqrt{\log N/k}$, the claim holds with $c_0 = c_2/16$. □

Lemma D.12 (Sharp maximal fluctuation)**.** *Let $X_1, \dots, X_N$ be as in Lemma D.11. If $k \to \infty$ and $N = N_k$ satisfy $\log k = o(\log N)$ and $\log N = o(k)$, then for every $\eta \in (0,1)$,*

$$\mathbb{P}\Big\{\max_{j \le N} X_j < (1-\eta)\sqrt{2\log N/k}\Big\} \to 0, \qquad \mathbb{E}\max_{j \le N} X_j \ge (1 - o(1))\sqrt{2\log N/k}.$$

Proof. Fix η and let $a = (1-\eta)\sqrt{2\log N/k}$, which tends to 0. Let $m = \lceil k(1+a)/2 \rceil$ and $x = m/k - 1/2 \in [a/2, a/2 + 1/k]$. Since $X_1 \ge a$ exactly when at least $k(1+a)/2$ of its signs are positive, Lemma D.15 gives

$$p := \mathbb{P}(X_1 \ge a) \ge \frac{1}{k+1}\exp\Big\{-\frac{2kx^2}{1-4x^2}\Big\}, \qquad 2kx^2 \le \frac{ka^2}{2} + 2a + \frac{2}{k} = (1-\eta)^2 \log N + o(1).$$

Because $x \to 0$ and $\log(k+1) = o(\log N)$, this gives $\log p \ge -(1-\eta)^2(1+o(1))\log N$, so $Np \ge N^{1-(1-\eta)^2(1+o(1))} \to \infty$ and $\mathbb{P}\{\max_j X_j < a\} \le (1-p)^N \le e^{-Np} \to 0$. For the mean, $\max_j X_j \ge a\mathbf{1}\{\max_j X_j \ge a\} + X_1\mathbf{1}\{\max_j X_j < a\}$, and the Cauchy–Schwarz inequality with $\mathbb{E}X_1^2 = 1/k$ gives

$$\mathbb{E}\max_j X_j \ge a\,\mathbb{P}\{\max_j X_j \ge a\} - k^{-1/2}\mathbb{P}\{\max_j X_j < a\}^{1/2} = a(1 - o(1)),$$

because $a\sqrt{k} \to \infty$. Since η is arbitrary, the claim follows. □

Lemma D.13 (Fixed-transcript empirical Bernstein)**.** *Let $X_1, \dots, X_k \in [0,1]$, $k \geq 2$, be i.i.d. with mean μ, empirical mean $\hat{\mu}$, and*

$$\hat{v} = \frac{1}{k(k-1)} \sum_{j=1}^{k} (X_j - \hat{\mu})^2.$$

Then for every $x > 0$,

$$\mathbb{P}\left\{ \mu - \hat{\mu} > \sqrt{2\hat{v}x} + \frac{7x}{3(k-1)} \right\} \leq 2e^{-x}.$$

This is Theorem 4 of Maurer and Pontil [31] with failure probability $2e^{-x}$, written for the variance of the empirical mean: their sample variance is $k\hat{v}$.

Lemma D.14 (An inequality of Fan et al. [16])**.** *If $\xi \geq -1$ and $\gamma \in [0,1)$, then*

$$\exp\{\gamma\xi - \psi(\gamma)\xi^2\} \leq 1 + \gamma\xi, \qquad \psi(\gamma) := -\log(1-\gamma) - \gamma.$$

This is the pointwise inequality in the proof of Lemma 4.1 of Fan et al. [16], eq. (4.12). It holds because $x \mapsto (\log(1+x)-x)/x^2$ is increasing on $(-1,\infty)$ and $\gamma\xi \geq -\gamma$, so that $\log(1+\gamma\xi)-\gamma\xi \geq -\psi(\gamma)\xi^2$.

Lemma D.15 (Binomial tails by the method of types)**.** *For integers $0 \leq m \leq k$,*

$$\mathbb{P}\big(\mathrm{Bin}(k,1/2) \geq m\big) \geq \frac{1}{k+1} \exp\big\{-k\, D(m/k \,\|\, 1/2)\big\},$$

where $D(p\|q)$ is the binary relative entropy. Moreover, $D(\frac{1}{2}+x \,\|\, \frac{1}{2}) \leq 2x^2/(1-4x^2)$ for $|x| < 1/2$.

Proof. Let $p = m/k$. The value m is a mode of $\mathrm{Bin}(k,p)$: if $0 < m < k$ then $(k+1)p \in (m, m+1)$, and if $m \in \{0,k\}$ the law is a point mass at m. So m has probability at least $1/(k+1)$ under $\mathrm{Bin}(k,p)$, and the likelihood ratio at m is $2^{-k}p^{-m}(1-p)^{m-k} = e^{-kD(p\|1/2)}$, with $0^0 = 1$ [11, §11.1]. Hence $\mathbb{P}(\mathrm{Bin}(k,1/2) \geq m) \geq \mathbb{P}(\mathrm{Bin}(k,1/2) = m) \geq e^{-kD(p\|1/2)}/(k+1)$. For the second claim,

$$D(\tfrac{1}{2}+x \,\|\, \tfrac{1}{2}) = \sum_{j \geq 1} \frac{(2x)^{2j}}{2j(2j-1)} \leq \frac{1}{2} \sum_{j \geq 1} (2x)^{2j} = \frac{2x^2}{1-4x^2}.$$

□

Lemma D.16 (Fano's inequality for a uniform prior)**.** *Let J be uniform on $\{1, \dots, M\}$ with $M \geq 2$, and let $\hat{J}$ be computed from an observation X and randomness independent of (J, X). Then $\mathbb{P}(\hat{J} \neq J) \geq 1 - (\mathrm{I}(J;X) + \log 2)/\log M$.*

This is Fano's inequality [11, §2.10].

