WHEN DOES POSTERIOR GEOMETRY MATTER FOR IN-CONTEXT DEMONSTRATION SELECTION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Posterior-based demonstration selection uses label-distribution compatibility to refine semantic retrieval. Understanding these selections requires identifying which posterior components can change the selected examples. We analyze candidate posterior tails, comprising probabilities assigned to labels other than the observed candidate label, at fixed query posteriors and observed-label probabilities. For Jensen–Shannon alignment, we derive exact attainable score intervals and establish a necessary and sufficient condition for selection invariance. Every non-invariant observed set admits a constructive counterexample involving at most two candidate tails. This characterization determines when posterior variation can reverse the ordering at the selection boundary. A five-task audit certifies 32.71% of selected sets under equal task and seed weighting, with rates ranging from 0.05% to 89.64%. In controlled nine-task comparisons, full posterior alignment improves macro accuracy over a query-label control by 0.83 percentage points on Mistral-7B, while the two selectors achieve similar means on Qwen2.5-7B. The certificate permits recovery of the full selector's demonstration set without candidate-tail access on certified queries. Code is available at `https://anonymous.4open.science/r/posterior-geometry-5376`.

1 INTRODUCTION

In-context learning (ICL) enables a large language model (LLM) to make predictions from demonstrations included in its prompt without parameter updates (Brown et al., 2020). Demonstration selection and ordering can substantially affect predictive performance (Liu et al., 2022; Rubin et al., 2022; Lu et al., 2022). Semantic retrieval provides a relevance criterion by selecting inputs similar to the query. However, semantic similarity does not ensure compatible label evidence: negation or a fine-grained intent distinction can change an example's label while largely preserving its input representation.

Output-aware selection incorporates label information beyond semantic relevance. Existing approaches score candidates through the target LLM (Peng et al., 2024; Zhang et al., 2025) or incorporate observed labels into retrieval (Chen et al., 2026a). Predictive accuracy evaluates the utility of a scoring rule. Explaining its selections requires a further step: identifying the posterior information that determines which examples enter the prompt. We therefore examine how posterior components affect individual selection decisions.

For posterior-based selectors, a candidate's observed label separates its distribution into the probability assigned to that label and the allocation among competing labels. We investigate whether the latter allocation can change demonstration membership when the former probability remains fixed. Posterior variation can alter scores while leaving the selected set unchanged. A membership change requires an excluded candidate to outrank a selected candidate. Characterizing this boundary identifies when competing-label probabilities determine selection and when this information can be omitted.

We formulate this analysis for our *full posterior alignment* (FPA) selector. An auxiliary posterior model (APM) estimates a distribution over task labels for each input. FPA combines semantic relevance with query label probabilities and Jensen–Shannon (JS) alignment between query and candidate posteriors. Candidate estimates are obtained by cross-fitting and anchored toward their observed labels. The analysis fixes query posteriors, candidate probabilities at their observed labels, and all other scoring inputs. It varies only the allocation among competing labels, termed the candidate's *posterior tail*.

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.

The central analytical challenge is to establish attainability. A general divergence bound provides a sufficient condition for invariance; failure to satisfy that condition can reflect either a possible membership change or slack in the bound. We derive exact attainable score intervals, which support a necessary and sufficient condition for invariance under all admissible tail reallocations. For a non-invariant observed set, the endpoint constructions provide a membership-changing perturbation involving at most two candidate tails. The analysis thereby resolves posterior dependence for individual queries and determines when candidate tails can be omitted.

Classification comparisons evaluate the predictive utility of the complete scoring rules. The LABEL-MASS baseline combines semantic relevance with the *query* probability at each candidate's observed label and uses no candidate posterior. In nine-task comparisons, FPA improves macro accuracy by 0.83 percentage points on Mistral-7B, with a paired 95% confidence interval above zero; the two selectors have similar means on Qwen2.5-7B (Table 1). The certificate determines when candidate tails can be omitted while preserving FPA membership.

The main contributions are:

1. We derive the exact attainable interval of the competing-label contribution to JS alignment and characterize the posterior distributions attaining both endpoints.
2. We establish a necessary and sufficient condition for Top-L invariance and construct a membership-changing perturbation of at most two candidate tails for every non-invariant observed set. The same analysis gives exact rank bounds and enables membership recovery without candidate-tail access on certified queries.
3. We quantify selection invariance in a five-task audit of 100,455 query–seed cases, certifying 32.71% of sets under equal task and seed weighting. Numerical checks validate the analytical constructions, and paired classification comparisons establish a 0.83 percentage points macro accuracy improvement on Mistral, with similar mean accuracies on Qwen.

2 RELATED WORK

Demonstration selection. Semantic and learned retrievers select examples relevant to a query (Liu et al., 2022; Rubin et al., 2022). Other approaches rank demonstration sets and orderings by description length (Wu et al., 2023), or use cross-entropy differences and conditional input likelihood for retrieval (Iter et al., 2023; Peng et al., 2024). D.Va evaluates candidates using a held-out memory example (Zhang et al., 2025), while TopK-SD incorporates class information into the retrieval representation (Chen et al., 2026a). We analyze how candidate-tail allocation determines membership under posterior alignment.

Stability and uncertainty in ICL. Transformer stability and generalization (Wang et al., 2026), robust adaptation after adversarial pretraining (Kumano et al., 2026), and uncertainty before response generation (Chen et al., 2026b) concern properties of the predictor. We examine invariance of the selected demonstration set under candidate-tail perturbations, with query posteriors and the remaining scoring inputs fixed.

Divergence and uncertain rankings. JS divergence (Lin, 1991), uncertain ranking (Wang et al., 2011), and robust selection (Chassein et al., 2018) provide the foundations for this analysis. Interval separation supports Top-k identification (Ré et al., 2006), and certain and possible ranking answers have been studied under data uncertainty (Feng et al., 2023). We derive the exact score range induced by fixed-mass JS tail variation and characterize the posterior distributions attaining its endpoints. These constructions connect the interval criterion to admissible posterior perturbations, providing invariance guarantees and explicit counterexamples.

3 OUTPUT-AWARE DEMONSTRATION SELECTION

3.1 SEMANTIC RETRIEVAL AND AUXILIARY POSTERIORS

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be a labeled memory, where x_i is an input and $y_i \in \{1, \ldots, C\}$ is its observed label, with $C \geq 2$. For a query x, a frozen encoder h produces embeddings with unit Euclidean

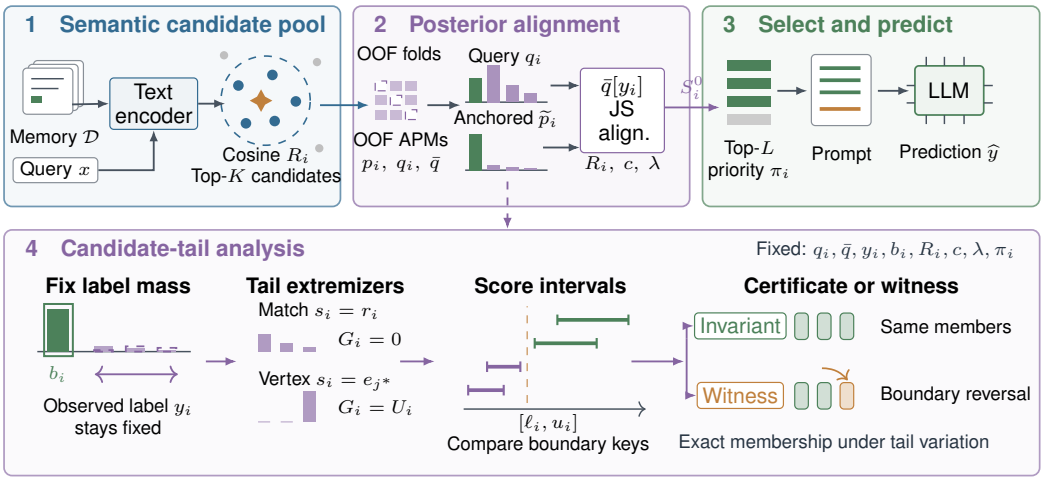

Figure 1: **Posterior alignment and selection invariance.** The upper path combines semantic retrieval and posterior alignment for demonstration selection. The lower path derives attainable score intervals under candidate-tail variation, supporting an invariance certificate or a constructive counterexample for the same selection rule.

norm. Semantic relevance is measured by the cosine score $R_i = h(x)^\top h(x_i)$. The index set $\mathcal{C}_K(x) \subseteq \{1, \dots, N\}$ contains the K highest-scoring candidates, with $1 \le K \le N$. The selector retains L candidates, with $0 \le L \le K$. Deterministic resolution of semantic ties assigns each candidate a distinct priority π_i according to semantic rank; smaller values take precedence.

