PROMPT SEARCH IS NOT PROMPT EXPRESSIVITY: CONTINUOUS PROMPTS BEAT GEPA, MIPROV2 AND SIMBA, AND AN EXACT VOCABULARY FRONTIER

Anonymous authors
Paper under double-blind review

ABSTRACT

When a prompt optimizer fails to find a good text prompt, it is tempting to conclude that none exists. We separate the two questions and give each its own evidence. Empirically, under a protocol frozen before any test row was scored, a 32-vector continuous prompt on three frozen 3–4B models achieves higher test accuracy than GEPA, MIPROv2 and SIMBA in all 18 model–dataset–regime cells (margins 7.8–34.3 points; all 18 H1 hypotheses rejected by the pre-registered procedure, with Holm adjustment over the 36 H1/H2 hypotheses at nominal level 0.05), and its lowest-accuracy run exceeds the highest-accuracy individual optimizer run in 17 of 18 cells. The ordering holds with every arm scored by one fixed decoder, and when GEPA and MIPROv2 receive the selected continuous configuration's example budget (still 5.9–18.4 points behind in the three LOW cells tested). Audits on frozen records document the scoring differences and show that removing normalised exact train–test duplicates changes the gaps by at most 0.22 points; the comparison is one of resource-unmatched pipelines with different search objectives, so it measures what the pipelines achieve, not what text can express. Theoretically, a representation limit can nonetheless be exact. In a fully specified coupled Gaussian attention model, the risk of the best legal text prompt for a $\tanh$ target converges to $q - \theta^2 + (\theta - s)_+^2$ in the dictionary's reach s, uniformly over all query-independent masses, so for $s < \theta$ no sequence of prompt lengths removes the reach penalty; the model and proof are given in full. A controlled block checks a finite paired-response certificate and shows which premise carries the gap. In a pretrained model the vocabulary constraint is exact at the first layer, where the median ratio of token to soft directional reach over sampled readouts is 11.7–12.5%; over continuation classes satisfying our stated label-consistency and readout assumptions, such task-agnostic first-layer quantities alone cannot yield a positive task-loss lower bound.

1 INTRODUCTION

A practitioner adapting a frozen classifier must choose between searching text prompts and fitting a few continuous vectors. Prompt optimizers such as GEPA (Agrawal et al., 2025), MIPROv2 (Opsahl-Ong et al., 2024) and DSPy's SIMBA (Khattab et al., 2023) search over instructions and demonstrations with a strong language model as the proposer; continuous prompt tuning (Lester et al., 2021) fits vectors by gradient. On three frozen 3–4B models a 32-vector continuous prompt leads all three optimizers by 7.8–34.3 accuracy points in each of 18 settings (Section 5.1). Such a gap invites two readings: that discrete search is hard, or that the task needs something text cannot say. Both conflate *what a search pipeline achieves* with *what the best legal text prompt can express*; an optimizer's shortfall alone does not establish a representation limit.

We answer the questions separately, with evidence of different kinds.

- **What the pipelines achieve** (Section 5.1). A pre-registered comparison on public intent and question-type benchmarks: the continuous prompt beats every text optimizer in all 18

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

cells; the lowest continuous run exceeds the best individual optimizer run in 17 of 18 cells, and the ordering survives duplicate removal, a per-class breakdown, one common decoder for every arm, and an approximately fourfold example budget for GEPA and MIPROv2 in three LOW cells. It is a comparison of pipelines — different objectives, budgets and search spaces — and we state what it does not establish.

- **What the best text prompt can express** (Section 3.1; controlled block, Section 5.2). An exact frontier in a coupled Gaussian attention model for every odd target: the best legal prompt is a slope-capped linear probe, $\mathcal{R}_f(s) = \|f\|^2 - \theta_f^2 + (|\theta_f| - s)_+^2$, uniformly over all query-independent masses, with matching legal integer attainment at the stated sufficient length. Length cannot substitute for vocabulary. The model and complete proof are in Appendix B; a controlled block checks a finite paired-response certificate and identifies which premise carries it.
- **What can be ruled out in a pretrained model** (Sections 3.2 and 5.3). At the first layer the vocabulary constraint is exact and certifiable; over continuation classes satisfying our stated label-consistency and readout assumptions, task-agnostic first-layer quantities alone cannot yield a positive task-loss lower bound (Proposition 7).

Contribution. Empirically, a frozen-protocol comparison, not a new prompt-tuning algorithm. Theoretically, an exact vocabulary-dependent risk frontier in a specified coupled Gaussian attention model, uniform over dictionary-adaptive, query-independent masses, with legal integer attainment and its threshold at $s = |\theta_f|$; it differs from fixed-length cardinality separations (Petrov et al., 2024a) and Gaussian-population attention linearity (Castin et al., 2025; Boursier & Boyer, 2025). The Birkhoff–Hopf coefficient (Birkhoff, 1957), two-atom extreme-point structure and the ingredients of the joint convergence are classical; we claim none of them.

Practical takeaway. On a frozen classifier with gradient access, fit a 32-vector continuous prompt alongside text search and retrieval, and select on held-out accuracy and measured resources: a text optimizer's failure is not evidence that no text prompt exists, and where a vocabulary limit can be certified, it is at the first layer (Section 3.2).

2 Related work, and what it pins down

Hsu & Lai (2026) establish soft-prompt approximation guarantees, even for a random transformer, and note the optimization difficulty of hard prompting; our result addresses the complementary question of the best discrete-class risk in a different, explicitly specified model, and does not contradict that observation. Petrov et al. (2024a) separate one continuous virtual token from any discrete token by *cardinality* at fixed length; ours is a statement about *risk*, uniform over all lengths and fixed by the vocabulary's reach, and in a model that reads every token no counting argument could replace it. Our model is the one-layer scalar instance of their equation (7) with prefix values confined to a vocabulary. Wang et al. (2023) give a qualitative "no length suffices" for soft prompts on a constructed model, and Meyer et al. (2025) a limit on the prompt-length axis; ours is on the vocabulary axis, and by Proposition 4 length cannot buy vocabulary. Smith et al. (2026) expand a pretrained model's tokenizer and adapt its weights for encoding efficiency; that differs from our fixed-backbone dictionary change, and our result does not predict its gains. On the empirical side, Erkan et al. (2026) use soft prompts to remove format-following failures from benchmark evaluation; our common label-constrained scorer addresses the same confound for the final comparison, and our audit documents the remaining scoring differences. The full comparison, including what Petrov et al. (2024b) imply for fixed pretrained models, is in Appendix K.

3 Theory

3.1 Setting and the closed form

An ordering among configured pipelines (Section 5.1) is not a bound on the best text prompt. We ask a separate mathematical question: in an explicitly specified attention model, what is the risk infimum over every legal, query-independent prompt?

We work in a coupled Gaussian attention model, specified in full in Appendix B: with $D = r^2$, the query, key and value matrices $W_Q, W_K, W_V \in \mathbb{R}^{D\times D}$ have independent $N(0, r^{-2})$ entries, the vocabulary embeddings follow the prescribed coupled Gaussian law, and a unit direction u is selected from the backbone before the vocabulary is drawn (Fail_r is the event that this selector fails). A dictionary atom is a token's fixed attention parameters (k_j, s_j, c_j): query-logit coefficient, logit offset and scalar value. A prompt induces nonnegative masses a_j (integer counts for a token prompt, of length $\sum_j a_j$), and on a fresh query $Y \sim N(0, I_r)$ the head outputs

$$F_a(Y) = \frac{e^{g(Y)} h_0(Y) + \sum_j a_j e^{s_j + k_j^\top Y/\sqrt{r}}\, c_j}{e^{g(Y)} + \sum_j a_j e^{s_j + k_j^\top Y/\sqrt{r}}},$$

where $g(Y)$ and $h_0(Y)$ are the query's own self-attention logit and value. The target is $f(Z)$ with $Z = u^\top Y$, and $\|f\|^2 = \mathbb{E} f(Z)^2$. A *legal* prompt uses only norm-legal atoms ($\|E_j\| \le 2r$; $\mathcal{A}_r$ is the set of such mass vectors), integer counts and a length cap. The masses may depend on the dictionary but not on the fresh query — this is the constraint that makes a prompt a prompt rather than a per-example parameter, and it is the constraint whose removal we later use as a control.

Theorem 1 (Vocabulary-risk frontier)**.** *Use the original model, norm-legal prompt classes and allowable selector of Appendix B, with single integer scale r (latent dimension, width through $D = r^2$, score normaliser, query amplitude and vocabulary scale). Let $r \to \infty$, $D = r^2$, $V_r \to \infty$ and $s_r = \log V_r/\sqrt{r} \to s \in [0, \infty)$. Fix an odd $f \in L^2(N(0,1))$ and define $\theta_f = \mathbb{E}[Z f(Z)]$ and*

$$\mathcal{R}_f(s) \;=\; \|f\|^2 - \theta_f^2 + \big(|\theta_f| - s\big)_+^2 \;=\; \min_{|d|\le s} \mathbb{E}\big(f(Z) - dZ\big)^2.$$

Let $\mathcal{F}_r$ contain the frozen backbone and dictionary. For every fixed $\epsilon > 0$, probability over their joint draw satisfies $\Pr\big(\mathsf{Fail}_r \cup \{\inf_{a\in\mathcal{A}_r} \mathbb{E}_Y[(F_a(Y) - f(u^\top Y))^2 \mid \mathcal{F}_r] < \mathcal{R}_f(s) - \epsilon\}\big) \to 0$. The masses may depend on $\mathcal{F}_r$, but not on the fresh query. This lower bound holds for every cap sequence, and a measurable legal integer witness of total length at most $\lceil 2r^2 V_r \rceil$ attains it: the real-mass and integer-count optima, and the length-capped integer optima for every cap $L_r \ge \lceil 2r^2 V_r \rceil$, converge jointly in probability to $\mathcal{R}_f(s)$.

Corollary 2 (The $\tanh$ case)**.** *For $f = \tanh$, $\theta = \mathbb{E}[Z \tanh Z] = 1 - q$ with $q = \mathbb{E}\tanh^2 Z$ (integration by parts), and $\mathcal{R}(s) = q - \theta^2 + (\theta - s)_+^2$; $q = 0.39429449$, $\theta = 0.60570551$.*

Corollary 3 (An optimal prompt attains a slope-capped linear response)**.** *There exists an asymptotically optimal legal integer-prompt sequence whose response converges in query-L^2, in probability, to $d_* Z$, where $d_* = \operatorname{sign}(\theta_f)\min\{|\theta_f|, s\}$. The vocabulary threshold is $|\theta_f|$, the target's first Hermite coefficient, and the floor $\|f\|^2 - \theta_f^2$ is the target's squared L^2 distance from its best linear approximation, its energy outside degree one. In the same asymptotic regime a mean-zero even target component contributes its full L^2 energy to the uniform lower bound, up to $o_P(1)$ (Appendix C).*

Proof sketch. Parity splits the risk into even and odd parts; the odd part of every prompt's response is $Z\, d_a(X) + o(1)$ in L^2, uniformly in the masses, with a slope clipped at $\rho_r \to s$, a step that does not see the target; and $\mathbb{E}(f(Z) - Zd)^2 = \|f\|^2 - \theta_f^2 + (d - \theta_f)^2$ for every clipped slope. The witness of the $\tanh$ case has one free parameter, the tilt slope, and its response converges to $d_* Z$ for any prescribed $d_* \in [0, s)$, whatever the target. Appendix C gives the proof relative to the $\tanh$ case.

The positive-part term carries the content. For $s \ge |\theta_f|$ it vanishes and the achievable risk equals the linear floor $\|f\|^2 - \theta_f^2$: the vocabulary is rich enough to buy the best linear probe, and in this finite-s regime nothing beyond a linear probe is bought. For targets with nonzero higher-order odd Hermite energy the floor is positive — for $\tanh$ it is $q - \theta^2 = 0.0274$ — so what a rich vocabulary removes is the *reach penalty* $(|\theta_f| - s)_+^2$, not the error; a linear target has zero asymptotic risk once its slope is reachable. For $s < |\theta_f|$ the penalty is strictly positive and depends on s alone — not on the prompt length, not on the optimizer, not on the compute budget. The quantifiers are ordered: the masses are fixed, possibly after seeing the dictionary, before the fresh query is drawn.

Proposition 4 (Length does not substitute for vocabulary)**.** *Below the threshold $s < |\theta_f|$, no sequence of prompt lengths removes the asymptotic reach penalty $(|\theta_f| - s)^2$, for every odd target.*

At finite r a larger length cap can still reduce integer-realization or optimization error. Section 5.2 checks a different, finite no-self paired-response certificate; it does not test Proposition 4.

The exact two-query response ceiling, an observable-dependent refinement of the Birkhoff–Hopf bound, is given in Appendix A; deterministic finite-r lower functionals are Theorem B.3.

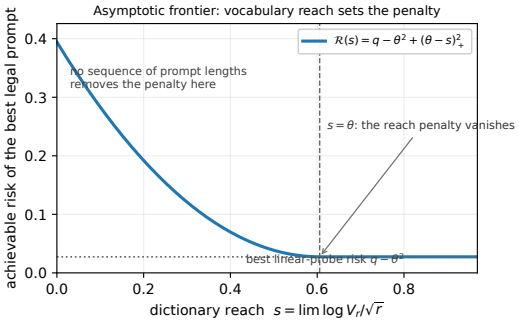


Figure 1: The analytic $\tanh$ frontier of Corollary 2: the minimum asymptotic squared risk $q-\theta^2+(\theta-s)_+^2$ against the dictionary reach $s = \lim \log V_r/\sqrt{r}$, a closed form, not a fit or experimental data ($q = 0.39429449$, $\theta = 0.60570551$ by 200-node Gauss–Hermite quadrature; $\theta = 1 - q$ holds to eight decimals). Left of $s = \theta$ no sequence of prompt lengths removes the reach penalty $(\theta - s)^2$; at and above it the risk is the linear-probe floor $q - \theta^2 = 0.02741533$.

3.2 The vocabulary limit is exact at the first layer

The vocabulary constraint is exact at the first layer of the frozen model, where every token prompt's keys and values are vocabulary items, and there it is a theorem with a measured, certified size (§3.2–3.3); under Proposition 7's assumptions, task-agnostic first-layer quantities alone cannot yield a positive task-loss lower bound (§3.4); at the output, the pre-registered confirmation measures the teacher–discrete gap and what length buys, and we do not call that a vocabulary limit (§5.4).

Fix a decoder-only transformer whose first attention layer forms, per key–value group, unrotated keys $k_t = W_K\, n(E_t) + b_K$ and values $v_t = W_V\, n(E_t) + b_V$ from the normalised embedding $n(e) = \varphi(\|e\|)\,(g \odot e/\|e\|)$ $(n(0) = 0)$, $\varphi(\rho) = \sqrt{d}\,\rho/\sqrt{\rho^2 + d\epsilon}$ increasing (RMSNorm with gain g), rotates queries and keys by position (RoPE) and masks causally. A prefix occupies L positions; a content query sits at a later position i; $o_{\text{live}}(i)$, $D_{\text{live}}(i)$ are the head output and softmax mass of the prompt-free context, prompt-independent at the first layer, and b_P the softmax average of the prefix values. Soft rows are norm-capped at $\text{cap} = \max_t \|E_t\|$ and pass through the same normalisation, as in the protocol. The vocabulary Σ is nonempty and $\epsilon > 0$; for $L = 0$, $o_P = o_{\text{live}}$ and $\lambda_P = 0$; $\hat{E}_t = 0$ when $E_t = 0$, and such a token contributes 0 to the maximum defining $\kappa(w)$.

Theorem 5 (First-layer reach of the token class)**.** *(i) For every token prompt of every length L, $o_P(i) = o_{\text{live}} + \lambda\,(b_P - o_{\text{live}})$ with $b_P \in \text{conv}\{v_t : t \in \Sigma\}$ and $0 \le \lambda \le \lambda_{\max}(L, i) = Le^{s^*}/(Le^{s^*} + D_{\text{live}})$, s^* the largest score of any (token, prefix position) at query i; in particular $o_P(i) \in \text{conv}(\{o_{\text{live}}(i)\} \cup \{v_t\})$. (ii) For norm-capped soft rows, $b_P \in \mathcal{E} = b_V + \varphi(\text{cap})\, W_V\, \text{diag}(g)\, B$ (B the unit ball, $\mathcal{E} \supseteq \text{conv}\{v_t\}$) and the same decomposition holds, with $\lambda_P \le Le^{s^*_{\text{soft}}}/(Le^{s^*_{\text{soft}}} + D_{\text{live}})$, $s^*_{\text{soft}}(L, i)$ the supremum of the score over norm-capped soft embeddings and allowed prefix positions; the token s^* is not asserted to bound soft-prefix scores. (iii) For every readout direction w of the head,*

$$\sup_{\text{token prompts, all } L} \langle w, b_P - b_V\rangle \;=\; \kappa(w)\cdot\varphi(\text{cap})\,\|g \odot W_V^\top w\|,$$

$$\kappa(w) \;=\; \max_{t\in\Sigma} \frac{\varphi(\|E_t\|)}{\varphi(\text{cap})}\,\cos\angle\big(E_t, g \odot W_V^\top w\big),$$

with $\kappa(w) = 0$ when $\varphi(\text{cap})\,\|g \odot W_V^\top w\| = 0$, while a single soft token attains $\varphi(\text{cap})\|g \odot W_V^\top w\|$; the token supremum is attained at length one.

The token-value convex envelope is independent of prefix length; the actual attention mass and attainable mixtures b_P can depend on length, positions and the token sequence, so we assert a length-independent enclosure and an exact directional support maximum, not equality with the attainable set. For each head, length buys mass, not reach (it does not enlarge that head's token-value convex envelope; with several heads the joint reachable set is contained in the product of the head-wise enclosures), and the reach is the extreme-value quantity $\kappa(w)$, whose reference value for a fixed direction independent of an i.i.d. uniform spherical token cloud of equal norms, with $\log V = o(d)$, is $\sqrt{2\log V/d}$ — the mechanism of Theorem 1. The hypotheses hold for Qwen2.5, Falcon3, Llama-3, SmolLM2 and Granite-3 and fail for absolute-position embeddings, where values depend on position (proofs: Appendix D).

3.3 A CERTIFIED FLOOR FOR THE FIRST-LAYER RESPONSE OF EVERY TOKEN PROMPT

Proposition 6 (Certified first-layer floor)**.** *Under the hypotheses of Theorem 5, for a content query i, a head and any fixed target output o_T, every token prompt of length at most L satisfies $\|o_P(i) - o_T\| \geq \operatorname{dist}(o_T, o_{\text{live}} + [0, \lambda_{\max}(L, i)](\operatorname{conv}\{v_t\} - o_{\text{live}})) \geq \operatorname{dist}(o_T, \operatorname{conv}(\{o_{\text{live}}\} \cup \{v_t\}))$, and any direction $w \neq 0$ with $\langle w, v_t \rangle \leq \beta$ for all t and $\langle w, o_{\text{live}} \rangle \leq \beta$ certifies the last distance from below by $(\langle w, o_T \rangle - \beta)_+ / \|w\|$.*

This bounds the whole legal class without an optimiser, for the quantity the compiler J minimises and the fidelity contrasts measure; it says nothing about task loss (§3.4).

3.4 WHY THE FIRST LAYER IS WHERE THE THEOREMS STOP

Task loss here is the normalized multiclass Brier loss $B(p, y) = \frac{1}{2} \sum_{k=1}^{K} (p_k - \mathbf{1}\{k = y\})^2$, averaged over labelled examples.