D.2 A mutual-information transfer lemma

Lemma D.17. *Let $S \sim P^n$ and let T be any random variable jointly distributed with S. Suppose that for every fixed t, the random variable $g(t,S)$ is ν-sub-Gaussian under $S \sim P^n$ and has mean zero. Then*

$$|\mathbb{E}g(T,S)| \leq \sqrt{2\nu\, \mathrm{I}(T;S)}.$$

Proof. Let $\bar{T}$ be an independent copy of T, independent of S. For any $\lambda > 0$, Lemma D.1 gives

$$\lambda \mathbb{E}g(T,S) \leq \mathrm{I}(T;S) + \log \mathbb{E} \exp\{\lambda g(\bar{T},S)\}.$$

Since $g(t,S)$ is ν-sub-Gaussian and mean zero for every fixed t, the last logarithm is at most $\lambda^2\nu/2$. Optimizing over λ gives the upper-tail bound. Applying the same argument to $-g$ gives the absolute-value statement. □

D.3 Proof of Theorem 4.1

Fix u and a transcript t. Conditional on $U = u$, the selected indices $i_1(t), \dots, i_k(t)$ are fixed numbers when t is fixed. Therefore

$$\mathrm{gap}_u(t, S) = L_u(t) - \frac{1}{k} \sum_{j=1}^{k} \ell_t(Z_{i_j(t)})$$

is mean zero and σ^2/k-sub-Gaussian under a fresh draw $S \sim P_u^n$ by Lemma D.3. Lemma D.17, conditional on $U = u$, yields

$$|\mathbb{E}[\mathrm{gap}_u(T, S) \mid U = u]| \le \sqrt{\frac{2\sigma^2}{k} \mathrm{I}(T; S \mid U = u)}.$$

Averaging over U and using Jensen's inequality gives

$$|\mathbb{E}[\mathrm{gap}_U(T, S)]| \le \sqrt{\frac{2\sigma^2}{k} \mathrm{I}(T; S \mid U)}.$$

For $[0, 1]$ losses, Lemma D.2 gives $\sigma^2 = 1/4$.

D.4 Sub-gamma context transport

Theorem 4.1 uses only a sub-Gaussian envelope for the ghost-context cumulant. A Bernstein-type envelope gives a variance-adaptive version.

Definition D.18 (Context cumulant envelope)**.** *For a fixed query u and transcript t, let*

$$\Lambda_u(\lambda; t) := \log \mathbb{E}_{S^\circ \sim P_u^n} \exp\{\lambda \mathrm{gap}_u(t, S^\circ)\},$$

where S° is independent of t. The pipeline has a (v, b) context-cumulant envelope if

$$\Lambda_u(\lambda; t) \le \frac{\lambda^2 v}{2(1 - b\lambda)} \quad \text{for all } u,\ t \in \mathrm{supp}(\bar{K}_u),\ \lambda \in (0, 1/b).$$

The sub-Gaussian case is recovered by taking $b = 0$ and $v = \sigma^2/k$.

The envelope asks only that the score of a fixed transcript on a fresh store concentrate like an average of k losses.

Theorem D.19 (Sub-gamma context transport)**.** *Assume the coupled and decoupled laws above are well-defined and that the pipeline has a (v, b) context-cumulant envelope, with the convention $1/0 = \infty$. Then*

$$\mathbb{E}[\mathrm{gap}_U(T, S)] \le \inf_{\lambda \in (0, 1/b)} \left\{ \frac{\mathrm{I}(T; S \mid U)}{\lambda} + \frac{\lambda v}{2(1 - b\lambda)} \right\}.$$

In particular,

$$\mathbb{E}[\mathrm{gap}_U(T, S)] \le \sqrt{2v\, \mathrm{I}(T; S \mid U)} + b\, \mathrm{I}(T; S \mid U).$$

The same statement holds for the lower tail after replacing gap *by* $-$gap *and using its cumulant envelope.*

Proof. Condition on $U = u$. Under the coupled law Q_u and the decoupled law $\bar{Q}_u$, Lemma D.1 gives, for every $\lambda \in (0, 1/b)$,

$$\lambda \mathbb{E}_{Q_u} \mathrm{gap}_u(T, S) \le \mathrm{KL}(Q_u \| \bar{Q}_u) + \log \mathbb{E}_{\bar{Q}_u} \exp\{\lambda \mathrm{gap}_u(T, S)\}.$$

The first term is $\mathrm{I}(T; S \mid U = u)$. The second term is bounded by the context-cumulant envelope, since under $\bar{Q}_u$ the transcript and the ghost store are independent:

$$\log \mathbb{E}_{\bar{Q}_u} \exp\{\lambda \mathrm{gap}_u(T, S)\} \le \frac{\lambda^2 v}{2(1 - b\lambda)}.$$

Dividing by λ, averaging over U, and optimizing over λ gives the variational bound. The relaxed expression

$$\sqrt{2v\, \mathrm{I}(T; S \mid U)} + b\, \mathrm{I}(T; S \mid U)$$

follows from Lemma D.5 applied with $x = \mathrm{I}(T; S \mid U)$. The lower-tail statement applies the same argument to $-$gap. □

D.5 Proof of Theorem 4.2

Fix u. Under the product law where T and S' are conditionally independent given $U = u$, Lemma D.4 gives

$$\mathbb{P}(\mathrm{gap}_u(T, S') > \epsilon \mid U = u) \le \exp\left\{-\frac{k\epsilon^2}{2\sigma^2}\right\}.$$

The definition of approximate max-information transfers this event to the joint law:

$$\mathbb{P}(\mathrm{gap}_u(T, S) > \epsilon \mid U = u) \le e^{\kappa} \exp\left\{-\frac{k\epsilon^2}{2\sigma^2}\right\} + \beta.$$

Set $\epsilon = \sqrt{2\sigma^2(\kappa + \log(1/\delta))/k}$. The lower-tail statement follows by applying the same argument to $-\mathrm{gap}$.

D.6 Weighted contexts, clustered stores, and latent tasks

These three statements are the extensions summarized in Remark 4.3.

Proposition D.20 (Weighted context transfer)**.** *Suppose a transcript t specifies selected indices $i_1(t), \dots, i_m(t)$ and nonnegative weights $w_1(t), \dots, w_m(t)$ with $\sum_j w_j(t) = 1$. Define $\widehat{L}^w_S(t) := \sum_{j=1}^m w_j(t)\ell_t(Z_{i_j(t)})$. If $\sum_{j=1}^m w_j(t)^2 \le 1/k_{\mathrm{eff}}$ for every transcript in the support of the pipeline, and the fixed-transcript loss is σ^2-sub-Gaussian, then*

$$\left|\mathbb{E}\left[L_U(T) - \widehat{L}^w_S(T)\right]\right| \le \sqrt{\frac{2\sigma^2}{k_{\mathrm{eff}}}\,\mathrm{I}(T; S \mid U)}.$$

Thus attention concentration or repeated demonstrations reduce the effective evidence even when the displayed context contains many tokens.