For a probability vector u, $u[j]$ denotes its probability at label j. We write $\Delta^{d-1} = \{u \in \mathbb{R}^d : u[j] \ge 0,\ \sum_{j=1}^d u[j] = 1\}$ for the probability simplex on d coordinates. All logarithms are base two. Entropy is $H(u) = -\sum_j u[j] \log_2 u[j]$, with $0 \log_2 0 = 0$.

An auxiliary posterior model (APM) estimates label probabilities independently of target-LLM inference. The memory is partitioned into $F \ge 2$ stratified folds, and classifier $P^{(f)}$ is trained with fold f excluded. For candidate i in fold $f(i)$, define

$$p_i = P^{(f(i))}(\cdot \mid x_i), \qquad q_i = P^{(f(i))}(\cdot \mid x), \qquad \bar{q} = \frac{1}{F} \sum_{f=1}^{F} P^{(f)}(\cdot \mid x). \tag{1}$$

All three vectors belong to Δ^{C-1}. The out-of-fold (OOF) estimate p_i and query posterior q_i use the same head; $\bar{q}$ averages query posteriors across heads. The index i in q_i identifies the candidate-associated head, with the query x fixed. Candidate posteriors are cached offline, and each query is evaluated by the F auxiliary heads. Figure 1 presents the selection procedure and the analysis in Sections 4 and 5.

3.2 Full posterior alignment

Let $e_{y_i} \in \mathbb{R}^C$ be the one-hot vector at the observed label y_i. We anchor the raw candidate posterior toward this vector and compute a query concentration weight:

$$\rho_i = p_i[y_i], \qquad \widetilde{p}_i = \rho_i p_i + (1 - \rho_i) e_{y_i}, \qquad c = 1 - \frac{H(\bar{q})}{\log_2 C}. \tag{2}$$

Thus $\widetilde{p}_i$ is the anchored candidate posterior and $c \in [0, 1]$ balances direct label evidence against distributional alignment. We use Jensen–Shannon (JS) divergence, defined in Section 4, to compare probability vectors. With base-two logarithms, $1 - \mathrm{JS}$ is an alignment score in $[0, 1]$.

For any candidate posterior $v \in \Delta^{C-1}$, define the scoring function

$$S_i(v) = \lambda R_i + (1 - \lambda)\big[c\bar{q}[y_i] + (1 - c)\{1 - \mathrm{JS}(q_i, v)\}\big], \qquad 0 \le \lambda \le 1. \tag{3}$$

Our full posterior alignment (FPA) selector evaluates the observed score $S_i^0 = S_i(\widetilde{p}_i)$ and retains the L largest keys $(S_i^0, -\pi_i)$. Lexicographic comparison prioritizes larger scores and resolves equal scores in favor of smaller π_i. The target LLM predicts a label $\widehat{y}$ from a prompt containing the selected demonstrations, as shown in the upper path of Figure 1. Experiments use $F = 3$, $K = 30$, $L = 8$, and $\lambda = 0.5$.

The LABEL-MASS baseline retains semantic relevance and direct query label evidence:

$$S_i^{\mathrm{LM}} = \lambda R_i + (1 - \lambda)\bar{q}[y_i].$$

The probability $\bar{q}[y_i]$ is evaluated under the query posterior at the candidate's observed label. Both selectors use observed memory labels without accessing the query's gold label.

3.3 UNCERTAINTY IN THE CANDIDATE TAILS

Let $b_i = \widetilde{p}_i[y_i]$ denote the anchored mass at the candidate's observed label. Its *posterior tail* consists of the probabilities at all other labels. We reserve p_i for the raw estimate, $\widetilde{p}_i$ for the observed anchored estimate, and v_i for a posterior allowed to vary in the analysis. The admissible sets are

$$\mathcal{U}_i = \{v_i \in \Delta^{C-1} : v_i[y_i] = b_i\}, \qquad \mathcal{U} = \prod_{i \in \mathcal{C}_K(x)} \mathcal{U}_i. \tag{4}$$

An element $\mathbf{v} = (v_i)_{i \in \mathcal{C}_K(x)} \in \mathcal{U}$ specifies one admissible posterior for every candidate. The query posteriors q_i, their mean $\bar{q}$, labels y_i, masses b_i, semantic scores R_i, weights c, λ, and priorities π_i remain fixed. Only the allocation of the remaining mass $1 - b_i$ varies.

The studied anchor gives $b_i = 1 - \rho_i + \rho_i^2 \in [3/4, 1]$; the analysis below also applies to arbitrary fixed masses $b_i \in [0, 1]$. The results apply to calibrated and uncalibrated posterior estimates. The Cartesian product in Equation (4) permits independent reallocations of the estimated candidate tails, so their score extrema are jointly attainable. The lower path of Figure 1 connects these perturbations to score intervals and membership guarantees.

4 EXACT SCORE INTERVALS UNDER TAIL VARIATION

4.1 SEPARATING LABEL MASS FROM COMPETING LABELS

For probability vectors r, s, Kullback–Leibler divergence is $\mathrm{KL}(r\|s) = \sum_j r_j \log_2(r_j/s_j)$, with zero contributions for $r_j = 0$ and value $+\infty$ if $r_j > 0 = s_j$. For $0 < \eta < 1$, weighted JS is

$$\mathrm{JS}_\eta(r, s) = \eta\,\mathrm{KL}(r\|m) + (1 - \eta)\,\mathrm{KL}(s\|m), \qquad m = \eta r + (1 - \eta)s. \tag{5}$$

At $\eta \in \{0, 1\}$ its continuous extension is zero; ordinary JS uses $\eta = 1/2$ (Lin, 1991). For a single candidate, suppress the index and write q, p for q_i, v_i, respectively. The symbol p denotes an admissible posterior; p_i remains the raw estimate. Ordering the observed label first gives

$$q = (a, (1 - a)r), \quad p = (b, (1 - b)s), \quad w = \frac{2 - a - b}{2}, \quad \eta = \frac{1 - a}{2 - a - b}. \tag{6}$$

Here $a, b \in [0, 1]$ are the query and candidate probabilities at the observed label, and $r, s \in \Delta^{C-2}$ are normalized distributions over the other $d = C - 1$ labels. If a tail has zero mass, its normalized vector is arbitrary. When $a = b = 1$, set $w = 0$ and $\eta = 1/2$; the residual below is zero.

Lemma 1 (Separating the observed label)**.** *Grouping the candidate label and its complement gives*

$$\mathrm{JS}(q, p) = \underbrace{\mathrm{JS}((a, 1 - a), (b, 1 - b))}_{B(a,b)} + \underbrace{w\,\mathrm{JS}_\eta(r, s)}_{G(r,s;a,b)}. \tag{7}$$

The decomposition separates disagreement in observed-label mass from disagreement among competing labels. Grouping the KL terms gives the identity (Appendix A). At fixed q and b, tail reallocation changes only the residual G.

The generic bound $G \le wH_2(\eta)$, with binary entropy $H_2(\eta) = H((\eta, 1 - \eta))$, depends on the tail masses. Conditioning additionally on the query tail r gives the largest residual attainable for that query.