Proposition 7 (Conditional non-identifiability from first-layer geometry)**.** *Fix a legal prompt p^* and a nonempty finite labelled set $\mathcal{D}$ with K classes; let h_x be the first-layer representation of (p^*, x) at the readout position. Let $\mathcal{N}$ be a class of causal networks that share the frozen model's token embeddings and first layer and include all readouts of h_x by a linear projection, one hidden ReLU layer and a softmax, with unrestricted real weights and biases. If (i) $h_x = h_{x'}$ implies $y_x = y_{x'}$ on $\mathcal{D}$, then for every $\tau > 0$ some network in $\mathcal{N}$ of hidden width at most $\max\{1, |\mathcal{D}| - 1\}$ gives p^* empirical Brier loss at most $\frac{1}{2} K(K-1) e^{-2\tau}$ on $\mathcal{D}$. Conversely, if (i) fails on c disjoint pairs, every readout of h_x alone has loss at least $c/(2|\mathcal{D}|)$. Consequently, under (i), the task-agnostic first-layer quantities of Theorem 5, Proposition 6 and the two-query ceiling of Appendix A alone do not yield a strictly positive empirical Brier-loss lower bound uniformly over $\mathcal{N}$; this does not rule out task-specific bounds that use further properties of the frozen downstream network.*

The proposition varies the downstream network while holding the embeddings and first layer fixed; it does not construct a better prompt for the actual frozen model. Condition (i) is weaker than distinct representations and checkable for a given prompt and data set; the constructive proof (Appendix D) uses weights of order τ over the smallest projected gap, so it bounds what first-layer quantities can exclude, not what the frozen network does. The confirmation below shows the corresponding empirical pattern: better measured first-layer fidelity need not rank fitted methods by task loss.

4 EXPERIMENTAL SETUP

Optimizer comparison (E2). We ask the practitioner's question on public benchmarks: on a frozen model, does a 32-vector continuous prompt achieve higher accuracy than the text-prompt optimizers in current use? This measures what configured pipelines achieve. It is not a test of Theorem 1, Proposition 4 or Theorem 5, and no number here is a vocabulary limit.

Protocol. All of the following was fixed in a protocol frozen before any test row was scored (2026-09-22, SHA-256 `be5132ac...`); the evaluator refuses the test split under any other text. *Cells* (18): three frozen ≈3B instruction models (Qwen2.5-3B-Instruct, Phi-4-mini-instruct, Granite-4.1-3B) × three datasets (Banking77 (Casanueva et al., 2020), 77 intents, 3,080 test rows; MASSIVE en-US (FitzGerald et al., 2022), 60, 2,974; TREC coarse (Li & Roth, 2002), 6, 500) × two regimes (LOW, 10 training examples per class, 25 on TREC; FULL, the whole pool), with a 300-row selection split S. *Arms*: the continuous prompt (Lester et al., 2021), $L = 32$, three seeds, its rate, initialisation and trainer chosen on S; GEPA (Agrawal et al., 2025) and MIPROv2 (Opsahl-Ong et al., 2024) at `heavy` and SIMBA, from DSPy 3.3.1 (Khattab et al., 2023), with an external reflection model, three seeds each; zero-shot; one demonstration per class; and kNN in-context learning (Liu et al., 2021) with 16 or 64 demonstrations retrieved *per test query*. The pre-registered policy arm ("ours" in the protocol) is the continuous prompt, except in the eight cells where it did not beat kNN on S and a hybrid did better on S; the hybrid is the same 32 vectors trained in front of a plain block of the 16 retrieved neighbours (one seed).

Objectives and scoring. The objectives differ, and we state them per stage. The text optimizers search and select with the task model's free generation scored by exact match (GEPA also receives

the gold label as textual feedback); GEPA and MIPROv2 select on S, SIMBA on its training pool and never sees S. The continuous prompt trains on the gold label's likelihood and selects its checkpoint by label-constrained accuracy on S. On test every arm is scored by the same label-constrained scorer, except that a text arm is scored by native generation wherever S preferred that. A post hoc audit rescored every original E2 text program with the continuous prompt's decoder; under that one fixed decoder the continuous prompt still leads every optimizer in all 18 cells (Appendix I; the budget-expanded programs below are scored natively). This removes final-output-format confounding at test but does not align the search objectives or establish that formatting had no effect on candidate discovery.

Controlled block. This is a premise-ablation study of the finite no-self certificate of Appendix A, not a numerical test of Theorem 1. *No-self* means that attention averages dictionary values without the query's own self-attention term; a is a scalar key–value coupling, not a mass vector; and the balanced witness puts equal mass on the two extreme atoms, producing $\tanh(aZ)$ against the target $\tanh Z$. Each intervention breaks one premise — length, dictionary support, coupling or query independence — to show which one the certificate rests on. Its cells are checks of one computation, not independent pieces of evidence.

Design: 16 development dictionaries and 64 confirmation dictionaries, two laws (Rademacher, Gaussian) and three coupling values $a \in \{0.25, 0.5, 1\}$, giving 96 development and 384 confirmation base cells; the interventions below are evaluated within those cells. The 96 development cells were independently re-derived by code sharing nothing with the module under test — a different ratio optimiser, a different quadrature family, and closed forms written from the protocol statement rather than from the implementation. Realised risks of the balanced witness are in Table 3 (Appendix E).

Compiler study on a pretrained model. The floor measurements and the confirmation use a separate teacher-to-token compilation study on Qwen2.5-3B-Instruct and CLINC150: a continuous teacher C is fitted per domain and seed, and three compilers O, M and J turn it into token prompts, J by also fitting its first-layer response (Appendix G); it is not a test of E2's text optimizers.

5 Results

5.1 What the pipelines achieve: continuous prompts against text optimizers

Pre-registered outcome. Each cell tests two intersection–union hypotheses. H1: the continuous prompt beats each of GEPA, MIPROv2 and SIMBA. H2: the policy arm beats each of the seven text methods. Each comparison is a one-sided paired bootstrap over test rows (10,000 resamples, seed 24021, seed-averaged per-row correctness), a cell's p is the maximum over its comparators, and Holm runs over all 36 hypotheses. The pre-registered percentile-bootstrap procedure with Holm adjustment **rejects all** 36 at nominal level 0.05 (not an exact finite-sample family-wise guarantee); the largest Holm-adjusted p is 0.0164. The continuous prompt leads the best optimizer by 0.078–0.343, every lower 95% endpoint at least 0.055 (Table 1). Inference conditions on the fitted artifacts; the three-seed comparisons below are descriptive and do not make the row-bootstrap intervals cover retraining uncertainty.

How far the ordering holds (post hoc, frozen records). These diagnostics use the frozen test records; they fit nothing, select nothing and do not alter the pre-registered analysis, and their intervals are pointwise and conditional on the fitted artifacts. *Seeds.* The lowest-accuracy continuous-prompt run exceeds the highest-accuracy individual optimizer run in 17 of 18 cells; the exception is Phi TREC LOW (0.812 against GEPA's 0.844). *Duplicates.* Normalised exact test texts found in the training pools number 7, 21 and 11 on the three datasets; excluding them changes every H1 and H2 gap by at most 0.0022. *Classes.* Against the cell's best optimizer the continuous prompt is strictly better on a median 83% of classes (minimum 56%) and worse on a median 17%; class recall and the optimizers' payloads are analysed in Appendix H.

Retrieval is a different class, and the hybrid is a pipeline comparison. The bare continuous prompt has a higher point-estimate accuracy than every text baseline in $11/18$ cells; the best kNN

Table 1: Test accuracy (fractions), 18 cells, pre-registered analysis. B77: Banking77; MAS: MASSIVE en-US; LOW: 10 training examples per class (25 on TREC); FULL: the whole pool. Soft: continuous prompt, $L = 32$, mean of three seeds (worst seed). Best opt.: best of GEPA, MIPROv2, SIMBA (three-seed means). Policy: the S-selected arm, the continuous prompt unless marked. Δ_{H1} = soft − best opt.; Δ_{H2} = policy arm − best text method; 95% paired percentile intervals, best comparator recomputed per resample. ^{h}: hybrid (one seed). One-sided p per comparator is the share of the 10,000 paired resamples in which the difference is ≤ 0; a cell's p is the maximum over its comparators. No resample had a difference ≤ 0 (zero of 10,000, a Monte Carlo count, not an exact zero probability or an upper bound $p < 10^{-4}$) for every H1 and for every H2 except four: Granite B77 LOW/FULL (Holm 0.0081), Phi B77 LOW (0.0164) and Phi MAS FULL (0.0004); all six Qwen cells are in the zero group.

model	data	reg.	soft	best opt.	Δ_{H1} [95% CI]	policy	best text	Δ_{H2} [95% CI]
Qwen	B77	LOW	.695 (.681)	MIPROv2 .574	+.121 [.106,.136]	.871^h	kNN-64 .784	+.087 [.073,.101]
Qwen	B77	FULL	.829 (.802)	MIPROv2 .557	+.272 [.256,.286]	.934^h	kNN-16 .878	+.056 [.045,.065]
Qwen	MAS	LOW	.783 (.764)	SIMBA .675	+.108 [.095,.121]	.783	kNN-64 .751	+.032 [.016,.042]
Qwen	MAS	FULL	.850 (.805)	MIPROv2 .674	+.176 [.162,.189]	.897^h	kNN-64 .855	+.042 [.031,.051]
Qwen	TREC	LOW	.872 (.838)	MIPROv2 .747	+.125 [.091,.160]	.872	kNN-16 .804	+.068 [.030,.096]
Qwen	TREC	FULL	.968 (.964)	SIMBA .756	+.212 [.181,.240]	.968	kNN-64 .872	+.096 [.067,.125]
Phi	B77	LOW	.767 (.744)	MIPROv2 .557	+.209 [.193,.223]	.843^h	kNN-16 .829	+.013 [.001,.026]
Phi	B77	FULL	.910 (.905)	MIPROv2 .568	+.343 [.327,.359]	.935^h	kNN-16 .912	+.023 [.015,.032]
Phi	MAS	LOW	.792 (.784)	MIPROv2 .536	+.256 [.240,.269]	.802^h	kNN-16 .751	+.051 [.037,.065]
Phi	MAS	FULL	.879 (.875)	SIMBA .625	+.254 [.241,.268]	.879	kNN-64 .856	+.023 [.013,.034]
Phi	TREC	LOW	.889 (.812)	GEPA .782	+.107 [.077,.136]	.889	GEPA .782	+.107 [.077,.136]
Phi	TREC	FULL	.972 (.970)	GEPA .816	+.156 [.131,.183]	.972	GEPA .816	+.156 [.129,.178]
Granite	B77	LOW	.830 (.827)	MIPROv2 .636	+.194 [.179,.209]	.873^h	kNN-16 .858	+.015 [.004,.026]
Granite	B77	FULL	.913 (.905)	SIMBA .628	+.285 [.269,.299]	.940^h	kNN-16 .930	+.010 [.002,.016]
Granite	MAS	LOW	.817 (.813)	SIMBA .691	+.125 [.112,.138]	.817	kNN-16 .779	+.038 [.025,.048]
Granite	MAS	FULL	.886 (.885)	MIPROv2 .692	+.194 [.181,.207]	.886	kNN-16 .864	+.023 [.013,.032]
Granite	TREC	LOW	.949 (.942)	GEPA .871	+.078 [.055,.102]	.949	GEPA .871	+.078 [.055,.102]
Granite	TREC	FULL	.931 (.860)	GEPA .829	+.103 [.075,.129]	.931	kNN-64 .850	+.081 [.047,.103]

point estimate is higher in all six Banking77 cells and in Qwen MASSIVE FULL. Its context depends on the query, which a prompt in our setting may not. In the eight cells where the policy selected the hybrid, the single-seed continuous-prefix pipeline outperformed the DSPy kNN-16 program by 0.010–0.149 with the same 16 retrieved neighbours (exact two-sided McNemar $p < 0.05$ in seven, 0.0505 in the eighth; Holm over these eight post hoc tests still rejects seven). Because the demonstration wrapper also differs, this does not isolate the contribution of the trained prefix, and training-seed robustness of the hybrid is not measured.

Accuracy–resource trade-off. The pipelines were not resource-matched. The 7.8–34.3-point H1 gains use 3.5–7.1 times GEPA's and 3.9–7.9 times MIPROv2's unweighted task-example presentations for the selected continuous configurations, and 9.5–26 times with their configuration search; the continuous prompt uses gradients but no external reflection model, which the text optimizers call (median 31–114 times per cell), and SIMBA used more task-model rows than the selected continuous prompt in half the cells. These counts do not match compute; the full ledger, with wall-clock and the hybrids' training, is in Appendix H. The observed advantage cannot be attributed separately to prompt representation, gradient access, optimization objective, selection procedure or resource allocation.

Giving the optimizers the continuous prompt's budget. In an addendum frozen before its runs, GEPA and MIPROv2 were rerun in three LOW cells, one per model and dataset (Granite MASSIVE, Phi TREC, Qwen Banking77; seeds 1–3), with the selected continuous configuration's allowance of 13,800 task-example presentations per seed (excluding its configuration search), about four times their default; they used 13.9–14.2k and 12.2–12.3k rows. Everything else was unchanged, and the

18 programs were frozen before one sealed test batch. The larger budget moved their test accuracy by -2.7 to $+9.1$ points; the continuous prompt still leads every budget-expanded optimizer, by 5.9–18.4 points, with every paired 95% interval above zero (Table 9; how the pipelines respond to more training data is in Appendix H). This is robustness to the example-presentation budget, not compute matching: the reflection model's calls are not matched.

One decoder for every arm, and certified search shortfall. Post hoc, all 306 text-method programs were scored on test by the label-constrained decoder of the continuous prompt (Table 10). Under this one fixed decoder the continuous prompt leads the best optimizer by 0.077–0.343 in all 18 cells (lowest paired 95% endpoint 0.051) and beats every non-retrieval text method in $18/18$. Two non-prompt references bracket the arms: the continuous prompt beats a per-class bias on the zero-shot label scores in $18/18$ cells and lies within 0.03 of a linear probe of the last-position state in $12/18$; on the intent datasets all three optimizers fall below the bias in $12/12$ cells, and one fixed demonstration per class beats the best optimizer in 10. For a fixed decoder, a scored legal prompt lower-bounds the best attainable empirical accuracy of the prompt class, so the best fixed text prompt among those scored (zero-shot, one-shot or another optimizer's program) certifies that each optimizer fell short of that class by up to 0.159 on the intent datasets (positive in 34 of 36 optimizer–cell pairs). A representation gap in accuracy would instead need an *upper* bound on that optimum, equivalently a lower bound on the minimum task loss; evaluating finite sets of prompts does not supply one.

Controls, incidents and what is not claimed. PEZ (Wen et al., 2023), at one untuned rate, falls below zero-shot in $18/18$ cells, and $L = 128$ at the $L = 32$ rate falls below $L = 32$ in $16/18$: outcomes under untuned finite-budget configurations, not estimates of the corresponding optima or tests of Proposition 4. One planned test evaluation was made per selected artifact; two technical re-evaluations and one unused accidental evaluation of a non-selected hybrid (0.786 against the selected 0.783) are logged (Appendix H). We do not claim: a vocabulary limit at the output; any prediction by the theory of the gaps' size or order; that the bare continuous prompt beats every text method; robustness of the hybrid beyond one seed; or any comparison at matched compute, or against LoRA or fine-tuning.

5.2 Controlled block: breaking one premise at a time

The right atoms are absent, not unfound. The usual way to argue "representation, not search" is from a gap that resisted optimisation — an argument from failure. Here it is a theorem: the interior atoms are confined to $(-0.9a, 0.9a) \times (-0.9, 0.9)$ by construction, so the extrema dominate every pairwise term, and the no-self block certificate of Appendix A, which follows from eq. (1) for any symmetric query law, certifies the balanced-extrema witness optimal over *all* nonzero finite non-negative masses of *any* support, which a pair search could never establish. The search only corroborates it: all $2{,}016$ atom pairs on a grid of 241 mass ratios, in every cell, improve on the witness by at most 6.4e-16 over 384 confirmation cells. Support repair inserts the target pair at unchanged dictionary size and the risk falls to at most 5.6×10^{-33}: within this dictionary there is nothing to find, with different contents at the same size nothing left to want. This does not license the practitioner's version: support repair is *told* which pair to insert, and the exponential V_r in the upper bound is a real condition, not a technicality that a lucky small atom set replaces.

The interventions that decompose the gap.

- **Repetition, length 2 to 128.** The balanced two-atom optimum is available under every tested cap $L \geq 2$, so no larger cap improves it: length invariance of the block, not a finite-length test of Theorem 1.
- **Support repair.** Replacing two interior atoms by the target pair at unchanged dictionary size collapses the risk to machine zero everywhere. The deficiency is in *which* atoms the dictionary contains, not how many.
- **Coupling swap.** Exchanging the values on the two extrema increases the certified optimal risk in all 384 confirmation base cells. Multiplicative comparisons are restricted to the 256 cells with $a < 1$ (minimum factor 1.27; mean 2.7 Rademacher and 2.95 Gaussian at

$a = 0.5$); in the 128 cells with $a = 1$ the unswapped optimum is zero, so only the additive increase is meaningful.

- **Query-dependent masses.** Allowing the mass vector to depend on the fresh query — the one operation the method is not permitted — fits exactly, maximum deviation 1.1e-16.

The separation is not an inference from these interventions; it is the certification itself. For every non-negative query-independent mass vector, of any support, the contrast between the responses at $\pm z$ is at most the exact same-pair ceiling, which under the box premise equals $2\tanh(az)$ and is carried by the two extrema; the paired-loss inequality then gives $R \geq \mathbb{E}[(\tanh Z - \tanh(aZ))^2]$, attained by the balanced pair. That bounds the exact minimiser, so no search could have done better, and the exhaustive pair sweep is a check on the implementation rather than evidence. The interventions identify the premises of this finite certificate; they do not validate the asymptotic frontier.

5.3 The first layer of the pretrained model, measured

Measured on Qwen2.5-3B-Instruct (float, not enclosures; Table 4): with $V = 151{,}651$, $d = 2048$ and $\sqrt{2\log V/d} = 0.108$, over 200 random readouts per key–value group $\kappa(w)$ has median 0.125 and 0.117, and on the joint value-by-key-sensitivity reach behind the slope budget of Theorem 1 the token fraction is 0.019; all 32 learned soft rows are certified outside the token value hull by a separating dual direction, with group-wise median certified distances 2.1–2.3 against median token value norms of 3.6–3.8. So, as a median over sampled random readouts, the directional support of the whole token class, at every length, is about 12% of that of a norm-capped soft token optimised separately for each direction. Off-hull rows alone do not place the head *output* outside the token-reachable set (averages can re-enter it); that output-level statement is the certified floor below — the first-layer geometry of prompt waywardness (Khashabi et al., 2022).

For the certified floor, measured on eight fitting frames of one development domain at four query positions and all 16 heads, with the replay validated against the export's σ_h^2 to 3.7×10^{-3} (Table 5): summed over heads and normalised by σ_h^2, the certified floor for prompts of length ≤ 32 is 0.121 against 0.496 for the compiled prompt J; this numerically evaluated lower certificate is 24% of J's measured discrepancy, concentrated on two heads, and it lower-bounds rather than decomposes the representation contribution. The length-dependent mass constraint raises the numerically evaluated lower certificate from 0.112 to 0.121, about 8%; this does not establish which constraints bind at the true token-prompt optimum.