Proof. Fix u and t. Under the ghost draw $S^\circ \sim P_u^n$, the weighted selected-context gap

$$L_u(t) - \sum_{j=1}^m w_j(t)\ell_t(Z^\circ_{i_j(t)})$$

is mean zero. Since the summands are independent for distinct selected indices and each centered loss is σ^2-sub-Gaussian, Lemma D.3 shows that the weighted average is

$$\sigma^2 \sum_{j=1}^m w_j(t)^2 \le \frac{\sigma^2}{k_{\mathrm{eff}}}$$

sub-Gaussian. Applying Lemma D.17 conditionally on $U = u$ and then Jensen's inequality over U gives the result. If the implementation allows repeated indices, the same statement holds after merging repeats into their total weights; repeated displays can only increase $\sum_j w_j(t)^2$ and therefore reduce k_{eff}. □

Proposition D.21 (Clustered-store effective length)**.** *For each query u, suppose the store is partitioned into public blocks $B_1, \dots, B_G$. Conditional on $U = u$, the blocks are independent, while examples inside a block may be arbitrarily dependent and each selected example has marginal law P_u. For a transcript t, define the selected block mass*

$$\pi_g(t) := \frac{1}{k}\sum_{j=1}^k \mathbf{1}\{i_j(t) \in B_g\}.$$

If $\ell_t \in [0, 1]$ and $\sum_{g=1}^G \pi_g(t)^2 \le 1/k_{\mathrm{blk}}$ for all transcripts in the support of the pipeline, then

$$|\mathbb{E}[\mathrm{gap}_U(T, S)]| \le \sqrt{\frac{\mathrm{I}(T; S \mid U)}{2k_{\mathrm{blk}}}}.$$

Retrieving many chunks from one document therefore certifies like one heavily weighted block, not like many independent examples.

Proof. Fix u and t. Under the ghost law, define the block contribution

$$A_g(t, S^\circ) = \frac{1}{k} \sum_{j: i_j(t) \in B_g} \ell_t(Z^\circ_{i_j(t)}).$$

The variables $A_1, \dots, A_G$ are independent across blocks, although the summands inside each A_g may be dependent. Since $\ell_t \in [0, 1]$, each A_g lies in an interval of length at most $\pi_g(t)$. Lemma D.2 therefore gives

$$\log \mathbb{E} \exp \left\{ \lambda \sum_{g=1}^{G} (\mathbb{E}A_g - A_g) \right\} \le \frac{\lambda^2}{8} \sum_{g=1}^{G} \pi_g(t)^2 \le \frac{\lambda^2}{8k_{\mathrm{blk}}}.$$

Thus the fixed-transcript gap is $(4k_{\mathrm{blk}})^{-1}$-sub-Gaussian. Lemma D.17, conditional on U, gives

$$|\mathbb{E}\mathrm{gap}_U(T, S)| \le \sqrt{2(4k_{\mathrm{blk}})^{-1} \mathrm{I}(T; S \mid U)},$$

which is the stated bound. □

Proposition D.22 (Latent task capacity decomposition)**.** *Let $W \in \mathcal{W}$ be a latent task variable and suppose $S \mid (U = u, W = w) \sim P_{u,w}^n$ and $T \sim K_u(\cdot \mid S)$. Let $L_{u,w}(t) = \mathbb{E}_{Z \sim P_{u,w}} \ell_t(Z)$. If the centered fixed-transcript loss is σ^2-sub-Gaussian under every $P_{u,w}$, then*

$$\left| \mathbb{E}\left[L_{U,W}(T) - \widehat{L}_S(T) \right] \right| \le \sqrt{\frac{2\sigma^2}{k} \mathrm{I}(T; S \mid U, W)}.$$

Moreover, because the prompt kernel uses W only through (U, S),

$$\mathrm{I}(T; S \mid U) = \mathrm{I}(T; W \mid U) + \mathrm{I}(T; S \mid U, W).$$

The term $\mathrm{I}(T; W \mid U)$ is task-identification capacity; the second term is the within-task overfitting budget.

Proof. Condition on $(U, W) = (u, w)$. The store is i.i.d. from $P_{u,w}$, so the proof of Theorem 4.1 applies verbatim with U replaced by (U, W):

$$\left| \mathbb{E}[L_{U,W}(T) - \widehat{L}_S(T)] \right| \le \sqrt{\frac{2\sigma^2}{k} \mathrm{I}(T; S \mid U, W)}.$$

For the decomposition, use Lemma D.6 in two orders:

$$\mathrm{I}(T; S, W \mid U) = \mathrm{I}(T; S \mid U) + \mathrm{I}(T; W \mid U, S)$$

and

$$\mathrm{I}(T; S, W \mid U) = \mathrm{I}(T; W \mid U) + \mathrm{I}(T; S \mid U, W).$$

Because the prompt kernel is $K_u(\cdot \mid S)$, T is conditionally independent of W given (U, S); hence $\mathrm{I}(T; W \mid U, S) = 0$. Equating the two displays proves the identity. □

D.7 Proof of Theorem 4.4

Let $N = \lfloor e^\kappa \rfloor$ and let the data domain be $\mathcal{Z} = \{-1, +1\}^N$ with the uniform distribution. For each coordinate $j \in [N]$, define a prompt loss

$$\ell_j(z) = \frac{1 + z_j}{2}, \qquad L(j) = \frac{1}{2}.$$

Given $S = (Z_1, \dots, Z_k)$, the selector outputs

$$T = \arg\min_{j \in [N]} \frac{1}{k} \sum_{m=1}^{k} Z_{m,j}.$$

The transcript takes at most N values, so $\mathrm{I}(T; S) \le H(T) \le \log N \le \kappa$. For a store of size $n \ge k$ the selector reads only $Z_1, \dots, Z_k$, so $\mathrm{I}(T; S) = \mathrm{I}(T; Z_1, \dots, Z_k)$ and the gap is unchanged. Its gap is

$$L(T) - \widehat{L}_S(T) = -\frac{1}{2} \min_{j \in [N]} \frac{1}{k} \sum_{m=1}^{k} Z_{m,j} = \frac{1}{2} \max_{j \in [N]} \frac{1}{k} \sum_{m=1}^{k} (-Z_{m,j}).$$

These are independent centered Rademacher averages, so Lemma D.11 gives

$$\mathbb{E}\max_{j\in[N]}\frac{1}{k}\sum_{m=1}^{k}(-Z_{m,j})\geq c_0\sqrt{\frac{\log N}{k}}$$

whenever $\log N\leq k/16$. Since $N=\lfloor e^\kappa\rfloor\geq e^\kappa/2$ for $\kappa\geq 1$, we have

$$\log N\geq\kappa-\log 2\geq(1-\log 2)\kappa.$$

Absorbing this numerical factor into the universal constant proves the claim.