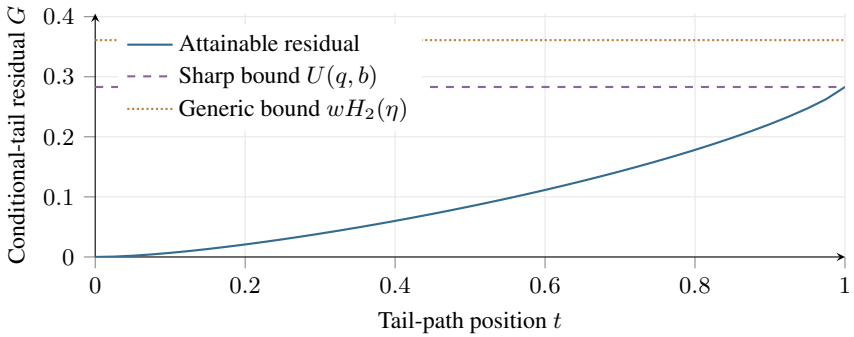

Figure 2: **Attainment of the sharp tail bound.** For $q = (0.2, 0.56, 0.20, 0.04)$ and $b = 0.8$, the path from $r = (0.7, 0.25, 0.05)$ to the least-probable query-tail vertex attains $U(q, b)$ at $t = 1$. The maximum lies below the generic entropy bound. Curves evaluate Equation (7) directly.

4.2 The extremal candidate tail

Theorem 2 (Exact attainable tail interval)**.** *Fix the complete query posterior q and the candidate mass b. Define $r_{\min} = \min_j r_j$, $m_z = 1 - \eta + \eta z$, and*

$$\Phi_\eta(z) = -m_z \log_2 m_z - \eta(1 - z) \log_2 \eta + \eta z \log_2 z. \tag{8}$$

Boundary values in Equation (8) are defined by continuity. As s varies over Δ^{C-2}, the attainable values of G are exactly

$$[0, U(q, b)], \qquad U(q, b) = w\Phi_\eta(r_{\min}). \tag{9}$$

The minimum is attained at $s = r$. A maximum is attained at $s = e_{j^*}$ for any $j^* \in \arg\min_{1 \le j \le d} r_j$, where $e_{j^*} \in \mathbb{R}^d$ is the corresponding tail one-hot vector. If either tail has zero mass, or $C = 2$, then $U = 0$.

For positive tail masses, convexity of $\mathrm{JS}_\eta(r, s)$ in s ensures that a maximum occurs at a simplex vertex. Direct expansion gives $\mathrm{JS}_\eta(r, e_j) = \Phi_\eta(r_j)$, with

$$\Phi'_\eta(z) = \eta \log_2 \frac{\eta z}{1 - \eta + \eta z} < 0, \qquad 0 < \eta < 1, \quad z \in (0, 1]. \tag{10}$$

The maximizing vertex therefore corresponds to a least-probable query coordinate. The matched tail attains zero, and continuity along $(1 - t)r + te_{j^*}$ establishes attainment of every intermediate value. Appendix A.2 provides the complete proof, including boundary cases. Figure 2 evaluates the residual along this path to the extremal tail.

Dependence on the query tail. The extremal allocation concentrates the entire candidate-tail mass on a least-probable competing label under the query. Because $\Phi_\eta(0) = H_2(\eta)$ and Φ_η decreases strictly for $0 < \eta < 1$, the generic bound is tight when the query tail has a zero coordinate, but strictly loose when that tail has full support and both tail masses are positive. For a uniform tail over $d = C - 1$ labels, the exact bound becomes $w\Phi_\eta(1/d)$.

4.3 From tail divergence to an exact score set

For candidate i, let $a_i = q_i[y_i]$, $r_i = q_{i,-y_i}/(1 - a_i)$, and $s_i = v_{i,-y_i}/(1 - b_i)$, where the subscript $-y_i$ removes the observed-label coordinate and preserves the order of the others. Choose a normalized tail arbitrarily if its denominator is zero. Define $w_i = (2 - a_i - b_i)/2$ and $\eta_i = (1 - a_i)/(2 - a_i - b_i)$, with $\eta_i = 1/2$ when $a_i = b_i = 1$. Write $B_i = B(a_i, b_i)$, $G_i = G(r_i, s_i; a_i, b_i)$, $U_i = U(q_i, b_i)$, and abbreviate $S_i(v_i)$ as S_i. The score then separates as

$$M_i = \lambda R_i + (1 - \lambda)\big[c\bar{q}[y_i] + (1 - c)(1 - B_i)\big], \tag{11}$$

$$\kappa = (1 - \lambda)(1 - c), \qquad S_i = M_i - \kappa G_i. \tag{12}$$

Theorem 2 gives the exact interval

$$\ell_i = M_i - \kappa U_i \;\le\; S_i \;\le\; u_i = M_i. \tag{13}$$

The matched tail attains the upper score, and the maximum-divergence tail attains the lower score. Independent candidate-tail reallocations give the exact joint score range $\prod_i [\ell_i, u_i]$. Joint endpoint attainment establishes necessity in the invariance criterion; an outer bound alone provides sufficiency.

The width κU_i combines the tail weight and attainable divergence. It vanishes for binary labels, zero tail mass, or $\kappa = 0$. Membership also depends on score separation: a wide interval can remain above all excluded candidates, while a narrow interval can intersect the selection boundary.

For binary labels, fixing the observed-label mass also fixes the sole remaining probability. The zero interval width expresses the absence of allocation freedom, while query and candidate label masses still contribute to the score. With more than two labels, distinct allocations can share the same remaining mass; U_i quantifies the resulting variation in the tail contribution.

The coarse score M_i removes the tail residual but retains the Bernoulli divergence B_i and entropy mixture; LABEL-MASS instead uses $\bar{q}[y_i]$ as its output term. For an invariant set, matching candidate and query tails reduces FPA to M_i while preserving membership. The coarse and full scores therefore select the same members on certified queries.

5 CERTIFYING SELECTION AND CONSTRUCTING COUNTEREXAMPLES

5.1 NECESSARY AND SUFFICIENT INVARIANCE CONDITION

Fix a query and its candidate index set $\mathcal{C}_K(x)$. Define the lower and upper keys of candidate i as $k_i^- = (\ell_i, -\pi_i)$ and $k_i^+ = (u_i, -\pi_i)$. The relation $>_{\text{lex}}$ denotes the lexicographic order from Section 3, with the second coordinate resolving equality in the first. All subsequent minima and maxima of keys use this order. Distinct priorities preclude equal keys for different candidates.

Theorem 3 (Top-L invariance with score ties)**.** *For $0 < L < K$, let $A \subseteq \mathcal{C}_K(x)$ be a set with $|A| = L$. Then A is the Top-L set for every admissible configuration $\mathbf{v} \in \mathcal{U}$ if and only if*

$$\min_{i \in A} k_i^- \;>_{\text{lex}}\; \max_{j \in \mathcal{C}_K(x) \setminus A} k_j^+. \tag{14}$$

For $L = 0$ or $L = K$, membership is trivially invariant.

Under the stated condition, every selected lower key exceeds every excluded upper key. Since admissible scores remain within their respective intervals, every selected candidate precedes every excluded candidate. The Top-L set is therefore A throughout $\mathcal{U}$.

For the converse, let i^- attain the smallest lower key in A and j^+ the largest upper key in $\mathcal{C}_K(x) \setminus A$. Violation of the condition implies $k_{j^+}^+ >_{\text{lex}} k_{i^-}^-$ because the priorities are distinct. A vertex at a smallest coordinate of r_{i^-} attains ℓ_{i^-}, while the matched tail r_{j^+} attains u_{j^+}. These endpoints are simultaneously admissible, with all other tails unchanged. The resulting order places j^+ ahead of i^- and hence produces a Top-L set different from A.

Figure 1 summarizes the certificate and counterexample constructions, and Algorithm 1 implements them. The attainable endpoint posteriors realize a membership change within $\mathcal{U}$.

For example, let $L = 1$, $[\ell_1, u_1] = [0.5, 0.5]$, $[\ell_2, u_2] = [0.2, 0.5]$, and $\pi_1 < \pi_2$. Candidate 1 remains selected despite the zero scalar margin, whereas reversing the priorities removes this guarantee.