5.4 The confirmation, in brief

A pre-registered compiler study on a protected split (Appendix G) reproduces a teacher–discrete gap of 0.114–0.143 Brier at 32 positions. The compiler with the best measured first-layer fidelity, J, is worse than direct search and PEZ, and its differences from O and M are not resolved: better first-layer fidelity need not rank fitted methods by task loss. Quadrupling the discrete length recovers 11–14% of the O and J compilers' gaps (intervals include zero under intent clustering), and a hand-written $\approx$ 280-token few-shot prompt, not length-matched, reduces the gap of the best 32-token searched arm from 0.114 to 0.066, about 42%. What may be called a vocabulary limit on the real model is the first-layer statement of §3.2–3.3; the output gap is not.

6 Discussion and limitations

The closed form stays inside its model: the scale r has no counterpart in a pretrained transformer, the task direction is not generic for a trained head, token marks are not i.i.d. Gaussian, rotary position embedding breaks the premise that integer mass means copies of one atom, and the model has one head. No verdict about real tokenizers follows in either direction. What transfers is the first-layer geometry, exact and certified, but under Proposition 7's assumptions not by itself a task-loss lower bound. The constructive half of Theorem 1 is a limit without a rate. The compiler study and its transfer rest on two retained model families (Qwen, Falcon); three further families were excluded post hoc after exploratory runs (Appendix F); E2 separately covers Qwen, Phi and Granite. E2 is resource-unmatched, is no representation limit, and has a one-seed hybrid (Appendix F).

AI USE STATEMENT

Large-language-model assistants were used in this project as productivity tools, in the same way as other research software, with every step directed and verified by the authors. They helped develop and refine the theoretical models, the conceptual framing and the hypotheses; assisted in exploring and refining the formulation of mathematical statements and in writing and checking proofs; discussing the experimental design and the analysis; writing and testing the training, evaluation and audit code under the authors' supervision; summarising and interpreting results; and drafting, editing and translating text, which the authors reviewed. They were also used for literature search, figure preparation and reference formatting. Human verification was applied throughout, and no claim of the paper rests on the assistants alone: all proofs were verified by the authors; the controlled block's development cells were re-derived by an independent implementation sharing no code with the module under test; every reported number is regenerated from frozen records by the released scripts; and every reference was checked against its source. The assistants supplied no experimental measurements, benchmark labels or invented observations; the text prompts and native predictions under comparison are, by design, generated by the models studied. The authors take full responsibility for the content of this paper, including any text, claims or code produced with the aid of these tools.

ETHICS STATEMENT

This work uses public benchmarks (Banking77, MASSIVE, TREC, CLINC150) and openly available task models; it involves no human subjects and no personal data. The original E2 analysis was fixed before any test row was scored, with one planned test evaluation per selected artifact; technical re-evaluations and the post hoc common-decoder and budget evaluations are disclosed in the appendices and do not extend the pre-registered error-control family. All pre-registered outcomes are reported.

REPRODUCIBILITY STATEMENT

The anonymous supplementary material opens with a README and a manifest with per-file SHA-256; the frozen records reproduce every table without rerunning any model, which is distinct from refitting (the resolved reflection backend of the text optimizers was not logged). It contains the frozen E2 protocol (full SHA-256), the selection and analysis scripts, the per-row test records behind every table, the post hoc audit scripts and their outputs, the controlled-block code, the budget addendum (hashed before its runs) with its programs, ledgers and test records, the common-decoder scoring code and records, and a manifest with per-file SHA-256; the complete proof of Theorem 1 is Appendix B. The controlled block's 96 development cells carry an independent re-derivation that shares no code with the module under test; its one disagreement was traced to the Monte Carlo error of the prescribed Gaussian measure, whose spread across dictionaries (1.29–1.42% relative) is reported in Table 3 rather than resolved away. The E2 row bootstrap conditions on the fitted artifacts and does not cover retraining or model-family selection uncertainty.

REFERENCES

L. A. Agrawal, S. Tan, D. Soylu, N. Ziems, R. Khare, K. Opsahl-Ong, A. Singhvi, H. Shandilya, M. J. Ryan, M. Jiang, C. Potts, K. Sen, A. G. Dimakis, I. Stoica, D. Klein, M. Zaharia, and O. Khattab. GEPA: Reflective prompt evolution can outperform reinforcement learning. arXiv:2507.19457, 2025.

G. Birkhoff. Extensions of Jentzsch's theorem. *Transactions of the American Mathematical Society*, 85:219–227, 1957.

E. Boursier and C. Boyer. Softmax as linear attention in the large-prompt regime: A measure-based perspective. arXiv:2512.11784, 2025.

I. Casanueva, T. Temčinas, D. Gerz, M. Henderson, and I. Vulić. Efficient intent detection with dual sentence encoders. arXiv:2003.04807, 2020.

V. Castin, P. Ablin, J. A. Carrillo, and G. Peyré. A unified perspective on the dynamics of deep transformers. arXiv:2501.18322, 2025.

S. N. Cohen and E. Fausti. Hyperbolic contractivity and the Hilbert metric on probability measures. arXiv:2309.02413, 2023.

S. Erkan, B. Boll, K. Kersting, B. Deiseroth, and L. Parcalabescu. Soft-prompt tuning for fair and efficient LLM benchmark evaluation. arXiv:2606.12117, 2026.

J. FitzGerald et al. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. arXiv:2204.08582, 2022.

A. Hsu and R. Lai. Training-free universal approximation by prompting random transformers. arXiv:2608.09558, 2026.

A. F. Karr. Extreme points of certain sets of probability measures, with applications. *Mathematics of Operations Research*, 8(1):74–85, 1983.

D. Khashabi, X. Lyu, S. Min, L. Qin, K. Richardson, S. Welleck, H. Hajishirzi, T. Khot, A. Sabharwal, S. Singh, and Y. Choi. Prompt waywardness: The curious case of discretized interpretation of continuous prompts. In *NAACL*, 2022. arXiv:2112.08348.

O. Khattab, A. Singhvi, P. Maheshwari, Z. Zhang, K. Santhanam, et al. DSPy: Compiling declarative language model calls into self-improving pipelines. arXiv:2310.03714, 2023.

B. Lester, R. Al-Rfou, and N. Constant. The power of scale for parameter-efficient prompt tuning. arXiv:2104.08691, 2021.

X. Li and D. Roth. Learning question classifiers. In *COLING*, 2002. doi:10.3115/1072228.1072378.

J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen. What makes good in-context examples for GPT-3? arXiv:2101.06804, 2021.

M. Meyer, M. Michelessa, C. Chaux, and V. Y. F. Tan. Memory limitations of prompt tuning in transformers. arXiv:2509.00421, 2025.

K. Opsahl-Ong, M. J. Ryan, J. Purtell, D. Broman, C. Potts, M. Zaharia, and O. Khattab. Optimizing instructions and demonstrations for multi-stage language model programs. arXiv:2406.11695, 2024.

A. Petrov, P. H. S. Torr, and A. Bibi. When do prompting and prefix-tuning work? A theory of capabilities and limitations. In *International Conference on Learning Representations (ICLR)*, 2024a.

A. Petrov, P. H. S. Torr, and A. Bibi. Prompting a pretrained transformer can be a universal approximator. arXiv:2402.14753, 2024b.

J. T. H. Smith, T. Dakhran, A. Cabrera, S. S. Lee, P. Pak, A. Tadimeti, T. Seyde, M. Labonne, A. Amini, and M. Lechner. In-place tokenizer expansion for pre-trained LLMs. arXiv:2607.15232, 2026.

Y. Wang, J. Chauhan, W. Wang, and C.-J. Hsieh. Universality and limitations of prompt tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.18787.

Y. Wen, N. Jain, J. Kirchenbauer, M. Goldblum, J. Geiping, and T. Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. arXiv:2302.03668, 2023.

A THE TWO-QUERY CEILING AND THE FINITE COMPANION

This appendix gives the no-self two-query response ceiling. Theorem B.3 separately supplies deterministic finite-r lower functionals for the full-self model; those functionals are not an exact finite-r version of Theorem 1's frontier.

An observable-dependent refinement of the Birkhoff–Hopf bound. The largest response gap a prescribed mass vector can produce between two queries is at most the product bound $\mathrm{osc}(c)\tanh(\Delta/4)$, Δ the projective diameter of the image cone, which rests on the Birkhoff–Hopf contraction coefficient of Birkhoff (1957), not a result of ours; in the positive-kernel representation of the two queries as distinct coordinate rays the input Hilbert distance is infinite, so the strengthening of Cohen & Fausti (2023) supplies no further improvement. Our $(\sqrt{\rho}-1)/(\sqrt{\rho}+1)$ is that coefficient at projective diameter $\Delta = \log\rho$, by one line of algebra. Equation (1) instead gives the exact ceiling for the fixed observable c, keeping the value and logit differences at the same pair of indices; it can be strictly smaller than the product bound (in 73 of 80 random dictionaries) without improving the universal contraction coefficient. What it gives is not a smaller universal constant but the exact value for the observable at hand. For a nonempty dictionary of $V \geq 1$ atoms, fixed logits $\ell_i^{\pm}$ and query-independent scalar values c_i, define the no-self mixtures $f_{\pm}(a) = \sum_i a_i e^{\ell_i^{\pm}} c_i / \sum_i a_i e^{\ell_i^{\pm}}$, $\delta_i = \ell_i^+ - \ell_i^-$, $a \geq 0$, $a \neq 0$; equation (1) concerns these mixtures, not the full-self predictor of Appendix B. Then

$$\sup_{a\geq 0,\, a\neq 0} \big(f_+(a) - f_-(a)\big) \;=\; C_D \;:=\; \max_{1\leq i,j\leq V} (c_i - c_j)\,\tanh\tfrac{\delta_i-\delta_j}{4}, \tag{1}$$

the diagonal terms supplying 0, attained on at most two atoms — the extreme-point theorem for one moment constraint (Karr, 1983) followed, with $r_i := e^{\delta_i}$, by the identity $(\sqrt{r_i}-\sqrt{r_j})/(\sqrt{r_i}+\sqrt{r_j}) = \tanh((\delta_i-\delta_j)/4)$. Unlike the product bound, which is invariant under permuting the values across atoms, eq. (1) pairs $(c_i - c_j)$ with $(\delta_i - \delta_j)$ at the same indices, so it sees the coupling swap of Section 5.2: the product bound is unchanged while the exact ceiling drops from $2\tanh a$ to the interior maximum (0.21 to 0.66 at $a \leq 0.5$).

For a two-point query law the premise can be dropped entirely. The set of response pairs (f^-, f^+) attainable by non-negative masses is compact, and every boundary point is attained by a support of size at most two, because f^+ is linear-fractional in the masses and its extrema on a slice $\{f^- = x\}$ sit at vertices. An exact membership test therefore decides any dictionary: inside the region, risk zero is attained by at most three atoms and exhibited by a 3×3 rational system; outside, the exhaustive pair search is the global optimum over all supports. On the block's dictionaries the target is outside ($2\tanh 1 > 2\tanh a$ for $a < 1$), so the Rademacher certification is premise-free; a pair search can fail only in the inside case, which is decidable in advance. For continuous symmetric query laws the bound is tight under the linear-logit, box and aligned-extrema hypotheses of the certificate below.

No-self block certificate. Let Z have a symmetric law and define $G_b(z) = \sum_i b_i e^{k_i z} c_i / \sum_i b_i e^{k_i z}$ for $b \geq 0$, $b \neq 0$. Assume $0 < a \leq 1$, that the dictionary contains $(-a,-1)$ and $(a,1)$, and that every atom (k_i, c_i) lies in $[-a,a]\times[-1,1]$. Then $\delta_i = 2k_i z$, and for every $z \geq 0$ eq. (1) gives $\sup_b (G_b(z) - G_b(-z)) = \max_{i,j}(c_i - c_j)\tanh((k_i - k_j)z/2) = 2\tanh(az)$. Writing $\mathbb{E}(G_b(Z) - \tanh Z)^2 = \frac{1}{2}\mathbb{E}[(G_b(Z) - \tanh Z)^2 + (G_b(-Z) + \tanh Z)^2] \geq \frac{1}{4}\mathbb{E}(G_b(Z) - G_b(-Z) - 2\tanh Z)^2$ and using $\tanh z \geq \tanh(az)$ for $z \geq 0$ gives $\mathbb{E}(G_b(Z) - \tanh Z)^2 \geq \mathbb{E}(\tanh Z - \tanh(aZ))^2$. Equal unit counts on the two extrema give $G_b = \tanh(az)$ for every query, attaining equality at total length two.

B Complete model and proof of Theorem 1

This appendix states the coupled Gaussian attention model in full and proves the $\tanh$ frontier of Theorem 1 (Theorem 1 below; equation and section numbers refer to this appendix), together with the deterministic finite companion (Theorem B.3). The odd-target extension is derived from it in Appendix C. The proof was checked by an independent mathematical audit; its two qualifications of the conditional tail displays (5.1) and (7.3), which restrict them to covariance-cap events, are incorporated below.

Original model, physical predictor, and admissible classes

Coordinates and original Gaussian matrices Let $r \geq 2$ be an integer and $D = r^2$. Fix, before sampling, an ordered injection

$$(j_1, \dots, j_{r+1}) \subseteq \{1, \dots, D\}.$$

Using the standard ambient coordinate vectors, define

$$J_p = [e_{j_1}^{(D)}, \dots, e_{j_r}^{(D)}], \qquad J_n = e_{j_{r+1}}^{(D)}, \qquad J = [J_p, J_n].$$

This gives an explicit coordinate-level definition for every fixed original coordinate list. The theorem below holds for every such fixed ordered list. Its order is retained throughout. In particular, the selector is never rerun after a rotation or diagonalization.

Let

$$W_Q, W_K, W_V \in \mathbb{R}^{D \times D}$$

be independent, with independent entries distributed as $N(0, r^{-2})$. Independently, let the original vocabulary embeddings satisfy

$$E_i \overset{\text{iid}}{\sim} N(0, I_D), \qquad 1 \le i \le V.$$

The physical query embedding is

$$x(Y) = \sqrt{r}\, J_p Y + r J_n, \qquad Y \sim N(0, I_r).$$

The scalar readout is the original row $e_D^{(D)\top} W_V$, and attention scores are divided by r.

Define

$$B = \begin{bmatrix} J^\top W_Q^\top W_K \\ e_D^{(D)\top} W_V \end{bmatrix}, \qquad T = BJ, \qquad \Sigma = BB^\top,$$

and partition each native mark as

$$BE_i = (k_i, s_i, c_i) \in \mathbb{R}^r \times \mathbb{R} \times \mathbb{R}.$$

Partition the rows of T as p, n, c and its columns as p, n. Set

$$A = \operatorname{sym} T_{pp}, \qquad \xi = \sqrt{r}(T_{pn} + T_{np}^\top), \qquad \gamma = r T_{nn},$$

$$v = \sqrt{r}\, T_{cp}^\top, \qquad \nu = r T_{cn}.$$

Then direct expansion of the original query-self score and value gives

$$g(Y) = \frac{x(Y)^\top W_Q^\top W_K x(Y)}{r} = Y^\top A Y + \xi^\top Y + \gamma,$$

$$h_0(Y) = e_D^{(D)\top} W_V x(Y) = \nu + v^\top Y.$$

For a finite nonnegative mass vector $a \in [0, \infty)^V$, the predictor is

$$\boxed{F_a(Y) = \frac{e^{g(Y)} h_0(Y) + \sum_{i=1}^V a_i e^{s_i + k_i^\top Y/\sqrt{r}} c_i}{e^{g(Y)} + \sum_{i=1}^V a_i e^{s_i + k_i^\top Y/\sqrt{r}}}.} \tag{1.1}$$

All mass choices are made before the fresh query. They may depend on the entire frozen backbone and dictionary. No bound on their finite magnitudes is imposed.

An integer mass n_i means n_i repetitions of the same original ID. Its physical prompt length is $\sum_i n_i$. The empty prompt satisfies

$$F_0 = h_0,$$

not $F_0 = 0$.

Norm-legal classes An original ID is norm-legal when

$$\|E_i\| \le 2r.$$

Write $\mathcal{I}_r = \{i : \|E_i\| \le 2r\}$. Define

$$\mathcal{A}_r = \{a \in [0,\infty)^V : a_i = 0 \text{ for } i \notin \mathcal{I}_r\},$$

$$\mathcal{N}_r = \mathcal{A}_r \cap \mathbb{Z}^V_{\ge 0}, \qquad \mathcal{N}_{r,L} = \left\{ n \in \mathcal{N}_r : \sum_i n_i \le L \right\}.$$

Every coordinate is finite, and all three classes contain the empty vector.

For a selected unit direction u, put

$$f(Y) = \tanh(u^\top Y), \qquad R_f(a) = \mathbb{E}_Y (F_a(Y) - f(Y))^2.$$

The three optimal risks are

$$R_r^{*,\mathrm{real}} = \inf_{a \in \mathcal{A}_r} R_f(a), \qquad R_r^{*,\mathrm{count}} = \inf_{n \in \mathcal{N}_r} R_f(n),$$

$$R_r^{*,L} = \inf_{n \in \mathcal{N}_{r,L}} R_f(n).$$

The lower proof will hold over the larger class using every original ID, including any norm-illegal ID. Restricting to norm-legal classes can only increase that lower bound.

THE COUPLED COVARIANCE, SELF, AND ORIGINAL-NORM LAW

Let

$$C_Q = (W_Q J)^\top (W_Q J).$$

Then

$$C_Q \sim r^{-2} \,\mathrm{Wishart}_{r+1}(D, I), \qquad S_B = r^{-2} \,\mathrm{diag}(C_Q, 1).$$

Conditional on $W_Q J$, the columns of B are independent centered Gaussian vectors with covariance S_B. This conditional distribution depends on $W_Q J$ only through C_Q. The columns at the original coordinates selected by J give T, while the remaining columns give an independent Wishart sum. Thus, conditional on C_Q,

$$T = S_B^{1/2} G_J, \qquad W_\perp \sim \mathrm{Wishart}_{r+2}(D - r - 1, S_B),$$

where G_J has independent standard Gaussian entries and $T, W_\perp$ are conditionally independent, and

$$\boxed{\Sigma = TT^\top + W_\perp.} \tag{2.1}$$

This is not independence of Σ and the self-generating matrix T. In fact, $TT^\top$ is part of Σ.

For $r \geq 2$, $D \geq r+2$, and B has full row rank almost surely. Conditional on such a B, let $z_i = BE_i$. Gaussian orthogonal projection gives

$$E_i = B^\top \Sigma^{-1} z_i + E_{i,\perp},$$

where the two terms are orthogonal and conditionally independent. Equivalently,

$$z_i = \Sigma^{1/2}\zeta_i, \qquad \zeta_i \sim N(0, I_{r+2}),$$

and

$$\boxed{\|E_i\|^2 = \|\zeta_i\|^2 + \rho_i, \qquad \rho_i \sim \chi^2_{D-r-2}.} \tag{2.2}$$

Here ρ_i is independent of ζ_i, conditional on B; pairs are independent across original IDs. A zero-degree chi-square denotes the point mass at zero.

In particular,

$$\mathrm{Cov}\big(\|E_i\|^2, z_i z_i^\top \mid \mathscr{B}\big) = 2\Sigma.$$

Consequently, independently redrawing a full χ^2_D norm preserves its marginal distribution but changes the joint object. A later partial disclosure must preserve the already disclosed residual radius, not add a new full radius. These coupling requirements are mathematical premises, not properties authenticated merely by a finite sampling seed or a stored-file hash.