For the constant, suppose $k\to\infty$ with $\log k=o(\kappa)$ and $\kappa=o(k)$. Then $\kappa\to\infty$ and $\log N\geq\log(e^\kappa-1)=\kappa-o(1)$, so $\log k=o(\log N)$ and $\log N=o(k)$. Lemma D.12, applied to the gap above, gives

$$\mathbb{E}\left[L(T)-\widehat{L}_S(T)\right]\geq\frac{1-o(1)}{2}\sqrt{\frac{2\log N}{k}}=(1-o(1))\sqrt{\frac{\kappa}{2k}},$$

and the same lower bound for the gap itself with probability tending to one. Theorem 4.1 bounds the expected gap by $\sqrt{\kappa/(2k)}$, so its constant is exact in this regime. The transcript takes at most N values, so by Lemma D.7 its approximate max-information is at most $\log(N/\beta)$, and at fixed β and δ the certificate of Theorem 4.2 is also $(1+o(1))\sqrt{\kappa/(2k)}$.

D.8 Proof of Proposition 4.5

This is Lemma D.6 applied repeatedly:

$$\mathrm{I}(T_1,\dots,T_R;S\mid U)=\sum_{r=1}^{R}\mathrm{I}(T_r;S\mid U,T_1,\dots,T_{r-1}).$$

Substituting this identity into Theorem 4.1 gives the stated bound.

D.9 Profiled multi-round transport

Corollary D.23 (Profiled multi-round transport)**.** *Let $C_r:=\mathrm{I}(T_r;S\mid U,T_{<r})$ be the capacity profile of an R-round prompt search, and suppose the final transcript has a (v,b) context-cumulant envelope after conditioning on the full history. Then*

$$\mathbb{E}[\mathrm{gap}_U(T,S)]\leq\inf_{\lambda\in(0,1/b)}\left\{\frac{\sum_{r=1}^{R}C_r}{\lambda}+\frac{\lambda v}{2(1-b\lambda)}\right\}.$$

Thus an apparently harmless sequence of small validation probes behaves like a single prompt selector whose complexity is the whole capacity profile, not only the last retrieved context.

Proof. By Proposition 4.5,

$$\mathrm{I}(T;S\mid U)\leq\mathrm{I}(T_1,\dots,T_R;S\mid U)=\sum_{r=1}^{R}C_r,$$

where the inequality is Lemma D.6's data-processing statement if the final prompt transcript is a post-processing of the full round transcript. Substituting this bound into Theorem D.19 yields the result. □

D.10 Capacity of noisy score releases

Proposition D.24 (Capacity of noisy score releases)**.** *Fix a round history $H=h$. Suppose the pipeline computes a score vector $a_h(S)\in\mathbb{R}^m$ and releases $Y_h=a_h(S)+G$, $G\sim N(0,\rho^{-1}I_m)$, before applying arbitrary post-processing to choose candidates. If $S^\circ\sim P_u^n$ is an independent ghost store, then*

$$\mathrm{I}(Y_h;S\mid U=u,H=h)\leq\frac{\rho}{2}\,\mathbb{E}\left[\|a_h(S)-a_h(S^\circ)\|_2^2\mid U=u,H=h\right].$$

In particular, if $\|a_h(S)-a_h(S^\circ)\|_2\leq\Delta_h$ almost surely, the release spends at most $\rho\Delta_h^2/2$ nats of context capacity, regardless of how many deterministic filters are applied afterward.

Proof. Fix u, h and suppress them from the notation. Lemma D.9 gives the displayed mutual-information bound. The almost-sure diameter bound gives the second claim immediately. Any deterministic or randomized post-processing of Y_h can only reduce mutual information by Lemma D.6. □

D.11 Proof of Remark 4.6

On $\{B \geq k\}$, where B is the number of store items on which the fixed prompt has zero loss, the pipeline selects k such items; otherwise it selects any k items. B is Binomial$(n, 1-p)$ with mean $\mu = n(1-p) \geq 2k$, so Lemma D.10 with $\delta = 1/2$ gives

$$\mathbb{P}(B < k) \leq \mathbb{P}(B \leq \mu/2) \leq e^{-\mu/8} \leq e^{-k/4}.$$

On $\{B \geq k\}$ the selected score is 0 and the risk is p. If a surcharge satisfied $b < p$, the certified event $\{L \leq \widehat{L}_S + b\}$ would be contained in $\{B < k\}$, whose probability is at most $e^{-k/4} < 1 - \delta$. Hence $b \geq p$ for every $p \leq 1 - 2k/n$, and in particular $b \geq 1 - 2k/n$.

D.12 Proof of Theorem 6.1

For a fixed u and fixed t, the sub-Gaussian assumption implies

$$\mathbb{E}_S \exp\left\{\lambda(L_u(t) - \widehat{L}_S(t)) - \frac{\lambda^2\sigma^2}{2k}\right\} \leq 1.$$

Integrating over $t \sim \Pi_u$ and applying Lemma D.8, with probability at least $1 - \delta$,

$$\mathbb{E}_{t\sim\Pi_u} \exp\left\{\lambda(L_u(t) - \widehat{L}_S(t)) - \frac{\lambda^2\sigma^2}{2k}\right\} \leq \frac{1}{\delta}.$$

The change-of-measure inequality, which is Lemma D.1 with $P = \Pi_u$ and Q arbitrary,

$$\mathbb{E}_Q f(t) \leq \mathrm{KL}(Q\|\Pi_u) + \log \mathbb{E}_{\Pi_u} e^{f(t)}$$

with $f(t) = \lambda(L_u(t) - \widehat{L}_S(t)) - \lambda^2\sigma^2/(2k)$ gives, simultaneously for all Q,

$$\mathbb{E}_Q L_u(t) \leq \mathbb{E}_Q \widehat{L}_S(t) + \frac{\mathrm{KL}(Q\|\Pi_u) + \log(1/\delta)}{\lambda} + \frac{\lambda\sigma^2}{2k}.$$

Applying the preceding display with $Q = Q_\lambda$ gives

$$\mathbb{E}_{Q_\lambda} L_u(t) \leq \mathbb{E}_{Q_\lambda} \widehat{L}_S(t) + \frac{\mathrm{KL}(Q_\lambda\|\Pi_u)}{\lambda} + \frac{\log(1/\delta)}{\lambda} + \frac{\lambda\sigma^2}{2k}.$$

The condition $\int e^{-\lambda\widehat{L}_S(t)}\,\Pi_u(dt) < \infty$ of the theorem, which holds for instance when losses are bounded below, makes Q_λ a probability measure. Since Q_λ minimizes $Q \mapsto \mathbb{E}_Q\widehat{L}_S(t) + \mathrm{KL}(Q\|\Pi_u)/\lambda$, the first two terms on the right are bounded by the displayed infimum over comparator posteriors. This proves the oracle inequality.