5.2 DIRECTIONAL VARIATION, MARGINS, AND ATTAINABLE RANKS

Define the observed residual as $G_i^0 = \text{JS}(q_i, \widetilde{p}_i) - B_i$. Then the observed score from Equation (3) satisfies $S_i^0 = S_i(\widetilde{p}_i) = M_i - \kappa G_i^0$. The largest attainable decrease and increase from that score are, respectively,

$$d_i^- = \kappa(U_i - G_i^0), \qquad d_i^+ = \kappa G_i^0, \qquad [\ell_i, u_i] = [S_i^0 - d_i^-, S_i^0 + d_i^+]. \tag{15}$$

Both quantities are nonnegative because $0 \le G_i^0 \le U_i$. For a candidate i in the observed Top-L set and a candidate j outside it, the smallest attainable score difference is

$$\ell_i - u_j = (S_i^0 - S_j^0) - (d_i^- + d_j^+). \tag{16}$$

Invariance requires each selected–excluded pair to retain a positive score difference, or equality resolved in favor of the selected candidate. The criterion therefore compares the observed margin against the jointly attainable decrease of the selected score and increase of the excluded score.

Replacing U_i with the generic bound $w_i H_2(\eta_i)$ can only decrease ℓ_i, with u_i unchanged. Certification under the generic bound therefore implies certification under the sharp bound, but the converse need not hold. The sharp bound retains exactly the attainable divergence values, enabling a constructive counterexample for every non-invariant observed set.

Endpoint attainment also determines the exact range of each candidate rank. Rank 1 corresponds to the highest key, and $\#$ denotes set cardinality.

Corollary 4 (Attainable rank bounds)**.** *For candidate $i \in \mathcal{C}_K(x)$, the best and worst attainable ranks are*

$$r_i^{\text{best}} = 1 + \#\{j \in \mathcal{C}_K(x) \setminus \{i\} : k_j^- >_{\text{lex}} k_i^+\}, \tag{17}$$

$$r_i^{\text{worst}} = 1 + \#\{j \in \mathcal{C}_K(x) \setminus \{i\} : k_j^+ >_{\text{lex}} k_i^-\}. \tag{18}$$

Candidate i belongs to every Top-L set exactly when $r_i^{\text{worst}} \le L$, and to none exactly when $r_i^{\text{best}} > L$.

The superscripted quantities are scalar ranks; r_i without a superscript remains the normalized query tail.

Appendix A proves Corollary 4 and gives the corresponding worst-case additive selection rule (Corollary 5).

5.3 Certification without candidate-tail access

Using $a_i = q_i[y_i]$ from Section 4, define

$$z_i = \begin{cases} \dfrac{\min_{j \neq y_i} q_i[j]}{1 - a_i}, & a_i < 1, \\ 0, & a_i = 1, \end{cases} \tag{19}$$

where the minimum ranges over the original label coordinates. When $a_i < 1$, z_i is the smallest coordinate of the normalized query tail r_i. If $a_i < 1$ and $b_i < 1$, then $U_i = w_i \Phi_{\eta_i}(z_i)$; otherwise $U_i = 0$, and the value assigned to z_i is immaterial.

Given the query heads and candidate fold assignments $f(i)$, the quantities B_i and U_i depend only on the candidate label y_i and anchored mass b_i. The endpoints additionally require the fixed semantic score and mixture weights, without accessing the candidate-tail vector.

For each query head, storing its two smallest probability entries, including ties, and the index of a smallest entry suffices. The minimum excluding y_i is the smallest entry when its index differs from y_i, and the second smallest otherwise. Preprocessing requires $O(FC)$ time, followed by $O(K)$ for all endpoints; separate scans of candidate-associated query vectors require $O(KC)$.

Let A_- contain the L largest lower keys. Joint endpoint attainment implies that any invariant set equals A_-. Certification by Equation (14) therefore permits exact FPA selection without candidate tails; otherwise, the procedure evaluates S_i^0 from cached anchored posteriors. The masses b_i come from offline APM fitting. Ordering demonstrations by full scores requires evaluating those scores, including on certified queries. The empirical comparisons use the original reranker.

Selection costs and prompt order. Beyond query posteriors, endpoint storage and set certification require $O(K)$ space and time, respectively. Sorting to construct A_- requires $O(K \log K)$, and direct evaluation of all rank bounds requires $O(K^2)$. No target-LLM calls are required. Fixed semantic ordering, prompt template, and deterministic inference make invariant membership sufficient for identical prompt tokens and target outputs.

6 Experiments

We evaluate the analytical constructions, measure certificate coverage, and compare the predictive performance of the selection rules. Appendix C specifies the empirical protocols.

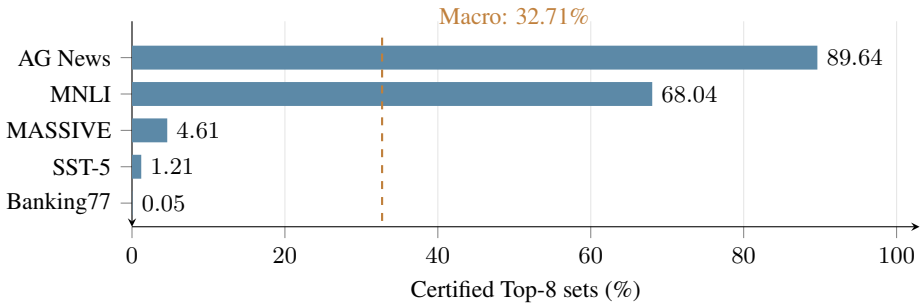

Figure 3: **Task variation in conditional Top-8 invariance.** Certification coverage across all five audited tasks. The dashed line is the equal-task/equal-seed mean of 32.71%; the query-weighted mean is 58.71%. Certified membership is invariant to candidate-tail reallocation at fixed query posteriors and candidate label masses.

6.1 Numerical verification

Verification protocol. We generated 2,500 query distributions paired with candidate label masses across $C \in \{2, 3, 5, 10, 77\}$, enumerated all tail vertices, and sampled five additional tails per pair. The evaluation also included 1,000 pools with $K = 30$ and $L = 8$, together with exhaustive endpoint enumeration for 200 smaller pools with $K = 6$ and $L = 2$. Appendix B specifies sampling and implementation details.

Attainability and membership guarantees. The decomposition and both interval endpoints agreed with direct JS calculations. Figure 2 confirms attainment of the derived maximum along the specified tail path. Of the 1,000 pools, 349 sets satisfied the invariance condition and retained their membership under 20 sampled tail configurations each. For all remaining 651 pools, Algorithm 1 constructed an admissible membership change. Exhaustive enumeration confirmed the rank bounds, all 3,000 size-two set certificates, and the additive minimax solutions. These checks confirm numerical agreement with the exact intervals, rank bounds, and constructive perturbations established analytically.

6.2 Certificate coverage on classification tasks

The classification audit comprised 100,455 query–seed rows from AG News, MNLI, MASSIVE, SST-5, and Banking77, with clean memory labels and eight demonstrations per prompt. Figure 3 reports the proportion of sets satisfying Theorem 3. Coverage was 32.71% under equal task and seed weighting and 58.71% under query-row weighting.

Certification rates varied substantially across tasks: 89.64% on AG News and 68.04% on MNLI, compared with less than 5% on each remaining task. The certificate compares score separation with candidate-specific interval widths for each query. The two aggregate rates differ because they weight tasks and queries differently.

On certified queries, candidate tails can be omitted while preserving FPA membership. The variation across tasks supports applying this reduction through the query-level certificate.

The lower-endpoint implementation check detected no membership violations. Separate APM retraining experiments, which jointly changed query posteriors, label masses, and tails, showed more set changes among initially uncertified queries (Appendix C.4).

6.3 Posterior information and prediction accuracy

The nine-task classification panel (suite A) used GTE-base-en-v1.5 retrieval, three-fold RoBERTa-base APMs, a shared Top-30 pool, and eight demonstrations per prompt. Three seeds varied APM training and fold assignment; target decoding was greedy. We compared semantic TopK, LABEL-MASS, and FPA on Qwen2.5-7B-Instruct and Mistral-7B-Instruct-v0.3. Invalid outputs were counted as errors, and macro accuracy assigned equal weight to each task. Appendix C.2 provides training settings and complete results.

Table 1: **Predictive performance and the label-mass control.** Nine-task macro accuracy (%); $\pm$ denotes sample SD across three seeds. TopK is deterministic. The final row gives paired FPA–Label-Mass differences and 95% confidence intervals in percentage points.