SELECTOR DEFINITION

Selector ingredients The selector uses the matrices

$$S_1 = (I - 2A)^{-1}, \qquad U = [\xi, v, S_1\xi, S_1 v], \qquad \eta = 2^{-20},$$

$$P_{\mathrm{sel}} = I - U(U^\top U + \eta I_4)^{-1} U^\top. \tag{3.1}$$

The matrix P_{sel} is a contraction, not generally an orthogonal projection.

The selector's spectral quantities are $\|A\|_{\mathrm{op}}$ and $\lambda_{\min}(\Sigma)$, with thresholds $1/25$ and $17/20$, and its ordered column tests are

$$(P_{\mathrm{sel}})_{jj} > 2/3, \qquad Q_j := \|AP_{\mathrm{sel}}e_j\|^2 - \|P_{\mathrm{sel}}e_j\|^2/6400 < 0,$$

implemented by a lower endpoint for the first quantity and an upper endpoint for the second.

Definition of the selector The selector is defined as follows.

An allowable enclosure rule is a Borel-measurable, backbone-only rule returning finite closed intervals containing the exact quantities below. Its intervals have the indicated maximum half-widths.

Table 2: Deterministic enclosure tolerances and endpoint tests of the backbone-only selector: mathematical admissibility conditions, not empirical confidence intervals.

Exact quantity enclosed	Maximum half-width	Passing endpoint test
$\|A\|_{\mathrm{op}}$	$1/1000$	upper endpoint $\le 1/25$
$\lambda_{\min}(\Sigma)$	$1/1000$	lower endpoint $\ge 17/20$
$(P_{\mathrm{sel}})_{jj}$	$1/64$	lower endpoint $> 2/3$
Q_j	2^{-20}	upper endpoint < 0

First apply the spectral tests. When they pass, $I - 2A$ is positive definite, so (3.1) is defined. Among columns satisfying both column tests, select the smallest original index j, and set

$$\boxed{u = \frac{P_{\mathrm{sel}} e_j}{\|P_{\mathrm{sel}} e_j\|}.} \tag{3.2}$$

This is the exact normalized column, not the midpoint of an enclosure or a rounded stored direction.

If the spectral tests fail or no column passes, return a mathematical failure flag. For the sole purpose of defining random variables everywhere, assign $u = e_1$ on failure, but retain the failure flag in every theorem statement. This assignment does not convert failure into admission and does not specify a teacher fallback.

The rule may choose its sound intervals in any backbone-dependent manner, including adverse choices within the radius bounds. It may not use the dictionary or the fresh query. Such interval-valued maps exist mathematically—for example, rational-grid enclosures of the exact scalar quantities.

The asymptotic theorem also holds for any other measurable completion that agrees in both exact direction and success status with the literal first-column branch whenever that branch passes. Thus its conclusion does not depend on how equality, later-column ambiguity, or failure is resolved outside that branch.

This mathematical definition assumes the required sound intervals are supplied. It makes no assertion that a particular bounded-precision program produces them, terminates, or meets a resource ceiling.

Why two enclosure radii must be paid If a sound interval contains x and has half-width at most δ, its endpoints need only satisfy

$$\underline{x} \ge x - 2\delta, \qquad \overline{x} \le x + 2\delta. \tag{3.3}$$

The interval need not be centered at x.

Accordingly, on the exact backbone event

$$\|A\|_{\mathrm{op}} < 1/30, \qquad \lambda_{\min}(\Sigma) > .9,$$

$$(P_{\mathrm{sel}})_{11} \ge 3/4, \qquad Q_1 \le -1/12800, \tag{3.4}$$

every allowed enclosure passes the spectral tests and the first-column tests, because

$$1/30 + 2/1000 < 1/25, \qquad .9 - 2/1000 > 17/20,$$

$$3/4 - 2/64 > 2/3, \qquad -1/12800 + 2 \cdot 2^{-20} < 0.$$

The stronger thresholds in (3.4) are proof events, not replacement selector guards.

For completeness, a deterministic existence calculation is also robust to these interval errors. If

$$\|A\|_F^2 \le 11/20, \qquad r \ge 4096,$$

let $E = \{j : (P_{\mathrm{sel}})_{jj} \ge 3/4\}$. Since

$\mathrm{tr}(I - P_{\mathrm{sel}}) \le 4$, at most sixteen columns lie outside E, and

$$\sum_{j\in E} \|P_{\mathrm{sel}} e_j\|^2 \ge r - 20, \qquad \sum_{j\in E} \|AP_{\mathrm{sel}} e_j\|^2 \le 11/20.$$

Some $j \in E$ therefore satisfies

$$\frac{\|AP_{\mathrm{sel}} e_j\|^2}{\|P_{\mathrm{sel}} e_j\|^2} \le \frac{11}{20(r-20)}.$$

Since $\|P_{\mathrm{sel}} e_j\|^2 \ge 9/16$,

$$Q_j \le -\left(\frac{1}{6400} - \frac{11}{20(r-20)}\right)\frac{9}{16} < -2\cdot 2^{-20}.$$

This explains why two enclosure radii must be paid.

MAIN THEOREM AND BACKBONE REGULARITY

Theorem 1 in full: original-law task-risk frontier Suppose

$$V_r \to \infty, \qquad s_r := \frac{\log V_r}{\sqrt{r}} \to s \in [0,\infty), \qquad L_r = \lceil 2r^2 V_r \rceil.$$

Use the original jointly generated model of this appendix (original model and coupled law) and any allowable mathematical selector as defined above.

Let

$$q = \mathbb{E}\tanh^2 Z, \qquad \theta = \mathbb{E}[Z\tanh Z], \qquad Z \sim N(0,1).$$

Then

$$\theta = 1 - q, \qquad q > \theta^2,$$

and, for every $\varepsilon > 0$,

$$\boxed{\begin{aligned}&\Pr\Big(\mathsf{Fail}_r \cup \{\exists i \le V_r : \|E_i\| > 2r\} \\ &\qquad \cup \big\{\mathsf{Success}_r, \max_{j\in\{\mathrm{real},\mathrm{count},L\}} |R_r^{*,j} - \mathcal{R}(s)| > \varepsilon\big\}\Big) \longrightarrow 0,\end{aligned}} \tag{4.1}$$

where

$$\mathcal{R}(s) = q - \theta^2 + (\theta - s)_+^2.$$

There is a common high-probability event proving this conclusion for every sound enclosure choice described in the selector definition. Probability is over the original backbone and vocabulary. Each risk averages a separate fresh Gaussian query.

Gaussian/Wishart regularity The coupled law implies

$$\|\Sigma - I\|_{\mathrm{op}}, \quad \|A\|_{\mathrm{op}}, \quad |\operatorname{tr} A| = O_{\Pr}(r^{-1/2}), \tag{4.2}$$

and

$$\|A\|_F, \quad |\gamma|, \quad |\nu|, \quad \|\xi\|, \quad \|v\| = O_{\Pr}(1). \tag{4.3}$$

Here is a direct justification. For an $m \times n$ standard Gaussian matrix G, the chi-square moment-generating function and a $1/4$-net of the unit sphere give, for $0 < \epsilon < 1$,

$$\Pr\left(\left\|\frac{GG^\top}{n} - I\right\| > \epsilon\right) \le 2\,9^m e^{-n\epsilon^2/32}. \tag{4.4}$$

Indeed, the scalar quadratic forms are chi-square averages; the net changes the required scalar deviation by at most a factor two. With $m = O(r)$, $n = r^2$, (4.4) gives an $O_{\Pr}(r^{-1/2})$ operator deviation.

Apply this first to C_Q, and then to

$$\Sigma = \operatorname{diag}(C_Q^{1/2}, 1)\frac{GG^\top}{r^2}\operatorname{diag}(C_Q^{1/2}, 1).$$

A rectangular Gaussian net bound also gives

$\|G_{(r+1)\times r}\| = O_{\Pr}(\sqrt{r})$, hence

$\|T_{pp}\| = O_{\Pr}(r^{-1/2})$ and the same bound for A.

Conditional independence of the columns of T gives

$$\mathbb{E}(\operatorname{tr} A)^2 = \frac{1}{r}.$$

Direct entrywise second moments give

$$\mathbb{E}\|A\|_F^2 = \frac{r+1}{2r}, \qquad \mathbb{E}\|\xi\|^2 = 2,$$

$$\mathbb{E}\gamma^2 = \mathbb{E}\nu^2 = \mathbb{E}\|v\|^2 = 1.$$

These establish (4.2)–(4.3) without declaring the self coefficients independent of Σ.

First-column success and small longitudinal self The joint law is invariant under rotations of the r prediction coordinates:

$$A \mapsto R^\top A R, \qquad \xi \mapsto R^\top \xi, \qquad v \mapsto R^\top v.$$

Consequently $P_{\mathrm{sel}} \mapsto R^\top P_{\mathrm{sel}} R$. This is a distributional argument, not a rotation of the actual selector's coordinate order.

Almost surely $I - 2A$ is invertible; on its null exceptional set one may define an auxiliary $P_{\mathrm{sel}} = I$ for this argument. The singular values of U give

$$0 \preceq P_{\mathrm{sel}} \preceq I, \qquad \operatorname{tr}(I - P_{\mathrm{sel}}) \le 4.$$

Thus

$$\mathbb{E}[1-(P_{\mathrm{sel}})_{11}] = \frac{1}{r}\mathbb{E}\,\mathrm{tr}(I-P_{\mathrm{sel}}) \le \frac{4}{r}. \tag{4.5}$$

It follows that

$$1-(P_{\mathrm{sel}})_{11} = O_{\Pr}(r^{-1}), \qquad \|P_{\mathrm{sel}}e_1\|^2 \to 1.$$

Together with (4.2),

$$Q_1 \longrightarrow -\frac{1}{6400} \quad \text{in probability.}$$

Therefore the event (3.4) has probability tending to one. On that event, every allowable enclosure rule selects exactly the literal first column.

The dimensionally correct matrix inequality is

$$\boxed{P_{\mathrm{sel}}UU^\top P_{\mathrm{sel}} \preceq \eta(I-P_{\mathrm{sel}}).} \tag{4.6}$$

For a singular value σ of U, this reduces to

$$\frac{\sigma^2\eta^2}{(\sigma^2+\eta)^2} \le \frac{\eta\sigma^2}{\sigma^2+\eta}.$$

Consequently, on the first-column branch,

$$\|U^\top u\|^2 \le \frac{\eta[1-(P_{\mathrm{sel}})_{11}]}{\|P_{\mathrm{sel}}e_1\|^2} = O_{\Pr}(r^{-1}). \tag{4.7}$$

Conditional on the frozen object, decompose

$$Y = X + Zu, \qquad X \sim N(0, I-uu^\top), \quad Z \sim N(0,1),$$

independently. Write

$$h_0 = h + \beta Z, \qquad h = \nu + v^\top X, \quad \beta = u^\top v,$$

$$g = g_X + \ell Z + a_0 Z^2, \qquad \ell = u^\top\xi + 2(Au)^\top X, \quad a_0 = u^\top Au.$$

For every fixed finite p,

$$\|h\|_{L^p(X)} = O_{\Pr}(1), \qquad \|\ell\|_{L^p(X)} = O_{\Pr}(r^{-1/2}),$$

$$|\beta|,\ |a_0| = O_{\Pr}(r^{-1/2}). \tag{4.8}$$

These follow from (4.2), (4.3), (4.7), and Gaussian moments. No independence between u, A, ξ, v, Σ has been introduced.

SHARP LOWER BOUND, UNIFORMLY OVER ALL FINITE MASSES

Throughout the conditional probability and dictionary-expectation arguments, use the full pre-dictionary information

$$\mathscr{B} = \sigma(W_Q, W_K, W_V).$$

The selector's enclosures and direction, and all construction parameters, are measurable with respect to this field before dictionary sampling. Conditioning on a favorable dictionary event would change its law; such events enter through indicators, not as replacement sampling laws. The covariance-event qualifications below are essential.

A simultaneous native radial budget Define

$$T_V = \max_{i \le V} \{c_i^2 + (u^\top k_i)^2\}.$$

Choose a deterministic $\lambda_r \downarrow 1$ such that

$\Pr(\Sigma \preceq \lambda_r I) \to 1$, possible by (4.2).

On C_r, conditional on the full pre-dictionary backbone information $\mathscr{B}$, the pair $(c_i, u^\top k_i)$ is centered Gaussian with covariance at most $\lambda_r I_2$. Coupling it to a standard two-dimensional Gaussian gives

$$1_{C_r} \Pr(T_V > 2\lambda_r(\log V + \log r) \mid \mathscr{B}) \le \frac{1_{C_r}}{r}, \qquad C_r = \{\Sigma \preceq \lambda_r I\}. \tag{5.1}$$

This uses $\Pr(\chi_2^2 > 2t) = e^{-t}$ and a union bound, not key–value independence.

Since $s_r \to s < \infty$,

$$T_V = O_{\Pr}(\sqrt{r}), \qquad \rho_r := \frac{T_V}{2\sqrt{r}} \le s + o_{\Pr}(1). \tag{5.2}$$

Mass-independent derivatives Fix $X = x, Z = z$. On the interval $|t| \le |z|$, define

$$M_z = \max\{\sqrt{T_V},\ |h| + |\beta z|\},$$

$$D_z = \max\left\{\sqrt{T_V/r},\ |\ell| + 2|a_0 z|,\ \sqrt{2|a_0|}\right\}. \tag{5.3}$$

Every value is bounded by M_z. Every logit derivative is bounded by D_z, and the only nonzero second logit derivative, the self term, is bounded in absolute value by D_z^2.

Let S and N be the denominator and numerator of $F_a = N/S$ along this path. Positivity gives

$$|S'| \le D_z S, \quad |S''| \le 2D_z^2 S, \quad |S'''| \le 4D_z^3 S.$$

Because only the self value has derivative β,

$$|N'| \le (M_z D_z + |\beta|)S,$$

$$|N''| \le (2M_z D_z^2 + 2|\beta| D_z)S,$$

$$|N'''| \le (4M_z D_z^3 + 6|\beta| D_z^2)S.$$

Differentiating $N = F_a S$ successively therefore yields

$$|F_a'| \le 2M_z D_z + |\beta|,$$

$$|F_a''| \le 8M_z D_z^2 + 4|\beta| D_z,$$

$$\boxed{|F_a'''| \le 44M_z D_z^3 + 24|\beta| D_z^2.} \tag{5.4}$$

These constants do not depend on the masses, their total, or their support. The argument includes zero masses, the empty prompt, and arbitrarily large finite masses.

Uniform odd-part linearization Let

$$F_{a,o}(X, Z) = \frac{F_a(X + Zu) - F_a(X - Zu)}{2}, \qquad d_a(X) = \partial_Z F_a(X, 0).$$

Taylor's theorem gives

$$F_{a,o}(X, Z) = Z d_a(X) + \varepsilon_a(X, Z),$$

$$|\varepsilon_a| \le \frac{|Z|^3}{6} \sup_{|t| \le |Z|} |F_a'''(X, t)|. \tag{5.5}$$

From (4.8), (5.2), and (5.3),

$$\|M_Z\|_8 = O_{\Pr}(r^{1/4}), \qquad \|D_Z\|_{12} = O_{\Pr}(r^{-1/4}).$$

For example,

$$\||Z|^3 M_Z D_Z^3\|_2 \le \|Z^3\|_8 \|M_Z\|_8 \|D_Z\|_{12}^3 = O_{\Pr}(r^{-1/2}).$$

The term involving βD_Z^2 is smaller. Hence

$$\boxed{\sup_{a \ge 0} \|\varepsilon_a\|_2 = o_{\Pr}(1).} \tag{5.6}$$

The supremum is controlled by an explicit measurable majorant; no uncountable probability-one intersection is being taken.

Clipping the slope At $Z = 0$, differentiate the full attention ratio:

$$d_a = \mathrm{Cov}_\pi(\text{value}, \text{logit slope}) + \beta \pi_{\text{self}}.$$

For any probability weights,

$$|\operatorname{Cov}(C, B)| \le \frac{\operatorname{Var} C + r \operatorname{Var} B}{2\sqrt{r}}.$$

The native pairs contribute $c_i^2 + (u^\top k_i)^2$, and the self pair contributes

$$K(X) = h(X)^2 + r\ell(X)^2.$$

Therefore

$$|d_a(X)| \le |\beta| + \frac{\max\{T_V, K(X)\}}{2\sqrt{r}}. \tag{5.7}$$

Let $\bar{d}_a$ be d_a clipped to $[-\rho_r, \rho_r]$. Since

$$\|K\|_2 \le \|h\|_4^2 + r\|\ell\|_4^2 = O_{\Pr}(1),$$

we obtain

$$\sup_a \|d_a - \bar{d}_a\|_{L^2(X)} \le |\beta| + \frac{\|K\|_2}{2\sqrt{r}} = o_{\Pr}(1). \tag{5.8}$$

The exact scalar optimization Parity and the oddness of the target imply

$$R_f(a) \ge \|F_{a,o} - \tanh Z\|_2^2.$$

For any measurable $d(X) \in [-\rho, \rho]$, independence of X, Z gives exactly

$$\begin{aligned} \mathbb{E}(\tanh Z - Z d(X))^2 &= q - 2\theta\,\mathbb{E}d + \mathbb{E}d^2 \\ &= q - \theta^2 + \mathbb{E}(d-\theta)^2 \\ &\ge q - \theta^2 + (\theta - \rho)_+^2. \end{aligned} \tag{5.9}$$

Combining (5.6), (5.8), and the triangle inequality, there is a mass-independent $\delta_r = o_{\Pr}(1)$ such that

$$\boxed{R_f(a) \ge \left[\sqrt{q - \theta^2 + (\theta - \rho_r)_+^2} - \delta_r\right]_+^2 \quad \text{for every finite } a \ge 0.} \tag{5.10}$$

Equation (5.2) proves the required sharp asymptotic lower.

Gaussian integration by parts gives

$$\theta = \mathbb{E}\,\operatorname{sech}^2 Z = 1 - q.$$

Strict Cauchy–Schwarz gives $q > \theta^2$, since $\tanh Z$ is not almost surely a scalar multiple of Z.

Only the odd part in the selected direction has been linearized. The theorem does not assert global linearity of every predictor. Its even and transverse parts may remain nontrivial, but they cannot improve the lower in (5.9). This is the odd-part argument, not the finite positive-part envelope proved later.

A FINITE CLIPPED-LIKELIHOOD UPPER ESTIMATE

The following lemma supplies the quantitative upper needed for both asymptotic regimes.

Premises and construction Fix a backbone, its unit direction u, and its self coefficients. Conditional on this backbone, suppose the native marks are independent $N(0, \Sigma)$, with

$$\boxed{0 \prec \Sigma \preceq \lambda I, \qquad r \ge 4\lambda, \qquad I - 4A \succ 0.} \tag{6.1}$$

The cap is on the **full** covariance, including the original bias coordinate s, not merely on the (k, c) marginal.

Let $V \geq 1$, $M > 0$, $B_0 \geq 0$, $\alpha, \Delta \geq 0$ and $C \in \mathbb{R}$, all fixed before the dictionary is drawn. Write

$$X_i = (k_i, c_i), \quad \Gamma = \mathrm{Cov}(X_i \mid \mathscr{B}), \quad \mathcal{T} = \Gamma_{kk}/r, \quad j = \Gamma_{kc}/\sqrt{r}.$$

Set

$$m = (\sqrt{r}\,\alpha u, C), \qquad E = m^\top \Gamma^{-1} m, \qquad H_i = m^\top \Gamma^{-1} X_i,$$

$$w_i^{\pm} = e^{\pm H_i - E/2}, \qquad w_i = \frac{w_i^+ + w_i^-}{2}.$$

With

$$T_c = e^{E/2+\Delta},$$

clip the components separately:

$$w_i^{\mathrm{cl}} = \frac{\min(w_i^+, T_c) + \min(w_i^-, T_c)}{2}.$$

Use the actual original bias marks:

$$\boxed{a_i = e^{-s_i} w_i^{\mathrm{cl}}, \qquad n_i = \lfloor M a_i \rfloor.} \tag{6.2}$$

No bias law is substituted, and no key–value independence is assumed.