For a gate $\mathcal{A} = \mathcal{A}(S, u)$ with $\Pi_u(\mathcal{A}) > 0$, the gated posterior is $Q_\lambda^{\mathcal{A}} \propto \Pi_u \mathbf{1}_{\mathcal{A}} e^{-\lambda\widehat{L}_S}$. Let $\Pi_{u,\mathcal{A}} := \Pi_u(\cdot \cap \mathcal{A})/\Pi_u(\mathcal{A})$. Every Q with $Q(\mathcal{A}) = 1$ satisfies

$$\mathrm{KL}(Q\|\Pi_u) = \mathrm{KL}(Q\|\Pi_{u,\mathcal{A}}) + \log\frac{1}{\Pi_u(\mathcal{A})},$$

so on such laws $Q \mapsto \mathbb{E}_Q\widehat{L}_S + \lambda^{-1}\mathrm{KL}(Q\|\Pi_u)$ differs from $Q \mapsto \mathbb{E}_Q\widehat{L}_S + \lambda^{-1}\mathrm{KL}(Q\|\Pi_{u,\mathcal{A}})$ by a constant. The Gibbs variational principle [10] minimizes the latter at $Q \propto \Pi_{u,\mathcal{A}} e^{-\lambda\widehat{L}_S} = Q_\lambda^{\mathcal{A}}$. The first display holds simultaneously for all data-dependent $Q \ll \Pi_u$, so applying it at $Q_\lambda^{\mathcal{A}}$ gives the gated oracle comparison, with every KL measured to Π_u; the gate charge $\log(1/\Pi_u(\mathcal{A}))$ enters through the identity above.

D.13 Rate optimality of SCAR

Proposition D.25 (Rate optimality of SCAR)**.** *Let $\ell_t \in [0,1]$, let the library have $M \geq 2$ prompt recipes scored on the same k selected examples, and let $L^\star = \min_t L_u(t)$.*

(i) With the uniform prior, no gate, $\sigma^2 = 1/4$, and $\lambda^\star = \sqrt{8k(\log M + \log(1/\delta))}$, with probability at least $1 - 2\delta$ SCAR satisfies

$$\mathbb{E}_{Q_{\lambda^\star}} L_u \leq L^\star + \sqrt{2(\log M + \log(1/\delta))/k}.$$

(ii) If $4 \leq M \leq e^{2k/3}$, there are laws $P_1, \ldots, P_M$ such that every rule that maps the k examples to a law Q on the library has

$$\max_j \mathbb{E}_j\big[\mathbb{E}_Q L_j - L_j^\star\big] \geq \frac{1}{4}\sqrt{\frac{3\log M}{32k}} \geq 0.07\sqrt{\frac{\log M}{k}}.$$

Proof. (i) Let $t^\star$ attain $L^\star$. On the event of Theorem 6.1, the oracle comparison at the point mass $Q = \delta_{t^\star}$, for which $\mathrm{KL}(\delta_{t^\star} \| \Pi_u) = \log M$, gives

$$\mathbb{E}_{Q_{\lambda^\star}} L_u \leq \widehat{L}_S(t^\star) + \frac{\log M + \log(1/\delta)}{\lambda^\star} + \frac{\lambda^\star}{8k} = \widehat{L}_S(t^\star) + \sqrt{\frac{\log M + \log(1/\delta)}{2k}}.$$

For the fixed recipe $t^\star$, Lemmas D.2 and D.4 give $\widehat{L}_S(t^\star) \leq L^\star + \sqrt{\log(1/\delta)/(2k)}$ with probability at least $1 - \delta$, and the two surcharges sum to at most $\sqrt{2(\log M + \log(1/\delta))/k}$.

(ii) Let $\mathcal{Z} = \{0,1\}^M$ and let recipe t have loss $\ell_t(z) = z_t$. Under P_j the coordinates are independent, coordinate j is Bernoulli$(\frac{1}{2} - \epsilon)$ and the others are Bernoulli$(\frac{1}{2})$, where $\epsilon^2 = 3\log M/(32k) \leq 1/16$. Then $L_j^\star = \frac{1}{2} - \epsilon$ and $\mathbb{E}_Q L_j - L_j^\star = \epsilon(1 - Q(j))$. Draw J uniformly from $[M]$, the examples from $P_J^{\otimes k}$, and $\hat{J}$ from Q; the average excess risk is $\epsilon\,\mathbb{P}(\hat{J} \neq J)$. For every law R, $\mathrm{I}(J;X) \leq \mathbb{E}_J \mathrm{KL}(P_{X|J} \| R)$ [34, Thm. 4.1 and Cor. 4.2]. With $R = P_0^{\otimes k}$, the law of M fair coordinates, Lemma D.15 gives

$$\mathrm{I}(J;X) \leq \max_j \mathrm{KL}\big(P_j^{\otimes k} \,\|\, P_0^{\otimes k}\big) = kD(\tfrac{1}{2} - \epsilon \,\|\, \tfrac{1}{2}) \leq \frac{2k\epsilon^2}{1 - 4\epsilon^2} \leq \frac{8}{3}k\epsilon^2 = \frac{1}{4}\log M.$$

Lemma D.16 then gives $\mathbb{P}(\hat{J} \neq J) \geq 1 - (\frac{1}{4}\log M + \log 2)/\log M \geq \frac{1}{4}$ for $M \geq 4$, so some j has expected excess risk at least $\epsilon/4$. □

The upper bound is the certificate of Theorem 6.1 turned into a risk guarantee, and the lower bound is the sample complexity of identifying the best of M options. In (ii) the excess risk lies in $[0, \epsilon]$, so for the worst j it is at least $\epsilon/8$ with probability at least $1/7$; hence for $2\delta < 1/7$ no rule has a $(1 - 2\delta)$-probability excess-risk bound below $\frac{1}{8}\sqrt{3\log M/(32k)}$, and in this sense the two parts match up to constants.

D.14 Proof of Theorem 6.2

The proof is a betting construction with the predictable plug-in mean of Waudby-Smith and Ramdas [43]; see also Howard et al. [23].