Selector	Qwen2.5-7B	Mistral-7B
Semantic TopK	84.81	82.56
LABEL-MASS	85.62 ± 0.10	86.66 ± 0.09
FPA	85.69 ± 0.09	87.49 ± 0.15
Difference	$+0.07$ $[-0.14, +0.27]$	$+0.83$ $[+0.62, +1.03]$

FPA achieved 85.69% and 87.49% accuracy on Qwen2.5-7B and Mistral-7B, respectively, compared with 84.81% and 82.56% for TopK (Table 1). On Mistral, FPA improved macro accuracy over LABEL-MASS by 0.83 percentage points (paired 95% CI: $[0.62, 1.03]$) and had higher means on eight of nine tasks. On Qwen, the difference was $+0.07$ points with a 95% CI of $[-0.14, +0.27]$ (Appendix C.2).

Among the eight Mistral tasks with higher FPA means, the increases ranged from 0.71 percentage points on QNLI to 1.14 on Subj. Banking77 was the exception, with a 0.27-point decrease.

The paired Mistral result supports the predictive benefit of the full scoring rule over direct query label evidence. The certificate addresses its information requirements: M_i retains the coarse divergence and entropy mixture, and the analysis isolates the tail residual with these terms fixed.

7 DISCUSSION AND CONCLUSION

Posterior dependence at the selection boundary. We characterize when candidate-tail geometry can change FPA selection. Exact attainable score intervals give a necessary and sufficient invariance criterion, constructive membership-changing perturbations, and exact rank bounds. Certification supports exact selection using candidate labels and anchored masses in place of candidate-tail vectors, with the original scores evaluated on uncertified queries.

Empirical findings. The five-task audit certified 32.71% of sets under equal task and seed weighting, with task rates from 0.05% to 89.64%. In the controlled nine-task comparison, FPA improved Mistral macro accuracy over LABEL-MASS by 0.83 percentage points, with a paired 95% CI of $[0.62, 1.03]$; the Qwen means remained close. These findings establish a predictive gain on Mistral and quantify where candidate-tail information can be omitted from selection.

Information requirements. The certificate supports a conditional implementation of FPA. Interval endpoints are computed from query posteriors, candidate labels and anchored masses, together with the fixed scoring inputs. The selector first tests the set with the largest lower keys. Certification permits membership recovery without candidate-tail access; otherwise, the original reranker evaluates the cached posteriors. This procedure recovers FPA membership in both cases and gives the certificate an operational role in determining which posterior information to access.

Scope and implications. Certification applies to candidate-tail reallocations at fixed query posteriors, label masses, and other scoring inputs in a finite label space. APM retraining changes the fixed inputs and requires recertification. With fixed prompt ordering and deterministic inference, certified membership preserves the target output. These results identify the posterior information sufficient for a selection decision and the tail perturbations that can change it.

ETHICS STATEMENT

This study analyzes demonstration selection using existing NLP benchmarks and synthetic probability distributions. Biases in the demonstration pool and auxiliary posterior model may persist even when selection is certified as invariant. Applications should assess performance disparities across relevant populations and protect sensitive information in demonstrations and prompts. Data and model reuse should follow applicable licenses and access conditions.

AI USE STATEMENT

Generative AI was used to assist with writing, code organization, reference lookup, and figure and table preparation. The submitting authors are responsible for the arguments, implementation, sources, and reported results.

REPRODUCIBILITY STATEMENT

Appendices A and B provide complete proofs and the numerical verification protocol. The supplementary code implements the score bounds, invariance certificates, and numerical checks. Accompanying CSV and JSON files contain the displayed summary data, and Appendix C specifies the empirical evaluation protocols.

REFERENCES

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pp. 38–45, 2020. doi: 10.18653/v1/2020.nlp4convai-1.5. URL https://aclanthology.org/2020.nlp4convai-1.5/.

André Chassein, Marc Goerigk, Adam Kasperski, and Paweł Zieliński. On recoverable and two-stage robust selection problems with budgeted uncertainty. *European Journal of Operational Research*, 265(2):423–436, 2018. doi: 10.1016/j.ejor.2017.08.013.

Haoyang Chen, Richong Zhang, and Junfan Chen. Rethinking label consistency of in-context learning: An implicit transductive label propagation perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(36):30226–30234, 2026a. doi: 10.1609/aaai.v40i36.40273. URL https://ojs.aaai.org/index.php/AAAI/article/view/40273.

Lihu Chen, Gerard de Melo, Fabian M. Suchanek, and Gaël Varoquaux. Query-level uncertainty in large language models. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3a07c3a67cfe50d3236b71fb674c7f30-Abstract-Conference.html.

Su Feng, Boris Glavic, and Oliver Kennedy. Efficient approximation of certain and possible answers for ranking and window queries over uncertain data. *Proceedings of the VLDB Endowment*, 16 (6):1346–1358, 2023. doi: 10.14778/3583140.3583151. URL https://www.vldb.org/pvldb/vol16/p1346-feng.pdf.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist,

Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4277–4302, 2023. doi: 10.18653/v1/2023.acl-long.235. URL https://aclanthology.org/2023.acl-long.235/.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 33–38, 2010. URL https://aclanthology.org/S10-1006/.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168–177, 2004. doi: 10.1145/1014052.1014073. URL https://www.cs.uic.edu/~liub/publications/kdd04-revSummary.pdf.

Dan Iter, Reid Pryzant, Ruochen Xu, Shuohang Wang, Yang Liu, Yichong Xu, and Chenguang Zhu. In-context demonstration selection with cross entropy difference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1150–1162, 2023. doi: 10.18653/v1/2023.findings-emnlp.81. URL https://aclanthology.org/2023.findings-emnlp.81/.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. A challenge dataset and effective models for aspect-based sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6280–6285, 2019. doi: 10.18653/v1/D19-1654. URL https://aclanthology.org/D19-1654/.

Soichiro Kumano, Hiroshi Kera, and Toshihiko Yamasaki. Adversarially pretrained transformers may be universally robust in-context learners. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ca0a348ce13a8cabbee9031112743776-Abstract-Conference.html.

Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115. URL https://doi.org/10.1109/18.61115.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 100–114, 2022. doi: 10.18653/v1/2022.deelio-1.10. URL https://aclanthology.org/2022.deelio-1.10/.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. URL https://arxiv.org/abs/1907.11692.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, 2022. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556/.

Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pp. 271–278, Barcelona, Spain, 2004. doi: 10.3115/1218955.1218990. URL https://aclanthology.org/P04-1035/.

Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. Revisiting demonstration selection strategies in in-context learning. In *Proceedings of the*

62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 9090–9101, 2024. doi: 10.18653/v1/2024.acl-long.492. URL https://aclanthology.org/2024.acl-long.492/.

Christopher Ré, Nilesh Dalvi, and Dan Suciu. Efficient top-k query evaluation on probabilistic data. Technical Report 2006-06-05, University of Washington, 2006. URL https://cs.stanford.edu/~chrismre/papers/Multisimulation_TR.pdf. Extended version.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2655–2671, 2022. doi: 10.18653/v1/2022.naacl-main.191. URL https://aclanthology.org/2022.naacl-main.191/.

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, 2013. URL https://aclanthology.org/D13-1170/.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, 2018. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019. URL https://iclr.cc/virtual/2019/poster/845.

Chonghai Wang, Li-Yan Yuan, and Jia-Huai You. On the semantics of top-k ranking for objects with uncertain data. *Computers & Mathematics with Applications*, 62(7):2812–2823, 2011. doi: 10.1016/j.camwa.2011.07.047.

Yingjie Wang, Yutian Zhou, Shi Fu, Yuzhu Chen, Yongcheng Jing, Leszek Rutkowski, and Dacheng Tao. Towards a theoretical understanding of in-context learning: Stability and non-I.I.D generalisation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/91b75715a2ae8f2542b78be31941eac3-Abstract-Conference.html.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, 2018. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101/.

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1423–1436, 2023. doi: 10.18653/v1/2023.acl-long.79. URL https://aclanthology.org/2023.acl-long.79/.

Qi Zhang, Zhiqing Xiao, Ruixuan Xiao, Lirong Gao, and Junbo Zhao. D.Va: Validate your demonstration first before you use it. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2580–2594, 2025. doi: 10.18653/v1/2025.acl-long.129. URL https://aclanthology.org/2025.acl-long.129/.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, pp. 649–657, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html.