For $t = (Y/\sqrt{r}, 0)$, write

$$b = t^\top \Gamma t = Y^\top \mathcal{T} Y.$$

Gaussian exponential tilting gives the population denominator and ratio

$$\mathcal{D}(Y) = e^{b/2} \cosh(\alpha Z), \qquad Z = u^\top Y,$$

$$\boxed{p(Y) = \frac{\mathcal{N}(Y)}{\mathcal{D}(Y)} = j^\top Y + C \tanh(\alpha Z).} \tag{6.3}$$

Let

$$A_{B_0} = \{\max_i |c_i| \leq B_0\}, \qquad \mathfrak{h} = 1 + B_0^2 + C^2 + \nu^2 + \|v\|^2,$$

and

$$\chi = \exp\left[-\frac{\Delta^2}{4(E + \alpha^2 + 1)}\right].$$

Define

$$J_{\mathrm{self}} = \mathbb{E}_Y \frac{e^{2g(Y)}(h_0(Y) - p(Y))^2}{\mathcal{D}(Y)^2}.$$

Lemma B.2: finite upper estimate Under (6.1),

$$\boxed{\mathbb{E}_{\mathrm{dict}}\left[1_{A_{B_0}}\|F_n - p\|_2^2\right] \le K_\lambda \mathfrak{h}\left(\frac{T_c}{V} + \chi + M^{-2}\right) + \frac{16 J_{\mathrm{self}}}{M^2 V^2},} \tag{6.4}$$

where

$$K_\lambda = 2^{16}(1+\lambda)^2 e^{8\lambda}.$$

When $m = 0$, use $w = 1$ without clipping and replace

$T_c/V + \chi$ by exactly $1/V$.

The proof follows.

Sampling fluctuations For each fixed query, write the normalized denominator and numerator errors as centered empirical fluctuations minus their clipping biases, $e_D = \zeta_D - \delta_D$ and $e_N = \zeta_N - \delta_N$. Conditional on $\mathscr{B}$,

$$\mathbb{E}_{\mathrm{dict}} e_D^2 = \mathrm{Var}_{\mathrm{dict}}(e_D) + \delta_D^2, \qquad \mathbb{E}_{\mathrm{dict}} e_N^2 = \mathrm{Var}_{\mathrm{dict}}(e_N) + \delta_N^2.$$

These exact identities explain why no factor two is required when combining the sampling and clipping estimates.

The identities

$$(w^{\mathrm{cl}})^2 \le T_c w, \qquad \frac{\cosh(2x)}{\cosh^2 x} \le 2$$

give normalized denominator variance at most

$$\frac{2T_c}{V} e^b, \tag{6.5}$$

and normalized numerator variance at most

$$\frac{2T_c}{V} e^b \left[\lambda + 2C^2 + 8(j^\top Y)^2\right]. \tag{6.6}$$

For (6.6), exponential tilting by $w^{\pm} e^{2t^\top X}$ produces a Gaussian value mean $\pm C + 2j^\top Y$, with variance $\Gamma_{cc} \le \lambda$.

Correlated clipping bias Under the component tilt

$$Q_\pm^Y = N(\pm m + \Gamma t, \Gamma),$$

the scalar $\pm H$ has mean $E \pm \alpha Z$ and variance E. Clipping removes weight only on

$$\{\pm H > E + \Delta\}.$$

For $E > 0$, integrating the sum of these two component tail probabilities over Z gives

$$2\bar{\Phi}\left(\frac{\Delta}{\sqrt{E+\alpha^2}}\right) \le \exp\left[-\frac{\Delta^2}{2(E+\alpha^2)}\right]. \tag{6.7}$$

To see how this controls the correlated numerator, let $a(Y) \in [0,1]$ be the mixture-weighted sum of the component tail probabilities. If δ_D, δ_N denote normalized denominator and numerator clipping biases, then

$$\delta_D^2 \le a(Y),$$

$$\delta_N^2 \le \left[\lambda + 2C^2 + 2(j^\top Y)^2\right] a(Y). \tag{6.8}$$

The second inequality is Cauchy–Schwarz under the tilted mixture; it does not separate value and likelihood.

Using Cauchy–Schwarz once more over Y, (6.7) yields the factor χ in (6.4). When $m = 0$, $w = 1$, so the clipping loss is exactly zero, including at $\Delta = 0$.

Exact rounding and self identity Put

$$f_i = Ma_i - \lfloor Ma_i \rfloor \in [0,1),$$

$$r_D = \frac{\sum_i f_i e^{s_i + t^\top X_i}}{MV\mathcal{D}}, \qquad r_N = \frac{\sum_i f_i e^{s_i + t^\top X_i} c_i}{MV\mathcal{D}}, \qquad u_s = \frac{e^g}{MV\mathcal{D}}.$$

Define the normalized empirical errors

$$e_D = \frac{V^{-1}\sum_i w_i^{\mathrm{cl}} e^{t^\top X_i}}{\mathcal{D}} - 1,$$

$$e_N = \frac{V^{-1}\sum_i w_i^{\mathrm{cl}} e^{t^\top X_i} c_i}{\mathcal{D}} - p.$$

Substitution into the original full-self ratio gives the exact identity

$$\boxed{F_n - p = \frac{e_N - p e_D - r_N + p r_D + u_s(h_0 - p)}{1 + e_D - r_D + u_s}.} \tag{6.9}$$

On A_{B_0}, $|r_N| \le B_0 r_D$. On

$$|e_D| \le 1/4, \qquad r_D \le 1/4,$$

the denominator in (6.9) is at least $1/2$.

For the complement, let

$$W = B_0^2 + h_0^2 + p^2.$$

The convex-value envelope gives $|F_n - p|^2 \le 2W$, while

$$1_{\mathrm{bad}} \le 16(e_D^2 + r_D^2).$$

Combining the good and bad events yields the global estimate

$$\boxed{1_{A_{B_0}} |F_n - p|^2 \le 16 e_N^2 + 48 W e_D^2 + 64 W r_D^2 + 16 u_s^2 (h_0 - p)^2.} \tag{6.10}$$

Thus rare denominator failure is controlled even when self is large or correlated with the other backbone quantities.

Full-covariance control of rounding Jensen's inequality and the joint Gaussian moment-generating function give

$$\mathbb{E}_{\mathrm{dict}} r_D^2 \le M^{-2} \exp\{2\Sigma_{ss} + 4j_s^\top Y + Y^\top \mathcal{T} Y\}, \qquad j_s = \Sigma_{ks}/\sqrt{r}. \tag{6.11}$$

This is the step for which a cap only on Γ is insufficient.

From $r \ge 4\lambda$,

$$(I - 2\mathcal{T})^{-1} \preceq 2I, \qquad \|j\|^2, \|j_s\|^2 \le \lambda/4,$$

$$\mathbb{E} e^{Y^\top \mathcal{T} Y} \le e^{2\lambda}. \tag{6.12}$$

The normalizing integral in (6.11) is at most $e^{8\lambda}$. Its tilted covariance is at most $2I$, and its tilted mean has norm at most $4\sqrt{\lambda}$. Hence

$$\mathbb{E}_Y \left[W e^{2\Sigma_{ss} + 4j_s^\top Y + Y^\top \mathcal{T} Y} \right] \le 16(1+\lambda)^2 e^{8\lambda} \mathfrak{h}. \tag{6.13}$$

For clarity, the constants in the remaining terms can also be checked directly. Under the centered tilt proportional to e^b, the covariance is at most $2I$. Combining (6.5)–(6.6) with the coefficients $16, 48$ in (6.10), the sampling contribution is bounded by

$$\frac{T_c}{V} e^{2\lambda} \left[32\lambda + 896\|j\|^2 + 256C^2 + 96B_0^2 + 96\nu^2 + 192\|v\|^2 \right],$$

which is at most

$$256(1+\lambda) e^{2\lambda} \mathfrak{h}\, T_c/V.$$

For clipping, the Gaussian L^2 bounds

$$\|W\|_2, \quad \|\lambda + 2C^2 + 2(j^\top Y)^2\|_2 \le 2(1+\lambda)\mathfrak{h}$$

and (6.7)–(6.8) give at most

$$128(1+\lambda)\mathfrak{h}\chi.$$

Equation (6.13), multiplied by $64M^{-2}$, gives the rounding contribution

$$1024(1+\lambda)^2 e^{8\lambda} \mathfrak{h}/M^2.$$

Each is covered by the stated K_λ.

The same quadratic self, not an independent substitute Let

$$S_2 = (I - 4A)^{-1}, \qquad J_0 = \frac{\exp\{2\gamma + 2\xi^\top S_2 \xi\}}{\sqrt{\det(I - 4A)}}.$$

Since $\mathcal{D} \ge 1$, Gaussian square completion gives

$$\boxed{J_{\mathrm{self}} \le 2J_0 \left[\left(\nu + 2(v-j)^\top S_2 \xi \right)^2 + (v-j)^\top S_2 (v-j) + C^2 \right].} \tag{6.14}$$

The tilted Gaussian has mean $2S_2\xi$ and covariance S_2.

Equations (6.10)–(6.14) prove Lemma B.2. The condition $I - 4A \succ 0$ is a premise of this exponential-envelope technique. It is not necessary for finiteness of the physical risk, whose convex-value envelope remains available for arbitrary finite A. The full covariance and integrability premises are those in equation (6.1).

ATTAINMENT, SLOW VOCABULARY GROWTH, LEGALITY, AND JOINT CONVERGENCE

Positive finite s Suppose $s > 0$. Choose

$$d_r = (\min\{\theta, s_r\} - 2r^{-1/8})_+,$$

$$\alpha_r = \sqrt{d_r}\, r^{-1/4}, \qquad C_r = \sqrt{d_r}\, r^{1/4}, \qquad \Delta_r = r^{3/8}, \qquad M = r. \tag{7.1}$$

Since $\Gamma = I + O_{\Pr}(r^{-1/2})$,

$$E/2 = d_r\sqrt{r} + O_{\Pr}(1).$$

For all sufficiently large r, $d_r > 0$, and

$$\log V - E/2 - \Delta \geq r^{3/8} + O_{\Pr}(1) \longrightarrow \infty. \tag{7.2}$$

Thus T_c/V decreases exponentially on the relevant scale. Also,

$$\chi \leq e^{-cr^{1/4}}$$

on backbone events of probability tending to one, for some fixed $c > 0$.

Fix any $\lambda > 1$. The premises (6.1) hold with probability tending to one. Take

$$B_0^2 = 4\lambda \log(2V).$$

Then

$$1_{\{\Sigma \preceq \lambda I\}} \Pr(A_{B_0}^c \mid \mathscr{B}) \leq \frac{1_{\{\Sigma \preceq \lambda I\}}}{2V}. \tag{7.3}$$

Here $\mathfrak{h} = O_{\Pr}(\sqrt{r})$.

The self factor is also controlled. On $\|A\|_{\mathrm{op}} \to 0$,

$$\log J_0 = 2\gamma + 2\xi^\top S_2 \xi + 2\,\mathrm{tr}\, A + O(\|A\|_F^2) = O_{\Pr}(1). \tag{7.4}$$

Hence J_0 is tight and (6.14) gives

$J_{\mathrm{self}} = O_{\Pr}(1 + C_r^2) = O_{\Pr}(\sqrt{r})$.

Conditional Markov applied to (6.4), with (7.2)–(7.4), proves

$$\|F_n - p\|_2 \longrightarrow 0 \quad \text{in probability.} \tag{7.5}$$

Moreover,

$$\|j^\top Y\|_2 = \|j\| \to 0,$$

and

$$|C_r \tanh(\alpha_r Z) - d_r Z| \le \frac{C_r \alpha_r^3}{3}|Z|^3 = \frac{d_r^2}{3\sqrt{r}}|Z|^3.$$

Therefore

$$F_n \longrightarrow d_* Z \quad \text{in query-}L^2\text{, in probability,} \qquad d_* = \min\{\theta, s\}.$$

Its task risk converges to

$$q - 2\theta d_* + d_*^2 = \mathcal{R}(s).$$

For clarity, a random conditional upper bound $H_r(\mathscr{B})$ tending to zero in probability is sufficient: on the good backbone event,

$$\Pr(\text{error} > \varepsilon, \text{good backbone}, A_{B_0}) \le \mathbb{E}\min\{1, H_r(\mathscr{B})/\varepsilon\} \longrightarrow 0.$$

No uniform unconditional expectation of the exponential self bound is needed.

Signed tilt The value coordinate C of the tilt may be negative. Every estimate of Lemma B.2 involves C only through C^2 (in (6.6), (6.8), $\mathfrak{h}$ and (6.14)), through the tilted value mean $\pm C + 2j^\top Y$, which enters squared, and through $E = m^\top \Gamma^{-1} m$, which is a parameter of the estimate; the population identity (6.3) is algebraic in C. Hence Lemma B.2 holds for every $C \in \mathbb{R}$. Replacing C_r by $-C_r$ in (7.1) leaves $\|m\|^2 = r\alpha_r^2 + C_r^2 = 2d_r\sqrt{r}$ unchanged, so $E/2 = d_r\sqrt{r} + O_{\Pr}(1)$ as before, since $\Gamma = I + O_{\Pr}(r^{-1/2})$; (7.2)–(7.5) therefore hold verbatim. For either sign of C_r, $|\Sigma_{sX}\Gamma^{-1}m| \le \|\Sigma_{sX}\|\,\|\Gamma^{-1}\|\,\|m\| = O_{\Pr}(r^{-1/4})$, so the length multiplier $e^{\Sigma_{ss}/2}\cosh(\Sigma_{sX}\Gamma^{-1}m)$ in (7.8) still converges to $e^{1/2}$, and the Taylor bound becomes $|C_r \tanh(\alpha_r Z) + d_r Z| \le |C_r|\alpha_r^3|Z|^3/3$. The witness then satisfies $F_n \to -d_* Z$ in query-L^2, in probability, with the same legality and length guarantees. This is the construction used for targets with a negative first Hermite coefficient; the masses remain nonnegative integers on the original dictionary.

$s = 0$, including arbitrarily slow $V \to \infty$ Set

$$m = 0, \qquad w = 1, \qquad M = r, \qquad \boxed{n_i = \lfloor re^{-s_i} \rfloor}, \qquad p(Y) = j^\top Y. \tag{7.6}$$

This is not the empty prompt.

Keep $B_0^2 = 4\lambda \log(2V)$. Equation (7.3) remains valid. When $C = 0$, (7.4) and (6.14) imply tightness of J_{self}.

For a large fixed K, choose backbone events E_K on which (6.1) holds and

$$\nu^2 + \|v\|^2 + J_{\text{self}} \le K.$$

They can be chosen so that

$$\lim_{K\to\infty} \limsup_{r\to\infty} \Pr(E_K^c) = 0.$$

The exact zero-clipping version of (6.4) gives

$$\begin{aligned} \Pr(\|F_n - p\|_2^2 > \varepsilon) \le\ & \Pr(E_K^c) + \frac{1}{2V} \\ & + \frac{1}{\varepsilon}\left\{ K_\lambda[1 + 4\lambda\log(2V) + K](V^{-1} + r^{-2}) + \frac{16K}{r^2 V^2} \right\}. \end{aligned} \tag{7.7}$$

For fixed K, the last line tends to zero because

$$\frac{1+\log V}{V} \to 0, \qquad \frac{\log V}{r^2} \to 0.$$

The first holds for every $V \to \infty$, however slowly. The second follows from $\log V = o(\sqrt{r})$. Letting $K \to \infty$ then proves

$$F_n \to 0 \quad \text{in query-}L^2\text{, in probability,}$$

and

$$R_f(n) \to q = \mathcal{R}(0).$$

No factor $\log r/V$ appears. Tight self quantities must not be replaced by unnecessary growing deterministic envelopes. No uniform unconditional expectation of the self exponential is needed.

Integer length Clipping decreases weights. Joint Gaussian integration, including the original bias, gives

$$\mathbb{E}_{\text{dict}} \sum_i n_i \le MV\, e^{\Sigma_{ss}/2} \cosh(\Sigma_{sX}\Gamma^{-1}m). \tag{7.8}$$

Here $X = (k, c)$.

For (7.1),

$$\|\Sigma_{sX}\| = O_{\Pr}(r^{-1/2}), \qquad \|m\| = O(r^{1/4}),$$

so the multiplier in (7.8) converges to $e^{1/2} < 2$. The same is true when $m = 0$. With $M = r$, on backbone events of probability tending to one,

$$\Pr\left(\sum_i n_i > 2r^2 V \mid \mathscr{B}\right) \le \frac{1}{r}. \tag{7.9}$$

This is a sufficient length bound; no optimality of its polynomial overhead is asserted.

Original-ID norm legality For each original embedding, the chi-square Chernoff bound gives

$$\Pr(\|E_i\| > 2r) \le \exp\left[-\frac{3-\log 4}{2} r^2\right] \le e^{-r^2/2}.$$

Therefore

$$\Pr\left(\max_{i \le V} \|E_i\| > 2r\right) \le V e^{-r^2/2} \to 0. \tag{7.10}$$

This uses original embeddings, not independently supplied norm marginals.

A fully defined legal witness can return the constructed count vector only when every used ID is legal and the length is at most L_r, and return the empty vector otherwise. Equations (7.9)–(7.10) show that this modification occurs with probability tending to zero.

Measurability and the sandwich For fixed finite coefficients,

$$|F_a(y)| \le \max\{|h_0(y)|, \max_i |c_i|\}. \tag{7.11}$$

The right side has finite Gaussian second moment. Pointwise continuity in finite masses and dominated convergence imply continuity of $R_f(a)$. The real-mass infimum equals the infimum over the corresponding dense set of rational mass vectors, including the zero vector. The random legality mask is measurable, so all three infima are measurable. Integer classes are countable; a fixed-length integer class is finite.

On the all-legal event,

$$R_r^{*,\mathrm{real}} \le R_r^{*,\mathrm{count}} \le R_r^{*,L_r}.$$

The sharp uniform lower bound of this appendix applies to all three. The legal length-bounded witness of this section supplies their common upper. The common first-column event supplies selector success independently of allowable interval errors. This proves (4.1).

Scope of the conclusion The theorem covers $0 \le s < \infty$, not $s = \infty$. It establishes representability in task risk, not efficient search or bounded-precision completion.

At $r = V = 4096$, its sufficient length is

$$L_r = 137{,}438{,}953{,}472,$$

not 64.

Nor does it establish an exact fixed-bias64 teacher-fidelity frontier. For the corresponding two-soft-token teacher,

$$H = \frac{e^g h_0 + 2e^{64}\sinh Z}{e^g + 2e^{64}\cosh Z},$$

one has

$$H - \tanh Z = \frac{e^g}{e^g + 2e^{64}\cosh Z}(h_0 - \tanh Z).$$

The available estimate is $e^{-64}O_{\Pr}(1)$, not automatically $o_{\Pr}(1)$. A small fixed allowance does not establish equality of exact limiting risks.

DETERMINISTIC FINITE COMPANION WITH THE LITERAL STORED TARGET

This section does not require Gaussian provenance of the fixed coefficients.