Lemma D.26 (Predictable plug-in supermartingale)**.** *Let $X_1, \ldots, X_k \in [0,1]$ be independent with common mean μ, and let each $\widehat{\mu}_{j-1} \in [0,1]$ be a Borel function of $X_1, \ldots, X_{j-1}$. For every $\gamma \in [0, 1)$, the partial products of $D_j := \exp\{\gamma(\mu - X_j) - \psi(\gamma)(X_j - \widehat{\mu}_{j-1})^2\}$ form a nonnegative supermartingale, and*

$$\mathbb{E}\exp\Big\{\gamma\sum_{j=1}^{k}(\mu - X_j) - \psi(\gamma)\sum_{j=1}^{k}(X_j - \widehat{\mu}_{j-1})^2\Big\} \leq 1.$$

Proof. Let $\mathcal{F}_{j-1} = \sigma(X_1, \ldots, X_{j-1})$ and $\xi_j = \widehat{\mu}_{j-1} - X_j \in [-1, 1]$. Then $\gamma(\mu - X_j) = \gamma(\mu - \widehat{\mu}_{j-1}) + \gamma\xi_j$ and $(X_j - \widehat{\mu}_{j-1})^2 = \xi_j^2$, and $\mathbb{E}[X_j \mid \mathcal{F}_{j-1}] = \mu$ by independence. Lemma D.14 gives

$$\mathbb{E}\big[D_j \,\big|\, \mathcal{F}_{j-1}\big] \leq e^{\gamma(\mu - \widehat{\mu}_{j-1})}\big(1 + \gamma(\widehat{\mu}_{j-1} - \mu)\big) = e^{-a}(1 + a) \leq 1, \qquad a = \gamma(\widehat{\mu}_{j-1} - \mu).$$

The tower property over $j = k, k-1, \ldots, 1$ proves the claim. □

Proof of Theorem 6.2. Fix u and $\gamma \in \Gamma$. For a fixed transcript t the selected indices are distinct constants, so under $S \sim P_u^n$ the selected losses $X_j(t)$ are i.i.d. in $[0,1]$ with mean $L_u(t)$, and $\widehat{\mu}_{j-1}(t) \in [0,1]$ depends only on $X_1(t), \dots, X_{j-1}(t)$. Lemma D.26 gives $\mathbb{E}_S M_\gamma(t,S) \le 1$ for

$$M_\gamma(t,S) = \exp\{\gamma k\big(L_u(t) - \widehat{L}_S(t)\big) - \psi(\gamma)\, k\, \widehat{V}_S(t)\}.$$

Since Π_u does not depend on S, Fubini's theorem gives $\mathbb{E}_S \mathbb{E}_{t\sim\Pi_u} M_\gamma(t,S) \le 1$, and Lemma D.8 gives $\mathbb{E}_{\Pi_u} M_\gamma > |\Gamma|/\delta$ with probability at most $\delta/|\Gamma|$. A union bound over the grid gives an event of probability at least $1-\delta$ on which $\mathbb{E}_{\Pi_u} M_\gamma \le |\Gamma|/\delta$ for every $\gamma \in \Gamma$. On that event, Lemma D.1 gives, for every $Q \ll \Pi_u$ and every $\gamma \in \Gamma$,

$$\gamma k\big(\mathbb{E}_Q L_u - \mathbb{E}_Q \widehat{L}_S\big) - \psi(\gamma)\, k\, \mathbb{E}_Q \widehat{V}_S \le \mathrm{KL}(Q\|\Pi_u) + \log\frac{|\Gamma|}{\delta},$$

and dividing by γk proves the bound. For fixed γ the right-hand side is, up to a constant, $\mathbb{E}_Q[\widehat{L}_S + \psi(\gamma)\widehat{V}_S/\gamma] + \mathrm{KL}(Q\|\Pi_u)/(\gamma k)$, which the Gibbs variational principle [10] minimizes at $Q \propto \Pi_u \exp\{-k(\gamma\widehat{L}_S + \psi(\gamma)\widehat{V}_S)\}$. □

Reading the certificate. Since $\psi(\gamma)/\gamma = \sum_{m\ge 2} \gamma^{m-1}/m \le \gamma/(2(1-\gamma))$, the surcharge at γ is at most $\gamma\, \mathbb{E}_Q \widehat{V}_S/(2(1-\gamma)) + C/(\gamma k)$ with $C = \mathrm{KL}(Q\|\Pi_u) + \log(|\Gamma|/\delta)$. At $\gamma^\circ = \sqrt{2C/(k\, \mathbb{E}_Q \widehat{V}_S)}$ this equals $(1 + \gamma^\circ/(2(1-\gamma^\circ)))\sqrt{2\, \mathbb{E}_Q \widehat{V}_S\, C/k}$, and the grid Γ_k of Section 6 contains a point within a factor $\sqrt{2}$ of γ° whenever $\gamma^\circ \in [0.9 \cdot 2^{-\lceil \log_2 k \rceil}, 0.9]$, which costs at most a factor $3/(2\sqrt{2}) < 1.07$ in the leading term. For i.i.d. losses with mean L and variance v,

$$\mathbb{E}\widehat{V}_S(t) = v\Big(1 + \frac{H_k - H_k^{(2)}}{k}\Big) + \Big(\frac{1}{2} - L\Big)^2 \frac{H_k^{(2)}}{k}, \qquad H_k = \sum_{j\le k} \frac{1}{j}, \quad H_k^{(2)} = \sum_{j\le k} \frac{1}{j^2} \le \frac{\pi^2}{6},$$

because $\mathbb{E}(X_j - \widehat{\mu}_{j-1})^2 = v + ((\tfrac{1}{2} - L)^2 + (j-1)v)/j^2$. The last term reflects the initial guess $\widehat{\mu}_0 = \tfrac{1}{2}$. The certificates of Figures 3(d) and 8 use Γ_k, $\delta = 0.05$, and the exact minimum over the grid.

Remark D.27 (A union-bound variant for countable libraries)**.** *If $\mathcal{T}$ is countable and $k \ge 2$, a union bound over Lemma D.13 gives a closed-form alternative. With $\widehat{v}_S(t) := \frac{1}{k(k-1)} \sum_j (X_j(t) - \widehat{L}_S(t))^2$, with probability at least $1-\delta$, simultaneously for all $Q \ll \Pi_u$,*

$$\mathbb{E}_Q L_u(t) \le \mathbb{E}_Q \widehat{L}_S(t) + \sqrt{2\, \mathbb{E}_Q \widehat{v}_S(t)\, c_Q} + \frac{7 c_Q}{3(k-1)}, \qquad c_Q = \mathbb{E}_Q \log\frac{1}{\Pi_u(t)} + \log\frac{2}{\delta}.$$

Since $\mathbb{E}_Q \log(1/\Pi_u(t)) = \mathrm{KL}(Q\|\Pi_u) + H(Q)$, this variant avoids the grid for point masses but pays the entropy of a spread-out posterior.