A Complete proofs and boundary conventions

A.1 Definitions and the decomposition

The proofs follow the notation of Sections 3–5. For a single candidate, q, p denote q_i, v_i with the index suppressed, while p_i denotes the raw APM estimate. For $r, s \in \Delta^{d-1}$, weighted JS is defined by Equation (5). We use $0 \log(0/v) = 0$ for $v \geq 0$, interpreting the zero-over-zero term by continuity. For $0 < \eta < 1$, a positive coordinate in either KL term implies a positive mixture denominator, and both weighted terms are finite. The continuous extension at $\eta = 0$ or 1 is identically zero. The entropy representation is

$$\mathrm{JS}_\eta(r, s) = H(\eta r + (1-\eta)s) - \eta H(r) - (1-\eta)H(s). \tag{20}$$

Write $t_q = 1 - a$ and $t_p = 1 - b$. If $t_q + t_p > 0$, the mixture of q and p has mass at the candidate label $(a+b)/2$, tail mass $w = (t_q + t_p)/2$, and normalized tail $m = \eta r + (1-\eta)s$ with $\eta = t_q/(t_q + t_p)$. KL contributions multiplied by zero tail mass are defined as zero by continuous extension. Expanding the first KL term gives

$$\mathrm{KL}\left(q \middle\| \frac{q+p}{2}\right) = a \log_2 \frac{2a}{a+b} + t_q \log_2 \frac{t_q}{w} + t_q \,\mathrm{KL}(r\|m). \tag{21}$$

Expanding the term for p replaces a, t_q, r with b, t_p, s. Averaging the first two terms of the two expansions gives $B(a, b)$; the remaining contribution is

$$\frac{t_q}{2}\,\mathrm{KL}(r\|m) + \frac{t_p}{2}\,\mathrm{KL}(s\|m) = w\big[\eta\,\mathrm{KL}(r\|m) + (1-\eta)\,\mathrm{KL}(s\|m)\big]. \tag{22}$$

This proves Equation (7). If $t_q = t_p = 0$, both posteriors are the point mass on the candidate label and the divergence is zero. If exactly one tail mass vanishes, $\eta \in \{0, 1\}$ and the residual is zero regardless of the unused normalized tail. The decomposition therefore extends to every boundary.

The generic bound $\mathrm{JS}_\eta(r, s) \leq H_2(\eta)$ follows by interpreting weighted JS as the mutual information between the mixture-component indicator and a sample from the mixture, or equivalently by applying the pointwise mixture inequalities to the KL terms. For $0 < \eta < 1$, disjoint supports of r, s attain this bound. A fixed query tail with full support precludes disjointness; Theorem 2 characterizes the exact attainable maximum under that constraint.

A.2 Sharp candidate-tail interval

The KL representation implies nonnegativity, and the matched tail $s = r$ attains the lower endpoint zero. The residual is identically zero for $w = 0$ or $\eta \in \{0, 1\}$. It remains to consider $w > 0$ and $0 < \eta < 1$.

For fixed r, convexity of $s \mapsto \mathrm{JS}_\eta(r, s)$ follows from the KL representation. Let $s_\theta = \theta s^{(1)} + (1-\theta)s^{(2)}$ and $m^{(j)} = \eta r + (1-\eta)s^{(j)}$. The mixture satisfies $m_\theta = \theta m^{(1)} + (1-\theta)m^{(2)}$. Convexity of KL in its second argument bounds $\mathrm{KL}(r\|m_\theta)$ by the corresponding convex combination; joint convexity gives the analogous bound for $\mathrm{KL}(s_\theta\|m_\theta)$. Their nonnegative weighted sum establishes the result.

Every simplex point has the form $s = \sum_j s_j e_j$. Hence

$$\mathrm{JS}_\eta(r, s) \leq \sum_j s_j\,\mathrm{JS}_\eta(r, e_j) \leq \max_j \mathrm{JS}_\eta(r, e_j). \tag{23}$$

For $z = r_j$, the mixture coordinate at j is $m_z = 1 - \eta + \eta z$, and every other coordinate is ηr_k. Using Equation (20),

$$\mathrm{JS}_\eta(r, e_j) = -m_z \log_2 m_z - \sum_{k \neq j} \eta r_k \log_2(\eta r_k) + \eta \sum_k r_k \log_2 r_k \tag{24}$$

$$= -m_z \log_2 m_z - \eta(1-z)\log_2 \eta + \eta z \log_2 z = \Phi_\eta(z). \tag{25}$$

For $z > 0$, differentiation gives Equation (10); the ratio inside the logarithm is strictly below one. Therefore Φ_η decreases on $(0, 1]$, and continuity at zero extends this ordering to $[0, 1]$. A vertex with $r_j = r_{\min}$ attains the maximum. Multiplication by w gives $U(q, b)$.

Continuity establishes attainment of the intermediate values. The path $s(t) = (1-t)r + te_{j^*}$ lies in the closed simplex for $t \in [0,1]$, with weighted JS equal to zero and U/w at its endpoints. The intermediate value theorem then gives the entire interval. For $C = 2$, the sole competing-label coordinate gives $r = s = (1)$ and $\Phi_\eta(1) = 0$, completing the proof.

If $0 < \eta < 1$, the sharp and generic bounds coincide when $r_{\min} = 0$, since $\Phi_\eta(0) = H_2(\eta)$. They differ strictly when $r_{\min} > 0$. For a uniform tail on $d = C - 1$ classes, the exact residual bound is $w\Phi_\eta(1/d)$.

A.3 Score intervals and universal membership

Substitution of Lemma 1 into Equation (3) gives $S_i = M_i - \kappa G_i$ with $\kappa \geq 0$. If $\kappa > 0$, the continuous image of the exact residual interval is precisely $[M_i - \kappa U_i, M_i]$. If $\kappa = 0$, both endpoints coincide. The Cartesian product in Equation (4) permits simultaneous realization of all individually attainable scores. The induced joint score set is therefore exactly $\prod_i [\ell_i, u_i]$.

Distinct fixed priorities induce a strict total order on candidate keys. Under Equation (14), each selected key is bounded below by its lower key, and each excluded key is bounded above by its upper key. Consequently, every member of A precedes every nonmember for all admissible tails, establishing invariance of the Top-L set.

For necessity, suppose the condition is violated. Let $i \in A$ minimize the lower key and $j \notin A$ maximize the upper key. The candidates are distinct, so their endpoint keys differ and satisfy $(u_j, -\pi_j) >_{\text{lex}} (\ell_i, -\pi_i)$. Theorem 2 provides posteriors attaining the lower score of i and the upper score of j. The product uncertainty set permits simultaneous attainment, with every other candidate retaining an arbitrary admissible tail, including its nominal tail. Candidate j then precedes i, contradicting the defining order of a Top-L set. Hence A is not invariant. Membership is unique for $L = 0$ and $L = K$.

A.4 Certificate algorithm

Algorithm 1 Top-L invariance certification and constructive counterexample

Require: Fixed $q_i, \bar{q}, y_i, b_i, R_i, c, \lambda, \pi_i$ for all $i \in \mathcal{C}_K(x)$.
Require: Observed anchored posteriors $\widetilde{p}_i$ and their Top-L set A; $0 < L < K$.
1: **for** each candidate $i \in \mathcal{C}_K(x)$ **do**
2: Compute B_i, U_i, ℓ_i, u_i using Section 4.
3: **end for**
4: $i^- \leftarrow \arg\min_{i \in A}(\ell_i, -\pi_i)$.
5: $j^+ \leftarrow \arg\max_{j \in \mathcal{C}_K(x) \setminus A}(u_j, -\pi_j)$.
6: **if** $(\ell_{i^-}, -\pi_{i^-}) >_{\text{lex}} (u_{j^+}, -\pi_{j^+})$ **then**
7: **return** Invariant.
8: **else**
9: Initialize $v_i \leftarrow \widetilde{p}_i$ for every candidate i.
10: Replace the normalized tail of v_{i^-} by a least-probable query-tail vertex.
11: Replace the normalized tail of v_{j^+} by its matching query tail r_{j^+}.
12: Preserve each label mass b_i and scale each replacement tail by $1 - b_i$.
13: **return v** as a Counterexample.
14: **end if**

Certification requires only A, the interval endpoints, and the fixed priorities. Counterexample construction additionally uses the observed posteriors to preserve all other candidates.