Theorem B.3: geometry and radial lower functionals (finite companion) Let $r \ge 1$, $V \ge 0$ be finite. Fix finite real coefficients

$$A = A^\top, \quad \xi, v \in \mathbb{R}^r, \quad \gamma, \nu \in \mathbb{R},$$

and native marks (k_i, s_i, c_i). Define F_a by (1.1).

Fix a literal target vector $w \ne 0$, and put

$$\boxed{\tau = \|w\|, \qquad e = w/\tau, \qquad P_\perp = I - ee^\top.} \tag{8.1}$$

The target remains $\tanh(w^\top Y)$, not $\tanh(e^\top Y)$.

Let

$$Y = X + Ze, \qquad X \sim N(0, P_\perp), \quad Z \sim N(0,1),$$

independently, and define

$$\beta = e^\top v, \qquad t_i = e^\top k_i,$$

$$H(X) = \nu + v^\top X, \qquad T_0(X) = \sqrt{r}\, e^\top \xi + 2\sqrt{r}\, (Ae)^\top X.$$

Use explicit empty-set conventions:

$$M_N = \max\left(\{0\} \cup \{(c_i - c_j)(t_i - t_j) : 1 \le i, j \le V\}\right),$$

$$T_N = \max\left(\{0\} \cup \{c_i^2 + t_i^2 : 1 \le i \le V\}\right).$$

Set

$$\mathcal{M}(X) = \max\left(\{M_N\} \cup \{(c_i - H(X))(t_i - T_0(X)) : 1 \le i \le V\}\right),$$

$$b_g(X) = |\beta| + \frac{\mathcal{M}(X)}{4\sqrt{r}},$$

$$b_r(X) = |\beta| + \frac{\max\{T_N, H(X)^2 + T_0(X)^2\}}{2\sqrt{r}},$$

and, for $t \ge 0$,

$$J_\tau(t) = \mathbb{E}\big[(\tanh(\tau|Z|) - t|Z|)_+^2\big].$$

Then, simultaneously for every finite query-independent $a \ge 0$,

$$\boxed{R_f(a) := \mathbb{E}(F_a(Y) - \tanh(w^\top Y))^2 \ge \lambda_g := \mathbb{E}_X J_\tau(b_g(X)) \ge \lambda_r := \mathbb{E}_X J_\tau(b_r(X)).} \tag{8.2}$$

Maximal one-sided covariance For every nonempty finite support $\{(C_i, T_i) : i \in I\}$,

$$\boxed{\max_{\pi \in \Delta_I} \mathrm{Cov}_\pi(C, T) = \frac{1}{4} \max_{i,j \in I} (C_i - C_j)(T_i - T_j).} \tag{8.3}$$

Proof. A maximizer exists by compactness. Let S be its positive-weight support. For every tangent vector δ supported on S, with $\sum_i \delta_i = 0$, positive and negative sufficiently small perturbations are feasible. At a maximum,

$$-2\left(\sum_i \delta_i C_i\right)\left(\sum_i \delta_i T_i\right) \le 0.$$

Thus the two linear functionals on the tangent space have nonnegative product everywhere.

If one functional is identically zero, its coordinate is constant on S, and the covariance is zero. Otherwise the functionals are positive scalar multiples: two independent functionals would have joint range $\mathbb{R}^2$, containing $(1, -1)$. Hence

$$T_i = \alpha C_i + b, \qquad i \in S, \quad \alpha > 0.$$

If $m = \min_{i \in S} C_i$, $M = \max_{i \in S} C_i$, then

$$\mathrm{Cov}(C, T) = \alpha \,\mathrm{Var}(C) \le \frac{\alpha (M - m)^2}{4}.$$

The variance inequality follows from

$$\mathrm{Var}(C) \le (\mathbb{E}C - m)(M - \mathbb{E}C) \le (M - m)^2/4.$$

The two support endpoints have pair product $\alpha(M - m)^2$, proving the upper. Equal weights on a maximizing pair attain its pair product divided by four. Taking $i = j$ includes the singleton and zero-maximum cases. □

This is a one-sided maximum, not an absolute-covariance formula. On

$\{(-1, 1), (1, -1)\}$, its right side is zero, while equal weights have covariance -1.

Well-definedness and integrability For each finite y, the denominator of $F_a(y)$ is finite and strictly positive because $e^{g(y)} > 0$. No query-uniform denominator floor is required. The envelope (7.11), omitting the native maximum when $V = 0$, proves finite Gaussian second moments and continuity in finite masses.

In particular, Theorem B.3 requires no condition $I - 4A \succ 0$.

Freeze the even self logit at its endpoint value Fix $X = x$ and $z \ge 0$. Write

$$g_\perp(x) = x^\top A x + \xi^\top x + \gamma, \qquad a_e = e^\top A e.$$

For $u \in [-z, z]$, define

$$\rho_0(u) = \exp\left(g_\perp(x) + a_e z^2 + \frac{T_0(x) u}{\sqrt{r}} \right),$$

$$\rho_i(u) = a_i \exp\left(s_i + \frac{k_i^\top x}{\sqrt{r}} + \frac{t_i u}{\sqrt{r}} \right).$$

Let $G_z(u)$ be the normalized weighted average with these weights and values

$$H(x), c_1, \ldots, c_V.$$

The quadratic term is frozen at $a_e z^2$; the self value is temporarily frozen at $H(x)$.

Differentiation gives

$$G_z'(u) = \frac{1}{\sqrt{r}} \,\mathrm{Cov}_{\pi_u}(C, T) \le \frac{\mathcal{M}(x)}{4\sqrt{r}}, \tag{8.4}$$

using the augmented support

$$\{(H(x), T_0(x)), (c_1, t_1), \ldots, (c_V, t_V)\}.$$

Zero-mass entries may be removed. Including unused entries only relaxes the maximum.

At $u = \pm z$, the frozen even self logit equals the actual even self logit. Restoring the affine self value gives exactly

$$F_a(x \pm ze) = G_z(\pm z) \pm \beta z\, \pi_0(\pm z).$$

Therefore the odd part satisfies

$$\begin{aligned} O_x(z) &:= \frac{F_a(x+ze) - F_a(x-ze)}{2} \\ &= \frac{1}{2}\int_{-z}^{z} G_z'(u)\, du + \frac{\beta z}{2}\{\pi_0(z) + \pi_0(-z)\} \\ &\le z\, b_g(x). \end{aligned} \tag{8.5}$$

The quadratic self term has not been differentiated and then dropped; the constructed path preserves both endpoints exactly.

One-sided distance and parity Let

$$E_x(z) = \frac{F_a(x+ze) + F_a(x-ze)}{2}, \qquad t_z = \tanh(\tau z).$$

The paired squared loss is

$$\frac{1}{2}\sum_{\sigma=\pm 1} \big(F_a(x+\sigma ze) - \sigma t_z\big)^2 = E_x(z)^2 + (O_x(z) - t_z)^2.$$

By (8.5),

$$(O_x(z) - t_z)^2 \ge \operatorname{dist}(t_z, (-\infty, z b_g(x)])^2 = (t_z - z b_g(x))_+^2.$$

No lower bound on $O_x(z)$ is needed. Integration over $|Z|$ and X proves $R_f(a) \ge \lambda_g$.

Radial domination Every augmented pair lies in the disk with squared radius

$$Q(x) = \max\{T_N, H(x)^2 + T_0(x)^2\}.$$

For two such points,

$$\Delta C\, \Delta T \le \frac{(\Delta C)^2 + (\Delta T)^2}{2} \le 2Q(x).$$

Thus $b_g(x) \le b_r(x)$. Since J_τ is nonincreasing,

$\lambda_g \ge \lambda_r$.

Every step used quantities independent of the particular mass vector. This proves the simultaneous deterministic quantifier, including empty prompts and $V = 0$. □

Why query independence is substantive Take $r = 1$, $g = h_0 = 0$, two values $c_\pm = \pm 2$, and zero native logits. Query-dependent masses

$$a_\pm(Y) = \frac{1 \pm \tanh Y}{2}$$

produce $F(Y) = \tanh Y$ exactly. Yet the fixed-mass geometry has $b_g = 0$, whose lower is $q > 0$. Such mass functions are therefore excluded.

Correlated, possibly degenerate integration law The pair used to integrate (8.2) is

$$\begin{pmatrix} H \\ T_0 \end{pmatrix} \sim N\left(\begin{pmatrix} \nu \\ \sqrt{r}\, e^\top \xi \end{pmatrix}, \begin{pmatrix} \|P_\perp v\|^2 & 2\sqrt{r}(P_\perp v)^\top (P_\perp Ae) \\ 2\sqrt{r}(P_\perp v)^\top (P_\perp Ae) & 4r\|P_\perp Ae\|^2 \end{pmatrix} \right). \tag{8.6}$$

This covariance is a Gram matrix. It may have rank two, one, or zero. The theorem integrates against the resulting Gaussian measure, not necessarily a two-dimensional density. Its cross-covariance cannot be discarded.

A legal empty upper For any designated admissible-ID subset and any $L \geq 0$, let $\mathcal{C}_L$ be its integer count vectors of total length at most L. Since $0 \in \mathcal{C}_L$,

$$\boxed{\lambda_g \leq \inf_{n \in \mathcal{C}_L} R_f(n) \leq R_f(0) = \nu^2 + \|v\|^2 + q_\tau - 2\beta\theta_\tau,} \tag{8.7}$$

where

$$q_\tau = \mathbb{E}\tanh^2(\tau Z), \qquad \theta_\tau = \mathbb{E}[Z\tanh(\tau Z)] = \tau(1 - q_\tau).$$

The identity follows by expansion and Gaussian integration by parts.

The zero function's risk q_τ is not automatically a legal upper. The empty prompt's output is the actual affine self h_0.

Relation to the sharp asymptotic theorem Covariance sharpness in (8.3) does not imply sharpness of the integrated risk bound. The optimizing simplex weights may vary with x, z and need not be realized by one fixed prompt.

Likewise, the finite theorem cannot replace the sharp uniform lower bound. For example,

$$J_1(1) = 0,$$

whereas the asymptotic residual $q - \theta^2$ is strictly positive. The sharper lower depends on uniform odd-part linearization and a fixed linear slope budget, not merely pointwise distance to an output interval.

C Proof of Theorem 1 relative to the $\tanh$ case

The $\tanh$ case is Theorem 1 of Appendix B, where the model is defined and the result proved in full; we use three of its statements, none of which involves the target. Fix the frozen object (backbone, selector, dictionary) and write $Y = X + Ze$ with $e = u$, $Z = u^\top Y \sim N(0,1)$ independent of X. For a finite query-independent mass vector a let $F_a(Y)$ be the predictor and $F_a = F_{a,\mathrm{e}} + F_{a,\mathrm{o}}$ its even and odd parts in Z at fixed X.

- (F1) *Uniform odd-part linearisation.* There are ρ_r with $\rho_r \leq s + o_{\Pr}(1)$ such that, on events of probability tending to one, for every finite query-independent a, $F_{a,\mathrm{o}}(X,Z) = Z\, d_a(X) + \varepsilon_a(X,Z)$ with a measurable d_a satisfying $|d_a| \leq \rho_r$ after a clipping that costs $o_{\Pr}(1)$ in L^2, and $\sup_a \|\varepsilon_a\|_{L^2} = o_{\Pr}(1)$. (Taylor remainder with the derivative constants $44/24$; the moment orders $\|M_Z\|_8 = O_{\Pr}(r^{1/4})$, $\|D_Z\|_{12} = O_{\Pr}(r^{-1/4})$.)
- (F2) *Even-part rigidity.* With the same constants, $\sup_a \|F_{a,\mathrm{e}}(X,Z) - F_a(X,0)\|_{L^2} = O_{\Pr}(r^{-1/4})$ (second-order Taylor term $\frac{1}{2}Z^2F''$, $\|Z^2F''\|_2 \leq \|Z^2\|_4(8\|M_Z D_Z^2\|_4 + 4|\beta|\,\|D_Z\|_4)$).
- (F3) *Witness.* For a prescribed $d \in [-s, s]$, use $d_r = (\min\{|d|, s_r\} - 2r^{-1/8})_+$ and the signed tilt $C_r = \mathrm{sign}(d)\sqrt{d_r}\, r^{1/4}$. For $d \neq 0$, the construction and estimates of Appendix B (signed form, paragraph "Signed tilt") give a legal integer witness n of length at most $\lceil 2r^2 V_r \rceil$ satisfying $F_n \to dZ$ in query-L^2, in probability. For $d = 0$, use the unclipped

zero-tilt construction (7.6); its estimate (7.7) remains valid for every finite s, since $(1 + \log V_r)/V_r \to 0$ and $\log V_r/r^2 \to 0$.

Lower bound. Let f be odd. For any a, since $F_{a,\mathrm{e}}$ is even and $F_{a,\mathrm{o}} - f$ is odd in Z, and Z is symmetric and independent of X, $\mathbb{E}(F_a - f)^2 = \mathbb{E}F_{a,\mathrm{e}}^2 + \mathbb{E}(F_{a,\mathrm{o}} - f)^2 \geq \mathbb{E}(F_{a,\mathrm{o}} - f)^2$. By (F1) and the triangle inequality in L^2, $\mathbb{E}(F_{a,\mathrm{o}} - f)^2 \geq \big(\|Zd_a(X) - f(Z)\|_2 - \|\varepsilon_a\|_2\big)_+^2$. For any measurable $d(X)$ with $|d| \leq \rho_r$, using $X \perp Z$, $\mathbb{E}Z^2 = 1$, and $\mathbb{E}[Zf(Z)] = \theta_f$,

$$\mathbb{E}\big(Zd(X) - f(Z)\big)^2 = \mathbb{E}d^2 - 2\theta_f\,\mathbb{E}d + \|f\|^2 = \|f\|^2 - \theta_f^2 + \mathbb{E}(d - \theta_f)^2 \geq \|f\|^2 - \theta_f^2 + (|\theta_f| - \rho_r)_+^2.$$

Hence, uniformly in a, $\mathbb{E}(F_a - f)^2 \geq \mathcal{R}_f(s) - o_{\Pr}(1)$, for all three classes at once.

Upper bound. Apply (F3) with $d = \operatorname{sign}(\theta_f)\min\{|\theta_f|, s\}$. Then $\|F_n - f\|_2^2 \to \|dZ - f(Z)\|_2^2 = \|f\|^2 - 2\theta_f d + d^2 = \mathcal{R}_f(s)$, using $\|F_n - f\|_2 \leq \|F_n - dZ\|_2 + \|dZ - f\|_2$ (for $\theta_f = 0$ this is the zero-tilt witness). Legality, length and measurability are those of the $\tanh$ case. □

Corollary 3, even part. For a general $f = f_0 + f_\mathrm{o} + f_\mathrm{e}$ with $\mathbb{E}f_\mathrm{e} = 0$, the parity split holds with f in place of its odd part: $\mathbb{E}(F_a - f)^2 = \mathbb{E}(F_{a,\mathrm{e}} - f_0 - f_\mathrm{e})^2 + \mathbb{E}(F_{a,\mathrm{o}} - f_\mathrm{o})^2$. By (F2), $F_{a,\mathrm{e}} = g_a(X) + o_{\Pr}(1)$ with $g_a(X) = F_a(X, 0)$, and for any measurable g, $\mathbb{E}(g(X) - f_0 - f_\mathrm{e}(Z))^2 = \mathbb{E}(g - f_0)^2 + \|f_\mathrm{e}\|^2 \geq \|f_\mathrm{e}\|^2$ by independence and $\mathbb{E}f_\mathrm{e}(Z) = 0$. The odd part is bounded as above; so for every $\varepsilon > 0$, with probability tending to one, all three optima are at least $\|f_\mathrm{e}\|^2 + \mathcal{R}_{f_\mathrm{o}}(s) - \varepsilon$. Whether the constant f_0 is attainable is not decided here. □

Hermite reading. For odd f with probabilists' Hermite coefficients $f_k = \mathbb{E}[f(Z)He_k(Z)]/k!$, $\theta_f = f_1$ and $\|f\|^2 - \theta_f^2 = \sum_{k \geq 3\ \mathrm{odd}} k!\, f_k^2$ (Parseval). Numerically (40-digit quadrature, float): for $\tanh$, $\tanh(2\cdot)$, $\tanh(5\cdot)$ the thresholds $|\theta_f|$ are 0.6057, 0.7295, 0.7852 and the linear floors 0.0274, 0.1031, 0.2264.

D Proofs for Sections 3.2–3.4

Lemma (first-layer identity). With $D_P = \sum_{p \in P} e^{s_p}$ and $b_P = \sum_p e^{s_p} v_{t_p}/D_P$, $o_P(i) = (D_{\mathrm{live}}\, o_{\mathrm{live}} + D_P\, b_P)/(D_{\mathrm{live}} + D_P) = o_{\mathrm{live}} + \lambda_P(b_P - o_{\mathrm{live}})$ with $\lambda_P = D_P/(D_P + D_{\mathrm{live}})$; this is the convex combination of Petrov et al. (2024a) with the prefix keys and values inserted, and it is exact because at the first layer every key and value depends only on its own token and position.