Proof. Fix u. For a fixed transcript t the selected indices are distinct constants, so under $S \sim P_u^n$ the selected losses are i.i.d. in $[0,1]$ with mean $L_u(t)$. For each t with $\Pi_u(t) > 0$ set $x_t := \log \frac{2}{\delta \Pi_u(t)}$. Lemma D.13 gives

$$L_u(t) - \widehat{L}_S(t) \le \sqrt{2\widehat{v}_S(t) x_t} + \frac{7 x_t}{3(k-1)}$$

except on an event of probability at most $2e^{-x_t} = \delta \Pi_u(t)$. A union bound over the countable set $\{t : \Pi_u(t) > 0\}$ yields an event of probability at least $1-\delta$ on which these inequalities hold for every such t simultaneously. On this event let $Q \ll \Pi_u$, possibly data-dependent; Q is supported on $\{t : \Pi_u(t) > 0\}$. Averaging over $t \sim Q$ and applying the Cauchy–Schwarz inequality $\mathbb{E}_Q \sqrt{2\widehat{v}_S(t) x_t} \le \sqrt{2\, \mathbb{E}_Q \widehat{v}_S(t)\, \mathbb{E}_Q x_t}$ gives the claim, because

$$\mathbb{E}_Q x_t = \mathbb{E}_Q \log\frac{1}{\Pi_u(t)} + \log\frac{2}{\delta}.$$

Finally $\mathbb{E}_Q \log(1/\Pi_u(t)) = \mathrm{KL}(Q\|\Pi_u) + H(Q) \ge \mathrm{KL}(Q\|\Pi_u)$, with equality when Q is a point mass. □

D.15 Proof of Proposition 7.1

Condition on X_I. The Gaussian prior and Gaussian likelihood imply posterior covariance

$$\Sigma_I = \left(\tau^{-2} I_d + \sigma_\xi^{-2} X_I^\top X_I\right)^{-1}.$$

The posterior mean is the ridge predictor, and the conditional mean-squared prediction error for $y_\star = x_\star^\top \theta + \xi_\star$ is the observation noise plus posterior variance:

$$\sigma_\xi^2 + x_\star^\top \Sigma_I x_\star.$$

If $X_I^\top X_I / k \succeq \alpha u u^\top + \beta(I_d - uu^\top)$, then along the query direction u,

$$u^\top \Sigma_I u \le \frac{1}{\tau^{-2} + k\alpha/\sigma_\xi^2}.$$

Multiplying by $\|x_\star\|^2$ gives the stated inequality.

D.16 Schur-complement form of the query precision

Proposition D.28 (Schur-complement form of query precision)**.** *Let $A_I = k^{-1} X_I^\top X_I$, let $u = x_\star/\|x_\star\|$, and write $\alpha_I = u^\top A_I u$, $b_I = (I_d - uu^\top) A_I u$, and $C_I = (I_d - uu^\top) A_I (I_d - uu^\top)$. Define the effective directional precision*

$$\psi_I := \tau^{-2} + \frac{k}{\sigma_\xi^2}\alpha_I - \frac{k^2}{\sigma_\xi^4} b_I^\top \left(\tau^{-2}(I_d - uu^\top) + \frac{k}{\sigma_\xi^2} C_I\right)^\dagger b_I.$$

Then the conditional Bayes risk can be written exactly as

$$\mathbb{E}[(y_\star - \widehat{y}_\star)^2 \mid X_I, x_\star] = \sigma_\xi^2 + \frac{\|x_\star\|^2}{\psi_I}.$$

The last term in ψ_I is the nuisance-coordinate leakage penalty induced by selecting examples that are aligned with the query direction but not orthogonal to irrelevant directions.

Proof. Let

$$B = \tau^{-2} I_d + \sigma_\xi^{-2} X_I^\top X_I = \tau^{-2} I_d + \frac{k}{\sigma_\xi^2} A_I.$$

Decompose $\mathbb{R}^d = \mathrm{span}\{u\} \oplus u^\perp$. In this block basis,

$$B = \begin{pmatrix} \tau^{-2} + k\alpha_I/\sigma_\xi^2 & k b_I^\top/\sigma_\xi^2 \\ k b_I/\sigma_\xi^2 & \tau^{-2}(I_d - uu^\top) + k C_I/\sigma_\xi^2 \end{pmatrix}.$$

The $(1,1)$ block of B^{-1} is the inverse of the Schur complement, which is exactly ψ_I^{-1}. Since $x_\star = \|x_\star\| u$, Proposition 7.1 gives

$$\mathbb{E}[(y_\star - \widehat{y}_\star)^2 \mid X_I, x_\star] = \sigma_\xi^2 + \|x_\star\|^2 u^\top B^{-1} u = \sigma_\xi^2 + \frac{\|x_\star\|^2}{\psi_I}.$$

□

D.17 Top-k Gaussian retrieval

Corollary D.29 (Top-k Gaussian alignment quantile)**.** *Assume $x_i \sim N(0, I_d)$ and the retriever selects the k examples with largest scores $(u^\top x_i)^2$, where $2k < n$. Let $q_{n,k} := F_{\chi_1^2}^{-1}(1 - 2k/n)$, where $F_{\chi_1^2}$ is the distribution function of a χ_1^2 random variable. Then with probability at least $1 - \exp(-k/4)$,*

$$\alpha_I = \frac{1}{k}\sum_{i \in I} (u^\top x_i)^2 \ge q_{n,k}.$$

Consequently, whenever the Schur leakage term in Proposition D.28 is at most an $\eta \in [0,1)$ fraction of $\tau^{-2} + k q_{n,k}/\sigma_\xi^2$, the query risk obeys

$$\mathbb{E}[(y_\star - \widehat{y}_\star)^2 \mid X_I, x_\star] \le \sigma_\xi^2 + \frac{\|x_\star\|^2}{(1-\eta)\left(\tau^{-2} + k q_{n,k}/\sigma_\xi^2\right)}.$$

Proof. For $x_i \sim N(0, I_d)$ and fixed u, the scores $R_i = (u^\top x_i)^2$ are i.i.d. χ^2_1. With $q_{n,k} = F^{-1}_{\chi^2_1}(1-2k/n)$, the number of indices satisfying $R_i \geq q_{n,k}$ is binomial with mean $2k$. Lemma D.10 with $\delta = 1/2$ gives

$$\mathbb{P}\left(\#\{i : R_i \geq q_{n,k}\} < k\right) \leq \exp(-k/4).$$

On the complementary event, at least k examples have score at least $q_{n,k}$, so the average of the top-k scores is also at least $q_{n,k}$. This proves the alignment inequality. The risk bound follows by substituting $\alpha_I \geq q_{n,k}$ and the assumed leakage bound into Proposition D.28. □

E Experiment Details

All experiment scripts are in the supplementary material and write the CSV files from which the figures are drawn. The synthetic scripts use fixed random seeds and reproduce Figures 3, 7, and 8 without external services; the second-order certificates of Figures 3(d) and 8 are recomputed with Theorem 6.2 as stated here, on the same library and seed as in the submission. The Banking77 figures and tables are computed on a CPU from cached model outputs. The API-backed scripts cover embedding retrieval, real LLM prompt-recipe selection, and real LLM validation-leakage stress tests; they require locally configured provider credentials and store only request metadata and model outputs needed for reproducibility. The scripts report the expected number of model calls before issuing requests, and all postprocessing is CPU-only. Compute for the Banking77 study is given in Appendix B.1.