A.5 Rank ranges and minimax selection

The configuration with candidate i at its upper endpoint and all competitors at their lower endpoints attains the best rank of i. A competitor precedes i in this configuration exactly when $k_j^- >_{\text{lex}} k_i^+$, which also implies precedence in every admissible configuration. This establishes the best attainable rank. The configuration with i at its lower endpoint and all competitors at their upper endpoints

establishes the worst attainable rank analogously. Membership in every Top-L set is equivalent to a worst rank of at most L; exclusion from every such set is equivalent to a best rank exceeding L.

Corollary 5 (Worst-case additive selection)**.** *Let A_- be the set of L candidates with the largest lower keys. Then*

$$A_- \in \arg \max_{\substack{A \subseteq \mathcal{C}_K(x) \\ |A|=L}} \min_{\mathbf{v} \in \mathcal{U}} \sum_{i \in A} S_i(v_i). \tag{26}$$

The priorities specify the choice among candidates with equal lower scores. If a universally invariant size-L set exists, it is A_-.

This objective maximizes the worst-case sum of reranker scores; target-LLM interactions among demonstrations are outside its additive formulation.

For any fixed set A, compactness, continuity, and simultaneous endpoint attainment over $\mathcal{U}$ imply

$$\min_{\mathbf{v} \in \mathcal{U}} \sum_{i \in A} S_i(v_i) = \sum_{i \in A} \min_{v_i \in \mathcal{U}_i} S_i(v_i) = \sum_{i \in A} \ell_i. \tag{27}$$

The maximizing sum contains the L largest lower scores, with boundary ties resolved by the fixed priorities. Any universally invariant set must also be selected when *all* candidate scores attain their lower endpoints. The unique set in that configuration is A_-, which is therefore the only possible invariant set. Theorem 3 characterizes whether it is invariant over the full uncertainty set.

A.6 Prompt order and numerical checks

Invariant membership preserves prompt tokens under fixed demonstration ordering and formatting. With unchanged inference settings, a deterministic decoder consequently produces the same output. Ordering by variable candidate scores can alter the prompt despite invariant membership.

The proofs use exact arithmetic and the specified priorities for ties. The numerical implementation evaluates these comparisons in floating point.

B Reproducible numerical validation

The numerical verification code is available at `https://anonymous.4open.science/r/posterior-geometry-5376`. The checks use synthetic posteriors and require no target-model calls.

Table 2: Numerical verification against analytical and exhaustive references. All comparisons satisfy a tolerance of 10^{-12}. The maximum error in the decomposition and sharp endpoints is 1.12×10^{-15} (rounded upward). The checks assess numerical implementation of the theoretical results.

Check	Evaluations	Reference
Tail decomposition	12,500	Direct JS calculation
Maximum over tail vertices	46,000	Theorem 2
Certified-set perturbations	6,980	Unchanged membership
Constructive counterexamples	651	Changed membership
All size-two set certificates	3,000	Exhaustive endpoint rankings
Best and worst ranks	200 pools	Exhaustive endpoint rankings
Additive minimax objective	200 pools	Exhaustive set optimization
Exact ties / degenerate tails	4 / 5	Closed-form boundary cases

Posterior-pair checks. For each $C \in \{2, 3, 5, 10, 77\}$, we generated 500 query distributions with alternating symmetric Dirichlet concentration parameters 0.3 and 2.0. Candidate labels were sampled uniformly, with masses b uniform on $[0, 1]$. Enumeration of all $C - 1$ candidate-tail vertices gave 46,000 endpoint evaluations. Five additional tails per pair were sampled from a symmetric Dirichlet distribution with concentration 0.6. The checks evaluated the decomposition, both sharp endpoints, and inclusion of sampled residuals in the interval (Table 2).

Retrieval-pool checks. For each $C \in \{3, 5, 10, 77\}$, we generated 250 pools with $K = 30$ and $L = 8$. Query distributions came from three fixed synthetic fold heads with alternating symmetric Dirichlet parameters 0.1 and 1.0. Raw candidate posteriors used parameter 0.7 and the anchor in Equation (2). Candidate labels were uniform, semantic scores were uniform on $[-0.2, 1]$, and $\lambda = 0.5$. Of the 1,000 pools, 496 used the entropy-defined concentration; the others fixed c at 0 or 0.99 in equal proportions. These settings tested the theorem across fixed concentration weights. The classification comparisons used the FPA rule defined in Section 3.

Under this sampling design, 349 of the 1,000 nominal sets satisfied the invariance condition. For each certified set, 20 independent configurations with symmetric Dirichlet tail parameter 0.5 gave 6,980 evaluations with no membership changes. For the remaining 651 sets, Algorithm 1 selected the two candidates from the violated endpoint comparison and assigned their extremal tails, preserving every other posterior. All 651 constructions changed membership.

Exhaustive interval checks. We generated 200 problems with $K = 6$ and $L = 2$, with lower endpoints uniform on $[-1, 1]$, interval widths uniform on $[0, 0.5]$, and priorities sampled as independent permutations. Enumeration of all 64 endpoint configurations per pool produced 12,800 complete rankings. These provided exhaustive references for the rank bounds in Corollary 4, all 15 size-two invariance certificates per pool, and the worst-case additive objective in Corollary 5. No discrepancies occurred in the 3,000 set tests or 200 rank/minimax problems. The interval tests checked the ordering criterion, and the posterior tests checked its realization through JS score intervals and extremal posteriors.

Boundaries and curve generation. The script separately checks four exact-tie examples, including the zero-margin case in Section 5, and five cases with zero tail mass or a binary label space. Figure 2 uses $q = (0.2, 0.56, 0.20, 0.04)$ and $b = 0.8$, giving $a = 0.2$, $r = (0.7, 0.25, 0.05)$, $w = 0.5$, and $\eta = 0.8$. The curves evaluate 41 points on the path $s(t) = (1 - t)r + t(0, 0, 1)$, $t \in [0, 1]$. The three plotted series are the direct JS residual, the sharp bound $0.5\Phi_{0.8}(0.05)$, and the generic bound $0.5H_2(0.8)$.

Compact calculations. The compact implementation stores the two smallest probabilities of each query head with their indices. The minimum excluding the candidate label is obtained from the smallest entry with a different label index, using the second entry when necessary. Equal probabilities are retained as distinct entries. Across 500 independently generated pools, it compares 15,000 interval calculations and 450,000 pairwise instances of Equation (16), including tied-minimum and zero-query-tail cases. All comparisons satisfied the numerical tolerance.

C Empirical protocols and complete results

C.1 Evaluation panels

The evaluation uses four panels. Suite A provides controlled classification comparisons, and suite B includes additional task structures. The posterior audit measures certificate coverage, and a separate panel compares selectors using target-LLM scores. Each macro average refers to the task set and protocol of its panel.

Table 3: Empirical panels, task composition, and evaluation objectives. Macro averages are computed separately for each panel. Appendix B specifies numerical verification.

Panel	Tasks	Role
Suite A	SST-2, SST-5, CR, Subj, AG News, MNLI, QNLI, Banking77, MASSIVE	Classification accuracy and baseline controls.
Suite B	SST-5, AG News, MNLI, QNLI, Banking77, MASSIVE, SemEval-D, MAMS, FEVER-NLI	Expanded task structures; separate summary tables.
Certificate audit	AG News, MNLI, MASSIVE, SST-5, Banking77	Conditional coverage and endpoint implementation check.
Target-aware panel	Five fixed panels of 512 queries each	Separate comparisons with TopK+ConE and D.Va.

C.2 SUITE A: DATA, MODELS, AND CONTROLS

The tasks cover sentiment and subjectivity (Socher et al., 2013; Hu & Liu, 2004; Pang & Lee, 2004), topic classification (Zhang et al., 2015), inference (Williams et al., 2018; Wang et al., 2019), and intent detection (Casanueva et al., 2020; FitzGerald et al., 2023). SST-2 uses the 6,920-example sentence-level training split from SetFit/sst2; MNLI and QNLI use labeled validation sets for evaluation; MASSIVE uses en-US. These splits define the evaluation protocol and differ in part from official leaderboard test sets.