Theorem 5. (i) b_P is a convex combination of the v_{t_p}; $\lambda \mapsto D/(D + D_{\mathrm{live}})$ is increasing and $D_P \leq Le^{s^*}$; the length-one prompt (t) has $b = v_t$, so every vertex of $\operatorname{conv}\{v_t\}$ is reached at length one. (ii) $v(e) = b_V + \varphi(\|e\|)\, W_V(g \odot \hat{e})$, and, with $\hat{0} = 0$ and φ continuous and increasing from $\varphi(0) = 0$, the set $\{\varphi(\|e\|)\hat{e} : \|e\| \leq \mathrm{cap}\}$ is the closed ball $\varphi(\mathrm{cap})B$ (including $\mathrm{cap} = 0$); a token value is the point of $\mathcal{E}$ at radius $\varphi(\|E_t\|)$ in direction $\hat{E}_t$. (iii) $\langle w, v_t - b_V\rangle = \langle W_V^\top w, n(E_t)\rangle = \varphi(\|E_t\|)\langle g \odot W_V^\top w, \hat{E}_t\rangle$; for a soft row the same with $\varphi(\|e\|) \leq \varphi(\mathrm{cap})$ and Cauchy–Schwarz, with equality at $\hat{e} \propto g \odot W_V^\top w$ and $\|e\| = \mathrm{cap}$; a convex combination cannot exceed the maximum over its vertices. □

Proposition 6. Theorem 5(i) places o_P in the two sets; the distance inequalities follow, and for any x in a set K and any w with $\langle w, k\rangle \leq \beta$ on K, $\|o_T - x\| \geq \langle w, o_T - x\rangle/\|w\| \geq (\langle w, o_T\rangle - \beta)/\|w\|$. In the computation the distances are bracketed by Frank–Wolfe on $\frac{1}{2}\|x - o_T\|^2$: for an explicit convex combination $x_k \in K$, with $d_k = \|o_T - x_k\|$ and $\mathrm{gap}_k = \max_{x \in K}\langle o_T - x_k, x - x_k\rangle$, if $d_k > 0$ then $(d_k - \mathrm{gap}_k/d_k)_+ \leq \operatorname{dist}(o_T, K) \leq d_k$ (the lower bound is the direction $w = o_T - x_k$ with $\beta = \langle w, x_k\rangle + \mathrm{gap}_k$), and if $d_k = 0$ then $\operatorname{dist}(o_T, K) = 0$. □

Proposition 7. Every network in $\mathcal{N}$ computes the same first-layer quantities, so any bound derived from them alone holds for every member of $\mathcal{N}$. Let $v_1, \ldots, v_m$ ($m \leq |\mathcal{D}|$) be the distinct values of h_x; by (i) each carries one label y_j. If $m = 1$, the constant logits $\ell_1 := \tau e_{y_1}$ are implemented by one hidden unit with zero output weight and output bias ℓ_1; suppose henceforth $m \geq 2$. For all u outside a finite union of hyperplanes the projections $z_j = \langle u, v_j\rangle$ are distinct; order them $z_1 < \cdots < z_m$. The piecewise-linear map $g(z) = \ell_1 + \sum_{j=1}^{m-1} c_j \max(0, z - z_j)$, with $c_1 = s_1$, $c_j = s_j - s_{j-1}$ and $s_j = (\ell_{j+1} - \ell_j)/(z_{j+1} - z_j)$, is one hidden ReLU layer of width $m - 1$ and satisfies $g(z_j) = \ell_j := \tau e_{y_j}$. With logits τe_y the softmax gives the true class $e^\tau/(e^\tau + K - 1)$ and

every other class $1/(e^\tau + K - 1)$, so $B = \frac{1}{2}K(K-1)/(e^\tau + K - 1)^2 \le \frac{1}{2}K(K-1)e^{-2\tau}$ on every example, and $\tau \to \infty$ drives the empirical loss to zero; a strictly positive lower bound valid for all members would contradict this. For $K \ge 2$ the loss is approached, not attained, with finite logits, and the weights grow as $\tau / \min_j(z_{j+1} - z_j)$. Conversely, a readout of h_x alone outputs one vector p on a pair with $h_x = h_{x'}$ and $y_x \ne y_{x'}$, and $\|p - e_y\|^2 + \|p - e_{y'}\|^2 \ge \frac{1}{2}\|e_y - e_{y'}\|^2 = 1$, so the pair contributes at least $\frac{1}{2}$ to the summed loss; c disjoint pairs give $c/(2|\mathcal{D}|)$. Both parts are checked numerically in the released code. □

E MEASUREMENTS

Table 3: Controlled block, 384 confirmation cells over 64 dictionaries. The realised risk of the balanced witness, mean and population standard deviation across dictionaries. The Rademacher prescribed measure is the exact two-point measure, so its realised risk is identical in every dictionary; the Gaussian prescribed measure is 2048 magnitudes with both reflections, so its realised risk carries the sampling error shown. The deterministic closed-form constants (by quadrature) are reproduced to 2e-9; the Gaussian means shown are Monte Carlo values over the prescribed measure.

law	a	mean risk	s.d.	relative s.d.	range
Rademacher	0.25	0.26695357	0	0.00%	—
Rademacher	0.50	0.08968647	0	0.00%	—
Gaussian	0.25	0.16103309	2.28e-3	1.42%	[0.15598, 0.16606]
Gaussian	0.50	0.05031894	6.49e-4	1.29%	[0.04861, 0.05170]

Searched compilers and the teacher's rows. The searched compilers barely beat an empty prefix (0.228–0.231 against 0.236), and the nearest-token projection of every teacher checkpoint is the common initialisation: the soft rows never left their initial tokens' Voronoi cells while their first-layer values left the token hull.

All entries are float32/float64 computations; certificates are exact inequalities evaluated in floating point, not interval enclosures. Weights: `Qwen/Qwen2.5-3B-Instruct` at the pinned revision, first-layer tensors and token embeddings only; development data only.

Table 4: First-layer reach of the token class relative to the soft class on Qwen2.5-3B-Instruct (§3.2); per key–value group where two numbers are given. κ: 200 random unit readouts, five principal directions, and the 32 teacher-row directions. Slope-budget and two-query ceilings: 24 random content-token query pairs × 16 heads × 2 readouts at RoPE offset -128 for both classes. Hull distances: Frank–Wolfe, 300 iterations, lower and upper certificates evaluated in floating point. These summaries describe the frozen backbone and sampled directions, not repeated training runs.

allowed vocabulary V; isotropic reference $\sqrt{2\log V/d}$	151,651; 0.108
$\kappa(w)$, random readouts: median (range)	0.125 (0.099–0.179); 0.117 (0.087–0.187)
$\kappa(w)$ on the five leading value principal directions	0.115, 0.107, 0.161, 0.150, 0.146; 0.516, 0.099, 0.142, 0.101, 0.096
$\kappa(w)$ in the 32 teacher-row directions, median	0.213; 0.214
teacher rows' own reach fraction, median (range)	0.340 (0.222–0.381); 0.311 (0.245–0.429)
rows exceeding every token along their own direction	32/32; 32/32
rows certified outside the token value hull	32/32; 32/32
$\mathrm{dist}(v_p, \mathrm{conv}\{v_t\})$: lower / upper certificate, median	2.336 / 2.342; 2.133 / 2.136
token value norm from b_V: median, max	3.81, 6.93; 3.64, 6.81
slope-budget reach $S_{\mathrm{tok}}/S_{\mathrm{soft}}$: median (5–95%)	0.019 (0.011–0.042)
two-query ceiling $C_D^{\mathrm{tok}}/C_D^{\mathrm{soft}}$: median (5–95%)	0.093 (0.057–0.227)

Table 5: Certified first-layer floor (Proposition 6) on eight fitting frames of the development domain *kitchen_and_dining*, four query positions each (first, two interior, final context), all 16 heads; each entry is a mean of squared discrepancies over the 32 problems of the head, divided by the export's σ_h^2. Found: the selected D0 prompts of direct search, J, O, M, replayed on the same frames (in-sample for those arms). Replay check: per-head σ_h^2 reproduced to 3.7×10^{-3} maximum relative error over all 120 fitting frames. Zero violations of found $\geq$ floor in 2,048 pairs; Frank–Wolfe certificate width median 0.051 in head-output units. Last column: the median over the head's 32 problems of the attention-mass bound $\lambda_{\max}(32)$ of Theorem 5(i), a mass, not a discrepancy. These are in-sample geometric measurements on fitting frames, not held-out Brier losses.

head	floor, any L	floor, $L \leq 32$	direct	J	O	M	$\lambda_{\max}(32)$
0	0.0001	0.0002	0.0116	0.0113	0.0115	0.0114	0.80
1	0.0041	0.0054	0.0311	0.0275	0.0281	0.0280	0.70
2	0.0716	0.0769	0.1301	0.1304	0.1301	0.1301	0.50
3	0.0357	0.0373	0.2088	0.2023	0.2038	0.2033	0.94
4	0.0004	0.0006	0.0085	0.0083	0.0084	0.0084	0.74
5	0	0	0.0017	0.0016	0.0017	0.0017	0.80
6	0	0	0.0815	0.0780	0.0798	0.0792	0.89
7	0.0001	0.0001	0.1833	0.0300	0.0305	0.0304	0.93
8–13	0	0	≤ 0.0045	≤ 0.0003	≤ 0.0005	≤ 0.0004	0.01–0.61
14	0	0	1.5739	0.0023	0.0086	0.0110	0.98
15	0	0	0.0388	0.0035	0.0135	0.0154	0.98
sum	0.112	0.121	2.276	0.496	0.517	0.520	

Finite-r status of the constructive half of Theorem 1 (mark-level simulation, float). With the proof's slack $d_r = (\min\{\theta, s\} - 2r^{-1/8})_+$ the witness has $d_r = 0$ for all $r \leq 1.4 \times 10^4$ at $s \geq \theta$. Without slack, for idealized independent standard Gaussian marks, the population-moment effective-sample-size proxy of the unclipped two-component weights decays like $e^{(s-2d)\sqrt{r}}$ up to constant factors; this proxy is distinct from the empirical ESS, which always lies in $[1, V]$ and need not follow it near weight collapse. The unclipped witness is constructive for $d < s/2$ and collapses near $d = s$ (risk of the order of the value scale squared). At $r = 1024$, $s = 0.3$, the tilt $d = s/4$ gives risk 0.322 with a response whose odd part is 92% linear in energy (nonlinear odd-energy fraction 0.080; slope 0.066 against the prescribed 0.075), between the zero-slope 0.381 and the asymptotic $\mathcal{R}(0.3) = 0.121$; multinomial rounding to K token positions, with repeated atoms permitted, tracks the real-mass witness once K exceeds a few $\sqrt{r}$ (excess 0.14, 0.06, 0.02 at $K = 2\sqrt{r}, 8\sqrt{r}, 32\sqrt{r}$).

F LIMITATIONS IN FULL

The closed form stays inside its model. The model's scale r has no counterpart in a pretrained transformer, and the obstruction is not a missing identification but four independent failures of hypothesis, none of them repairable by choosing r differently: the task direction is not generic for a trained head; token marks are neither i.i.d. Gaussian nor embedded in a $V \to \infty$ ensemble; rotary position embedding makes a repeated token's key position-dependent, which breaks the premise that integer mass means copies of one atom; and the model has one head, a scalar readout and affine self. Even the width identity $D = r^2$ has no analogue, since $\sqrt{d_{\text{model}}}$, $\sqrt{d_{\text{head}}}$, d_{head} and d_{model} are four different numbers in the model we use.

Consequently, **no verdict about real tokenizers follows in either direction.** Under some readings of r all three tokenizers we considered sit above the threshold; under another they sit below it. The identification is undefined, so the sign of the answer is undefined with it.

What does transfer is the first-layer geometry of §3.2–3.3: exact, certified, and, under Proposition 7's label-consistency and readout assumptions, not by itself a task-loss lower bound; this does not exclude task-specific bounds using additional properties of the frozen downstream network. The Gaussian Monte Carlo evaluations of the block have across-dictionary relative standard deviations of 1.29–1.42%; more independent dictionary replicates reduce the standard error of their mean, not the per-replicate variation or the accuracy of the deterministic quadrature constants. The Rademacher

Table 6: Protected confirmation (Appendix G): equal-intent Brier / accuracy over 4,500 records, ten domains, three seeds; family-S contrasts with nominal simultaneous 95% intervals (Bonferroni over eleven); primary contrasts Bonferroni over seven. Controls are descriptive. Two deviations of the fitting from its protocol: the compilers were fitted against the teacher's third checkpoint although the validation-selected teacher differs in 29 of 60 groups, and J's KL weight was selected among J's own candidates.

arm or contrast	Brier / accuracy	interval or note
C, 32 positions	0.088 / 0.894	seeds 0.090, 0.086, 0.087
C, 128 positions	0.122 / 0.832	seeds 0.082, 0.178, 0.106
direct 32; PEZ 32	0.202 / 0.763; 0.207 / 0.770	
M 32; O 32; J 32	0.228 / 0.737; 0.228 / 0.737; 0.231 / 0.735	
O 128; J 128	0.213 / 0.758; 0.211 / 0.765	
few-shot, one utterance per label (≈ 280 tokens)	0.154 / 0.830	not length-matched
empty prefix	0.236 / 0.741	
common initialisation, 32 / 128	0.233 / 0.235	Brier at 32 / 128
nearest-token projection of the teacher, 32 / 128	0.233 / 0.235	Brier at 32 / 128; equals the initialisation
G at 32: direct, PEZ, O, M, J	0.114, 0.119, 0.140, 0.140, 0.143	lower endpoints 0.105–0.134
G at 128: O, J	0.091, 0.089	$[0.083, 0.099]$, $[0.081, 0.097]$
D_C, D_O, D_J (128 minus 32)	$+0.034$, -0.015, -0.020	$[0.029, 0.039]$, $[-0.021, -0.008]$, $[-0.027, -0.014]$
Δ_O^{128} (J minus O at 128)	-0.002	$[-0.008, 0.004]$
Δ_{PEZ}, Δ_{direct} (J minus baseline)	$+0.025$, $+0.029$	$[0.017, 0.032]$, $[0.023, 0.036]$
Δ_O, Δ_M	$+0.003$, $+0.004$	$[-0.003, 0.010]$, $[-0.002, 0.009]$
accuracy J minus direct	-0.028	$[-0.036, -0.020]$; not non-inferior at 0.01
raw first-layer MSE/σ_h^2: J, O, M	0.048, 0.060, 0.049	bounded endpoints 2.7×10^{-7}, 5.1×10^{-7}

Table 7: Matched-length natural text on development validation rows (reading pre-stated before scoring). Natural text: the few-shot construction truncated to exactly 32 and 128 tokens, and untruncated; truncation changes content as well as length. Searched: mean over three seeds of each arm's selected-checkpoint validation Brier. Lower is better; these are development rows, not protected-test results.

domain	nat. 32	nat. 128	nat. full (≈ 280)	search 32	search 128
kitchen_and_dining	0.519	0.402	0.148	0.238 (direct)	0.255 (J)
banking	0.273	0.149	0.158	0.218 (M)	0.229 (O)

evaluations have no such variation, because their prescribed measure is exact. The constructive half of Theorem 1 is a limit statement without a rate: with the proof's constants the attaining witness is the zero-slope prompt for every $r \lesssim 10^4$, and a mark-level simulation at $r = 1024$ reaches one quarter of the reach s constructively (Appendix E); the lower bound, which carries Proposition 4, has no such gap. The pretrained-model evidence at the output is a comparison of finite-budget witnesses on ten domains and two development domains, with the scope of Appendix G.

Transfer, and the two ways it stopped. The gap reproduces on Falcon3-3B-Instruct on the two development domains (teacher–best-discrete Brier gap 0.124 $[0.046, 0.223]$ and 0.041 $[0.002, 0.095]$; two seeds, paired cluster bootstrap over intents, $50{,}000$ draws) and does not currently extend fur-

ther, for two different reasons. A mainstream instruction-tuned family fails a *precondition*: its chat template does not preserve an exact token boundary at the learned positions, so our protocol cannot instantiate the compiler comparison there; this is a protocol-compatibility limit, not a statement about its token-prompt class. Three further families pass all six mechanical preconditions and yield nothing usable: on two, every discrete arm and a random legal prompt land at the 0.467 uniform-probability baseline while the continuous teacher reaches 0.066–0.110; on the third a random prompt already reaches 0.13–0.21. **We exclude those cells rather than report their apparent gaps**: these outcomes do not identify whether a gap there arises from search failure, limitations of the evaluated text-prompt class, or other pipeline differences. The exclusion was decided after those runs, not pre-registered, and the pretrained-model claim rests on two families.

G THE PROTECTED CONFIRMATION IN FULL

Methods. The study compiles a 32- or 128-position continuous teacher C, fitted per domain and seed on intent classification (ten CLINC150 domains of 15 intents), into token prompts for the same frozen model (Qwen2.5-3B-Instruct), and compares five discrete arms: direct search, PEZ, and three teacher-to-token compilers O, M and J selected on development validation (pre-registered configurations; J additionally fits the teacher's first-layer response; the compilers' objectives are specified in the released method code). The task score is the normalized multiclass Brier loss $B(p, y) = \frac{1}{2}\sum_{k=1}^{15}(p_k - \mathbf{1}\{k = y\})^2$, where the fifteen labels are shown as the answer letters A–O, each a single token (asserted by the scorer), and p is the softmax of their next-token log-probabilities at the answer position; it is averaged with equal weight per intent, and the uniform-probability baseline is $7/15 \approx 0.467$, a reference prediction rather than the expected score of an arbitrary legal prompt; intervals are paired cluster bootstraps with $50{,}000$ draws, one $\mathrm{Exp}(1)$ weight per group of exact duplicate examples of the protected split, shared by every arm and fitted seed (and, as a sensitivity, clustered by intent); they condition on the fitted artifacts. The compilers were fitted against the teacher's third checkpoint, whereas the validation-selected teacher differs in 29 of 60 groups (Table 6); fidelity, the first-layer certificate and the task comparator all refer to that fitted-against checkpoint. The quantitative compiler confirmation in Table 6 is for Qwen2.5-3B-Instruct; the development-domain gap was also measured on Falcon3-3B-Instruct (Appendix F). Three additional mechanically admissible families were excluded after exploratory runs: on two, the discrete arms and random-prompt control scored near the 0.467 uniform-probability baseline; on the third, the random-prompt control scored 0.13–0.21. These exclusions were post hoc. This study is separate from E2, which reports all 18 pre-registered cells of its three models.

The protected $4{,}500$-example split (ten domains, eight unseen in development, three seeds) was opened once, after the analysis was hashed: seven primary contrasts (Bonferroni over seven, $50{,}000$-draw paired cluster bootstrap) test whether the response-aware compiler J beats four discrete baselines; a secondary family of eleven compares every discrete arm with the continuous witness C and reports length differences. Neither family tests Theorem 1, Proposition 4 or Theorem 5. Numbers are equal-intent Brier with nominal simultaneous 95% intervals (Table 6, which also lists two disclosed deviations of the fitting from its protocol).

The gap reproduces. Every discrete arm at length 32 is worse than the 32-position witness by 0.114–0.143 (lower endpoints ≥ 0.105; ≥ 0.075 when clustering by intent), and by 0.104–0.134 on the eight domains never seen in development. Both sides are finite-budget witnesses, so this orders them and decomposes nothing.

The compiler loses, and first-layer fidelity does not buy task loss. J is worse than PEZ by 0.025 $[0.017, 0.032]$ and than direct search by 0.029 $[0.023, 0.036]$, its differences from O and M are not statistically resolved by these intervals, and is not non-inferior in accuracy (-0.028 $[-0.036, -0.020]$ at margin 0.01): the protocol's question is answered in the negative. Yet J has the best raw first-layer fidelity of the compilers (MSE$/\sigma_h^2$ 0.048 against 0.060 and 0.049; the pre-registered bounded endpoints are inert at $B_h \approx 140$), and the two methods that never look at the first layer have the best task loss: better measured first-layer fidelity need not rank these fitted methods by task loss.

What length buys, exactly. Quadrupling the discrete length moved the compilers by $D_O = -0.015$ $[-0.021, -0.008]$ and $D_J = -0.020$ $[-0.027, -0.014]$ (intervals include zero under intent clustering): 11–14% of their gap to the 32-position witness. The gaps at 128 (0.091, 0.089) are smaller mainly because the comparator worsened, $D_C = +0.034$ $[0.029, 0.039]$ — one of the three 128-position fits had Brier loss 0.178, and five of ten domains worsened; these are outcomes under the evaluated finite-budget configurations, and the experiment does not identify the cause of the degradation. Legal prompts bound the discrete class from above at their own length (Table 7): a few-shot prompt with one fitting utterance per label ($\approx$ 280 tokens, not length-matched) scores 0.154 on the protected split, better than every searched arm and 0.066 from the 32-position witness; on development rows, with the reading pre-stated, the same text truncated to 32 tokens is worse than the reported searched arm in both domains (no same-length evidence of search shortfall), and at 128 tokens it beats the reported searched arm in one domain by 0.080 — a certified shortfall of that arm from the 128-token discrete optimum there — and loses in the other.

What may and may not be called a vocabulary limit on the real model. May: the first-layer statements of §3.2–3.3, exact, length-uniform and certified. May not: the output gap — at 32 tokens consistent with either explanation and not resolved by the first-layer certificates reported here, at 128 partly certified search shortfall, at 280 tokens a hand-written few-shot prompt, not length-matched, reduces the gap to the 32-position witness from 0.114 to 0.066 (about 42%). Length does not enlarge the first-layer token-value convex envelope; it can still change attention masses, mixtures, residual representations and downstream computation, and these experiments do not identify which changes cause the task improvement: "length cannot buy vocabulary" is the model's theorem about risk and the frozen model's theorem about first-layer reach, and at the output of a deep model length with the right content buys task loss, where we claim nothing.

H E2 AUDITS IN FULL

All numbers in this appendix are post hoc, computed from frozen records (no model run, no fitting, no selection) and do not alter the pre-registered analysis; intervals are pointwise and conditional on the fitted artifacts.