F Audit Checklist for Practitioners

Before using selected-context performance as evidence of deployment risk, a practitioner should be able to answer:

1. What data split was used to construct the final prompt transcript?
2. How many prompt recipes, retrievers, templates, and rerankers were tried?
3. Were validation examples, validation labels, model critiques, or manual edits reused across rounds?
4. Is the final context deterministic top-k, randomized retrieval, or a posterior sample?
5. What prior Π_u or candidate pool defines the KL budget?
6. Is there a fresh holdout split that was not used for any prompt-construction decision?

G Scope and Relation to Existing Bounds

Provenance of each result. The proof tools are standard, and we name them. Theorem 4.1 is Theorem 1 of Xu and Raginsky [46] applied to the empirical mean of the k selected examples; this selected-subsample step, which gives the variance proxy σ^2/k rather than one normalized by the store size, is its only nonstandard step. Theorem D.19 is an instance of the cumulant bounds of Jiao et al. [24] and Bu et al. [8]. Theorem 4.2 is approximate max-information transfer [14, Thm. 4] conditioned on the query. Proposition 4.5 is the chain rule, and Theorem 4.4 is a Rademacher maximum over a finite class, whose exact constant follows from binomial tails by the method of types [11]. Theorem 6.1 is fixed-λ Catoni-form PAC-Bayes [10]; we found no counterpart of its gate charge $\log(1/\Pi_u(\mathcal{A}))$ phrased for retrieval gating. Theorem 6.2 combines the predictable plug-in supermartingale of Waudby-Smith and Ramdas [43], built from an inequality of Fan et al. [16], with the PAC-Bayesian change of measure and a union bound over a finite grid; its countable-library variant (Remark D.27) uses the empirical Bernstein inequality of Maurer and Pontil [31]. Proposition D.25 combines the oracle inequality of Theorem 6.1 with Fano's inequality. Among the closest results, conditional supersample bounds [37] normalize by the full sample and use data-independent selection bits; variance-calibrated stability [17], information bounds for meta-learning [25], and second-order PAC-Bayes [33, 38] certify learned hypotheses against full-sample averages.

The certified object. The results can be read as an adaptive-data-analysis theorem specialized to prompt construction, but the specialization is doing statistical work. The transcript includes selected indices, prompt text, validation scores, and model-in-the-loop edits; the empirical quantity being certified is the selected-context average, not an arbitrary population loss of a learned hypothesis. This is what produces the explicit k dependence and makes repeated demonstrations, clustered chunks, and multi-round prompt edits auditable.

The mutual-information capacity is not meant to be estimated by reverse-engineering an opaque RAG system. The practical route is forward accounting: log a prior retriever, candidate gates, temperatures, noisy score releases, and validation reuse. Theorem 6.1 turns the logged KL to the prior, including the prior mass of any data-dependent gate, into a computable certificate term.

The conditional i.i.d. store assumption is the clean base case. Correlated corpora should be blocked, clustered, or evaluated with a fresh holdout; Proposition D.21 gives the simplest effective-block version. Finally, the theory is not pessimistic about useful adaptation: Proposition D.22 separates task-identification information from within-task overfitting, and the linear-regression application shows how good retrieval can improve prediction while still requiring stable certification.

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 Section 1 state the context-capacity formulation, the selected-context validity bounds and their sharpness, the PAC-Bayesian retrieval rules, the Banking77 and synthetic experiments, and the linear-regression instantiation; the contributions paragraph names the result behind each claim, and the scope is stated in Section 1 and Appendix A.4.

   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: Appendix A.4 discusses the conditional i.i.d. store assumption, dependence beyond clustered stores, finite-precision capacity accounting, the boundary for deterministic pipelines (Remark 4.6), and the single-task scope and conservatism of the Banking77 study; Table 4 states when each assumption holds in deployment.

   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: Each result states its assumptions (Sections 3–7). Every numbered result has a complete proof in Appendix D, and the auxiliary inequalities taken from the literature are stated there with citations.

   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: [Yes]

   Justification: Section 5 and Appendices B.1 and E specify the Banking77 protocol (splits, recipe library, models, temperatures, resampling, and certificate parameters) and the synthetic and API-backed protocols. The supplementary material contains the scripts, fixed seeds, cached model outputs, and the CSV files behind every figure and table.

   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: [Yes]

   Justification: The code, the cached Banking77 model outputs, and the reproduction commands are in the supplementary material. Banking77 is public, and every figure and table is recomputed on a CPU from the cached outputs.

   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: [Yes]

   Justification: No model is trained. The splits (store of 2,000, validation pool of 256, holdout of 1,024 or 512), the 64-recipe library, the decoding temperature, $\lambda^\star$, δ, β, the number of resamples, and the seeds are given in Section 5 and Appendices B–E.

   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: Figure 2(a) shows 95% intervals across validation draws. The Banking77 tables report means and 95th percentiles over 400 resampled selection events per cell, the fraction of 4,000 further draws certified below zero-shot, and the standard errors of the zero-shot holdout estimates; the events reuse fixed validation pools, as stated in Appendix B.1. Synthetic figures average 2,500 to 4,000 trials per point.

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: [Yes]

   Justification: Appendix B.1 reports the Banking77 compute: one T4 GPU for the embeddings, nine parallel H100 jobs for the 87,040 Qwen2.5-7B-Instruct calls (11 minutes), and 18 minutes for the 52,224 GPT-4.1-mini API calls. All other experiments and analyses run on a CPU (Appendix E).

   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 paper uses synthetic data, fictional-label tasks, and the public Banking77 benchmark. It involves no human subjects, private or scraped personal data, or model release.

   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: Appendix A.5 discusses more honest evaluation of RAG and prompt engineering and the risk of performative certification when prompt-construction logs omit selection steps.

    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 releases no model weights, scraped dataset, or other high-risk artifact.

    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: Banking77 (CC BY 4.0), bge-small-en-v1.5 (MIT), Qwen2.5-7B-Instruct (Apache 2.0), and vLLM (Apache 2.0) are cited with their licenses in Appendix B.1. GPT-4.1-mini and the OpenAI models of the API-backed protocols are used through the OpenAI API under its terms of use.

    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 released assets are experiment scripts, cached model outputs, and generated CSV files, documented in Appendix E and in the README of the supplementary material.

    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 paper does not involve crowdsourcing or research with 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 paper does not involve research with human subjects.

    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: LLMs are the objects of study: frozen Qwen2.5-7B-Instruct and GPT-4.1-mini are the in-context learners certified in the Banking77 study (Section 5 and Appendix B.1), and an OpenAI LLM is used in the API-backed stress tests (Appendix B). The theoretical method does not use an LLM as an algorithmic component.

    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.