Metric flow per stage. Task model during search: vLLM, temperature 0, free generation of the label field (DSPy ChatAdapter) for the text optimizers; the same frozen weights without generation for the continuous prompt. Training or proposal signal: exact match of the generated label after stripping and lower-casing (GEPA also receives "Wrong: predicted X, the correct label is Y"); cross-entropy on the gold label continuation for the continuous prompt. Candidate acceptance: native $0/1$ metric on minibatches (GEPA, MIPROv2) or sampled pool batches (SIMBA). Final selection: GEPA and MIPROv2 by native accuracy on S; SIMBA on its training pool; the continuous prompt by constrained accuracy on S, its configuration by seed-1 S accuracy. Test scoring: label-constrained for every arm, except a text arm scored natively where S preferred that. The reflection model was requested as `deepseek-v4-flash-0731` (temperature 1); the resolved backend was not logged. Native generation and constrained scoring differ by 1–3 points on S for final text programs; this does not bound differences in candidate ranking during search.

Hybrid ledger. Ten hybrids were trained; nine were scored on test: the eight the policy selected, and one it did not (Qwen MASSIVE LOW, 0.786, unused). One was never scored. A programmatic check confirms that the pre-registered H2 used exactly the policy artifacts (10 continuous, 8 hybrid), and that for every hybrid comparison the accuracy difference times the number of test rows equals corrected minus broken rows.

Resources in task-model units. Per cell, three seeds (medians over cells): the selected continuous prompt passed 54k training examples through forward and backward passes (36k LOW, 72k FULL) and scored 5.4k S rows; all its candidate configurations together, 144k. GEPA made 10.1k task-model rollouts and 31 reflection calls; MIPROv2 9.1k rows and 114 proposer calls; SIMBA 71k rows (13–26k LOW, 117–251k FULL) and 80 reflection calls. From the journal re-runs, the median reflection usage per run is 35k prompt and 20k completion tokens for MIPROv2 and 114k and 21k for SIMBA. These are unweighted counts; a backward pass, a native generation and a constrained scoring of K labels are different amounts of work.

Table 8: Selected hybrid pipeline against the DSPy kNN-16 program, same test rows and the same 16 retrieved neighbours (different wrapper), regenerated from the paired record file. Corrected: kNN wrong, hybrid right; broken: the reverse; difference $\times$ test rows $=$ corrected $-$ broken in every row. Exact two-sided McNemar test; the 95% intervals are percentile intervals from 10,000 paired resamples of test rows (seed 24021). Table 1 uses seed-averaged per-row correctness for the text comparators; this table instead compares each selected (seed-1) hybrid with the seed-1 kNN-16 artifact of the same cell (paired record file `AUDIT-FIXES-V2.json`), so its differences need not equal Table 1's H2 differences.

cell	regime	difference	95% CI	corrected / broken	p
Granite B77	FULL	$+0.0097$	$[0.0026, 0.0172]$	80 / 50	0.0107
Granite B77	LOW	$+0.0156$	$[0.0045, 0.0266]$	178 / 130	0.0073
Phi B77	FULL	$+0.0234$	$[0.0149, 0.0321]$	130 / 58	$1.6 \cdot 10^{-7}$
Phi B77	LOW	$+0.0130$	$[0.0003, 0.0256]$	219 / 179	0.0505
Phi MAS	LOW	$+0.0508$	$[0.0370, 0.0649]$	298 / 147	$7.0 \cdot 10^{-13}$
Qwen B77	FULL	$+0.0545$	$[0.0445, 0.0649]$	218 / 50	$3.5 \cdot 10^{-26}$
Qwen B77	LOW	$+0.1487$	$[0.1334, 0.1643]$	557 / 99	$2.7 \cdot 10^{-78}$
Qwen MAS	FULL	$+0.0454$	$[0.0343, 0.0565]$	211 / 76	$7.4 \cdot 10^{-16}$

Resource ledger of Section 5.1. The counts below describe the bare continuous-prompt comparison (H1) and exclude hybrid training and retrieval: ten hybrids were trained for H2 (4.5 h of training), of which the eight selected used 4.0 h; retrieval and indexing costs are not quantified, and the reflection backend actually resolved was not logged. Counted in unweighted task-example presentations per cell over three seeds, the selected continuous prompt used 3.5–7.1 times GEPA's and 3.9–7.9 times MIPROv2's rollouts, and 9.5–26 times with its whole configuration search on S included; SIMBA used more task-model rows than the selected continuous prompt in half the cells. The text optimizers in turn call an external reflection model (median calls per cell over three seeds: 31 for GEPA, 80 for SIMBA, 114 for MIPROv2). Wall-clock on mixed hardware: 36.2 h for the continuous prompt including selection (13.0 h for the chosen configurations), 17.1 h GEPA, 13.8 h MIPROv2, 43.7 h SIMBA. Retrospective checkpoint comparisons on S do not establish an efficiency advantage: in LOW, at similar unweighted task-example counts, the selected continuous prompt's selection score is a median 0.0 points from MIPROv2's.

Classes and payloads. On Banking77, seed-averaged minimum class recall is 0.31–0.69 for the continuous prompt against 0.00–0.04 for the strongest optimizer, which has 1–4 classes of zero empirical recall in five of six cells; the pattern is not universal, since on MASSIVE the continuous prompt also has a zero-recall class (Granite FULL, where every method has one). GEPA's programs carry no demonstrations (170–420-word instructions), MIPROv2's a median of four and SIMBA's two, against the one-per-class baseline's K; the association between demonstration coverage and per-class accuracy is inconsistent across datasets and does not identify the effect of coverage.

More data, different gains. Moving from LOW to FULL improves the continuous prompt by 7–14 points on the intent datasets, whereas the configured text optimizers show no consistent gain (GEPA -4 to $+4.5$, MIPROv2 -5 to $+6$, SIMBA -10 to $+9$). Because the contrast also changes the continuous prompt's training work and, for SIMBA, its task-model evaluation, it characterises these pipelines, not the capacity of the text class.

Retrospective checkpoint comparison on S. For each seed-cell the continuous prompt's best checkpoint S accuracy within the optimizer's task-row budget is compared with the optimizer's reported S accuracy. Only LOW has checkpoints inside the GEPA and MIPROv2 budgets (the first checkpoint is at 2,000 examples plus 300 S rows). Strict wins / ties / losses: against GEPA $20/0/7$ (median $+5.3$ points), against MIPROv2 $14/1/12$ (median 0.0), against SIMBA $25/0/2$ in LOW and $27/0/0$ in FULL. The early checkpoints belong to a configuration chosen with the whole later search, SIMBA never selects on S, the scorers differ, and both maxima are selection-biased: this is a diagnostic, not a matched-budget result.

Retrieval strata. The gold label is among the 16 retrieved neighbours for 93–100% of test rows. On Banking77 the continuous prompt trails kNN-16 on gold-retrieved rows by 0.4–6.6 points; on MASSIVE it leads (by 2–4 points, one cell at -0.4); on TREC it leads by 7–24. Gold-absent strata are too small to interpret, and gold presence is a coarse diagnostic of retrieval quality.

Test-split incidents. No artifact was re-run because of its value. Two written test values were voided and re-scored: a Qwen kNN-16 native pass exhausted the open-file limit and scored the affected rows as wrong (a bias in our favour), and a Phi GEPA pass was written as 0 while the inference server was down. Four post-freeze deviations wrote no test value. Unparseable native answers are counted wrong on every row, as they already were on S. A vectorised scorer gave identical argmax on check rows from S. Granite hybrids were trained early, but only the admitted ones were scored. Phi's Banking77 kNN-64 was redone on S with a longer context. CLINC150 is excluded because its test split is the protected set of Appendix G. COPRO was dropped before test, and LoRA was not run.

I BUDGET-EXPANDED OPTIMIZERS AND A COMMON DECODER

Budget-expanded optimizers. The addendum (supplement, `BUDGET-ADDENDUM-B.md`) was written and hashed before any of its runs. It fixed the cells by a public permutation, independent of outcomes, and gave GEPA a metric-call budget of 13,800 (about 3,400 at `heavy`) and MIPROv2 its `heavy` configuration scaled by $m = 4.53$ (38 candidates, 122 trials, minibatch 35, full evaluation every five steps). The reflector, temperature, pools, S, demonstration limits and seeds were unchanged, and each optimizer selected on S as in E2. One run was repeated once from scratch after its host's GPU was shown to corrupt generation; the failed run's logs are in the supplement. The text arms are scored natively in both columns of Table 9.

Common decoder. Records the protocol had already scored with the label-constrained decoder were reused; the remaining 108 programs were scored by the same code (`e2_common_decoder.py`) under the frozen protocol hash gate. Nothing was selected on these numbers, and Table 1 is unchanged.

Table 9: Budget-expanded text optimizers (post hoc addendum; three LOW cells, one per model and dataset, seeds 1–3). Test accuracy in %, seed means; the text arms are scored natively in both columns, the continuous prompt as in Table 1. Rows: task-model rows used in search per seed (allowance 13,800, the continuous prompt's per-seed presentations). Intervals: 95% percentile intervals from 10,000 paired row bootstraps over the test rows, conditional on the artifacts.

cell	optimizer	default	budget	budget − default	continuous	continuous − budget	rows / seed
Granite MAS	GEPA	69.3	66.7	−2.7 [−3.5, −1.8]	81.7	+15.0 [+13.8, +16.3]	13.96k
Granite MAS	MIPROv2	67.9	68.8	+0.9 [+0.2, +1.7]	81.7	+12.8 [+11.5, +14.1]	12.19k
Phi TREC	GEPA	78.2	82.9	+4.7 [+2.7, +6.8]	88.9	+5.9 [+3.1, +8.9]	13.96k
Phi TREC	MIPROv2	62.8	71.9	+9.1 [+6.2, +11.9]	88.9	+17.0 [+14.1, +19.9]	12.24k
Qwen B77	GEPA	50.8	51.1	+0.3 [−0.7, +1.3]	69.5	+18.4 [+17.0, +19.8]	14.14k
Qwen B77	MIPROv2	57.4	56.9	−0.5 [−1.2, +0.2]	69.5	+12.6 [+11.1, +14.1]	12.24k

J E2 REFERENCE-CLASS LADDER

K RELATED WORK IN FULL

Two recent results bracket our claim from opposite sides, and together they turn our separation from an assertion into a consequence.

The continuous side is universal. Hsu & Lai (2026) show, in an approximation-theoretic sense, that a *soft* prompt suffices to steer even a randomly initialised transformer to approximate any Hölder function on a compact manifold, with an explicit construction resting on a rank condition that Gaussian initialisation satisfies almost surely. Pretraining is optional; the soft prompt alone carries the

Table 10: Common-decoder audit (post hoc): every arm scored on test by the same label-constrained decoder, seed means. Bold: the best optimizer. Δ: continuous prompt − best optimizer, 95% paired row-bootstrap interval with the best comparator recomputed per resample (10,000 resamples).

model	data	reg.	soft	zero-shot	one-shot	kNN-16	kNN-64	GEPA	MIPROv2	SIMBA	Δ [95% CI]
Qwen	B77	LOW	.695	.436	.606	.722	.743	.447	**.564**	.554	+.131 [.116,.146]
Qwen	B77	FULL	.829	.436	.610	.877	.837	.464	.540	**.547**	+.282 [.268,.298]
Qwen	MAS	LOW	.783	.535	.683	.751	.731	.619	.665	**.675**	+.108 [.095,.121]
Qwen	MAS	FULL	.850	.537	.688	.852	.849	.603	**.674**	.662	+.176 [.162,.190]
Qwen	TREC	LOW	.872	.464	.661	.804	.824	.697	**.747**	.663	+.125 [.091,.161]
Qwen	TREC	FULL	.968	.464	.701	.842	.860	.743	**.796**	.756	+.172 [.143,.203]
Phi	B77	LOW	.767	.551	.596	.829	.757	.548	**.557**	.555	+.209 [.193,.224]
Phi	B77	FULL	.910	.550	.565	.912	.887	.540	**.568**	.553	+.343 [.327,.359]
Phi	MAS	LOW	.792	.455	.666	.751	.773	.566	**.619**	.610	+.173 [.159,.187]
Phi	MAS	FULL	.879	.462	.677	.841	.856	.555	.610	**.625**	+.254 [.240,.268]
Phi	TREC	LOW	.889	.566	.614	.702	.726	**.811**	.664	.742	+.077 [.051,.105]
Phi	TREC	FULL	.972	.566	.669	.772	.802	**.834**	.724	.777	+.138 [.111,.165]
Granite	B77	LOW	.830	.562	.700	.858	.837	.550	**.636**	.620	+.194 [.179,.209]
Granite	B77	FULL	.913	.561	.691	.930	.916	.568	.627	**.628**	+.285 [.269,.299]
Granite	MAS	LOW	.817	.644	.645	.779	.777	.682	.683	**.691**	+.125 [.112,.138]
Granite	MAS	FULL	.886	.648	.709	.864	.860	.654	**.692**	.687	+.194 [.181,.207]
Granite	TREC	LOW	.949	.534	.746	.716	.788	**.871**	.746	.771	+.078 [.055,.102]
Granite	TREC	FULL	.931	.534	.780	.848	.820	**.826**	.810	.670	+.105 [.077,.131]

Table 11: E2 reference-class ladder, test accuracy (post hoc). Zero-shot, best optimizer, one-shot and best text method: seed means under the S-selected scorer, as in Table 1. Calibration (per-class bias on the zero-shot label scores), probe (logistic regression on frozen features) and PEZ: seed 1, label-likelihood scorer. Soft: three-seed mean. These are descriptive comparisons across pipeline classes and scorers, not a nested expressivity hierarchy. Paired intervals for every step: `POSTHOC-E2.json`.

model	data	reg.	zero-shot	calib.	probe	soft	best opt.	one-shot	best text	PEZ
Qwen	B77	LOW	.501	.624	.834	.695	.574	.608	.784	.239
Qwen	B77	FULL	.500	.621	.924	.829	.557	.613	.878	.293
Qwen	MAS	LOW	.628	.716	.787	.783	.675	.683	.751	.398
Qwen	MAS	FULL	.630	.750	.881	.850	.674	.688	.855	.337
Qwen	TREC	LOW	.464	.690	.806	.872	.747	.661	.804	.206
Qwen	TREC	FULL	.464	.692	.964	.968	.756	.701	.872	.258
Phi	B77	LOW	.551	.639	.832	.767	.557	.596	.829	.378
Phi	B77	FULL	.550	.642	.928	.910	.568	.565	.912	.408
Phi	MAS	LOW	.427	.703	.783	.792	.536	.666	.751	.394
Phi	MAS	FULL	.427	.745	.883	.879	.625	.677	.856	.356
Phi	TREC	LOW	.432	.822	.882	.889	.782	.614	.782	.294
Phi	TREC	FULL	.440	.822	.958	.972	.816	.669	.816	.270
Granite	B77	LOW	.554	.648	.831	.830	.636	.691	.858	.299
Granite	B77	FULL	.551	.643	.915	.913	.628	.691	.930	.234
Granite	MAS	LOW	.644	.739	.813	.817	.691	.641	.779	.368
Granite	MAS	FULL	.640	.782	.881	.886	.692	.706	.864	.381
Granite	TREC	LOW	.534	.762	.784	.949	.871	.746	.871	.276
Granite	TREC	FULL	.534	.794	.956	.931	.829	.780	.850	.264

function. Their scope is stated plainly: they study soft prompting, and set discrete prompting aside as "a discrete optimization problem, which makes it difficult from an optimization standpoint".

Their observation concerns optimization difficulty. Our theorem concerns the discrete-class risk infimum in a different model; the two statements are compatible, and we do not conjoin the two results across models.

The closest prior theory, and what separates ours from it. Castin et al. (2025) and Boursier & Boyer (2025) give the linearity of attention over Gaussian tokens that Corollary 3 makes adaptive. Petrov et al. (2024a) give a theory of the capabilities and limitations of prompting and prefix-tuning,

and the relation to our result is close enough that it should be stated precisely rather than gestured at.

Their Theorems 1 and 2 construct transformers in which a single continuous virtual token realises exponentially more input–output maps than any discrete token. That is a *cardinality* separation, at *fixed* prefix length, and their own appendix notes that for the construction the choice of embedding matrix does not matter. Ours is a statement about *risk*, uniform over *all* lengths, whose value is a function of the vocabulary's geometry alone. The two cannot be substituted for one another. In our full-self model, Appendix B proves that the real-mass, integer-count and sufficiently length-capped optima have the same asymptotic limit, by an explicit clipped-weight construction, rounding and a common lower bound; we do not assert fixed-r density of integer-count responses in the real-mass class. Theorem 1 determines the discrete-class risk against the specified target; the fixed-bias soft teacher has a small self-attention discrepancy, and we do not identify its exact risk gap with $\mathcal{R}(s)$. We do not claim the existence of a continuous–discrete gap as new; what is new is the exact discrete-class risk in reach, its threshold at $s = \theta$, its independence of length, and a finite certificate.

Precisely, our model is the one-layer scalar instance of their equation (7) with the prefix values confined to a vocabulary. They give the form of the bias a prefix can add; the frontier is what that form can reach when the atoms must be drawn from a dictionary of reach s. We quantify a term they leave qualitative.

Their second result — that at an attention layer, for fixed incoming content representations, a prefix cannot alter the relative attention pattern over content, only scale it down and add a bias (later layers can still change through earlier representations) — constrains soft and hard prefixes *alike*, so it cannot separate them, and the same authors later show that this structure is compatible with universal approximation (Petrov et al., 2024b). No lower bound on task loss follows from that structure alone; it is the reason we record the pretrained-model gap as measured rather than decomposed.

Two results the frontier must be read against. Wang et al. (2023) prove (Theorem 2) that a given single-layer transformer has datasets no soft prompt of any length memorises — a qualitative "no length suffices" for the continuous class on a constructed model; Theorem 1 is the discrete class's residual as a closed form in reach on a random model. Castin et al. (2025) show that attention preserves Gaussian measures and Boursier & Boyer (2025) (Lemma 2.1) that softmax attention over a Gaussian population of tokens is linear in the query; Corollary 3 is the adaptive counterpart: over every reweighting of a finite dictionary drawn from that population the asymptotically optimal response is still linear, and reweighting buys a slope at most $\log V/\sqrt{r}$.

Prompting has other fundamental limits, on a different axis. Meyer et al. (2025) prove memory limitations of prompt tuning in *arbitrary* rather than specially constructed transformers, including masked self-attention, giving a formal account of the empirically familiar degradation on long prompts. Their limit is on the prompt-length axis; ours is on the vocabulary axis. Their results bound memorization capacity under their stated assumptions, including a bound independent of prompt length; our finite-s theorem instead identifies a vocabulary-dependent risk bound, and neither result diagnoses the cause of an individual failed training run.

Their critique of earlier universality results — that those rely on synthetically constructed transformers whose weights are designed to make prompt tuning work — does not apply to Hsu & Lai (2026), which is why we lean on the latter.

What we do not overlap with. Rate-distortion and quantisation theory bound distortion at a given rate; our result identifies the optimum within a specific constrained attention class as a function of dictionary reach, and we do not claim that threshold behaviour itself distinguishes it from rate-distortion theory. Two-atom optimality is classical extreme-point structure and we claim no priority for it.

The operation the theory is about is one practice already performs. Vocabulary *content* rather than size is the binding resource here. Smith et al. (2026) expand a pretrained model's tokenizer and adapt its weights for encoding efficiency; that intervention differs from our fixed-backbone dictionary change, and our result does not predict its gains. Our support-repair intervention is *told* which pair to insert, so identifying the missing atoms in a real dictionary remains open.