Semantic retrieval used normalized GTE-base-en-v1.5 embeddings, with eight demonstrations selected from a Top-30 pool. Three-fold cross-fitting trained RoBERTa-base APMs (Liu et al., 2019) for three epochs per fold, with learning rate 2×10^{-5}, weight decay 0.01, warmup ratio 0.06, maximum input length 256, and effective batch size 32. Three seeds varied APM training and fold assignment. Qwen2.5-7B-Instruct and Mistral-7B-Instruct-v0.3 used greedy decoding with at most 16 generated tokens, and invalid label outputs were counted as errors. The scoring mixture and anchor were developed on CR, SST-2, and Subj and fixed for the controlled runs.

C.2.1 CONTROL DEFINITIONS

LABEL-MASS combines semantic relevance with the query probability $\bar{q}[y_i]$, as defined in Section 3. Hard label match selects candidates whose observed labels equal $\arg\max_y \bar{q}[y]$, ordered by semantic relevance, and fills any remaining positions by semantic rank. Predicted-label filtering instead matches the most probable query label to the most probable candidate label, with the same fallback. Posterior confidence ranks candidates by $p_i[\arg\max_y \bar{q}[y]]$, breaking ties by semantic score. OOF distribution-only alignment uses $\lambda R_i + (1-\lambda)(1-\mathrm{JS}(q_i, p_i))$ without the anchor or entropy mixture. FPA evaluates Equation (3) at $v = \widetilde{p}_i$.

Statistics. Suite A reports macro means with equal task weights and sample standard deviations across three seeds. Paired 95% confidence intervals use the 2.5th and 97.5th percentiles of 100,000 bootstrap draws over seeds and test IDs, with tasks fixed and equally weighted. Differences are computed from paired outcomes before rounding. Reported aggregates and contrasts can differ slightly from calculations using rounded task means. For Mistral, the FPA–LABEL-MASS contrast is $+0.83$ percentage points with a 95% CI of $[0.62, 1.03]$; the Qwen contrast is $+0.07$ with a CI of $[-0.14, +0.27]$. The following tables report the task-level results.

Table 4: Suite A: Qwen2.5-7B accuracy (%). TopK is deterministic; other entries give mean ± sample SD over three selector seeds. The two blocks separate the primary comparison from auxiliary controls (Appendix C.2.1).

Task	TopK	Label-Mass	FPA
SST-2	95.72	96.34 ± 0.14	96.43 ± 0.20
SST-5	56.83	60.36 ± 0.39	60.94 ± 0.51
CR	94.95	94.41 ± 0.46	94.86 ± 0.41
Subj	85.95	86.55 ± 0.05	86.48 ± 0.16
AG News	91.89	93.91 ± 0.06	93.93 ± 0.05
MNLI	80.41	77.62 ± 0.04	77.85 ± 0.10
QNLI	84.53	87.06 ± 0.08	87.02 ± 0.16
Banking77	88.96	88.74 ± 0.25	88.04 ± 0.33
MASSIVE	84.03	85.56 ± 0.22	85.65 ± 0.45
Macro	84.81	85.62 ± 0.10	85.69 ± 0.09

Task	Hard match	Pred. filter	Post. conf.	OOF only
SST-2	96.30 ± 0.11	96.27 ± 0.15	95.94 ± 0.40	96.30 ± 0.27
SST-5	59.70 ± 0.45	57.90 ± 0.13	58.79 ± 0.32	58.34 ± 0.59
CR	94.33 ± 0.41	95.12 ± 0.15	93.88 ± 0.46	94.68 ± 0.53
Subj	86.55 ± 0.05	86.45 ± 0.18	88.73 ± 0.19	86.57 ± 0.29
AG News	93.85 ± 0.06	93.11 ± 0.09	93.90 ± 0.04	93.36 ± 0.05
MNLI	77.63 ± 0.03	79.82 ± 0.05	78.83 ± 0.07	79.28 ± 0.02
QNLI	87.04 ± 0.05	86.58 ± 0.14	87.17 ± 0.16	86.72 ± 0.16
Banking77	88.21 ± 0.06	88.48 ± 0.38	87.98 ± 0.46	89.07 ± 0.18
MASSIVE	85.57 ± 0.06	84.49 ± 0.27	84.47 ± 0.43	84.94 ± 0.12
Macro	85.46 ± 0.09	85.36 ± 0.04	85.52 ± 0.10	85.47 ± 0.08

Table 5: Suite A: Mistral-7B accuracy (%). TopK is deterministic; other entries give mean ± sample SD over three selector seeds. The two blocks separate the primary comparison from auxiliary controls (Appendix C.2.1).

Task	TopK	Label-Mass	FPA
SST-2	95.17	96.27 ± 0.24	97.25 ± 0.12
SST-5	51.40	59.58 ± 0.22	60.43 ± 0.10
CR	94.68	93.79 ± 0.15	94.88 ± 0.30
Subj	86.20	92.83 ± 0.03	93.97 ± 0.21
AG News	90.61	94.00 ± 0.04	94.90 ± 0.12
MNLI	73.91	81.89 ± 0.03	82.88 ± 0.32
QNLI	74.89	85.00 ± 0.12	85.71 ± 0.21
Banking77	91.66	90.93 ± 0.15	90.66 ± 0.02
MASSIVE	84.57	85.66 ± 0.19	86.71 ± 0.05
Macro	82.56	86.66 ± 0.09	87.49 ± 0.15

Task	Hard match	Pred. filter	Post. conf.	OOF only
SST-2	96.17 ± 0.27	95.79 ± 0.28	95.99 ± 0.10	95.97 ± 0.14
SST-5	58.55 ± 0.36	52.35 ± 0.82	52.67 ± 1.50	51.42 ± 0.61
CR	93.71 ± 0.15	94.59 ± 0.41	93.53 ± 0.31	93.71 ± 0.31
Subj	92.80 ± 0.05	91.13 ± 0.18	92.68 ± 0.45	91.72 ± 0.24
AG News	94.03 ± 0.05	92.07 ± 0.05	93.18 ± 0.06	92.32 ± 0.11
MNLI	81.86 ± 0.03	78.70 ± 0.10	80.73 ± 0.08	79.71 ± 0.13
QNLI	84.99 ± 0.11	79.96 ± 0.35	84.27 ± 0.27	80.87 ± 0.28
Banking77	89.70 ± 0.21	90.92 ± 0.34	89.10 ± 0.59	91.71 ± 0.07
MASSIVE	85.84 ± 0.07	85.94 ± 0.09	85.16 ± 0.20	85.63 ± 0.35
Macro	86.41 ± 0.01	84.61 ± 0.02	85.26 ± 0.25	84.78 ± 0.19

C.3 SUITE B AND COMPARISONS USING THE TARGET MODEL

Suite B shares six tasks with suite A and adds a 10-class view of SemEval-2010 Task 8 with relation direction removed (Hendrickx et al., 2010), MAMS aspect sentiment (Jiang et al., 2019), and FEVER-NLI with fixed evidence (Thorne et al., 2018). SemEval-D is evaluated by accuracy in the direction-removed label space, which differs from the official task and metric. FEVER-NLI evaluates claim classification with supplied evidence, excluding evidence retrieval.

This panel used three-fold RoBERTa-base, three selector seeds, $K = 30$, $L = 8$, and $\lambda = 0.5$. The parser selected the earliest uniquely recognized label and counted invalid outputs as errors. A sensitivity check required exactly one distinct response label. Comparisons retained each method's original prompt ordering rule and therefore include both membership and ordering effects.

Table 6: Suite B: mean accuracy (%) across three selector seeds. The task panel differs from suite A; its macro is computed separately. All task-level comparisons are retained.

	Qwen2.5-7B			Mistral-7B		
Task	TopK	Label-Mass	FPA	TopK	Label-Mass	FPA
SST-5	56.83	60.36	60.94	51.40	59.58	60.43
AG News	91.89	93.91	93.93	90.61	94.00	94.90
MNLI	80.41	77.62	77.85	73.91	81.89	82.88
QNLI	84.53	87.06	87.02	74.89	85.00	85.71
Banking77	88.96	88.74	88.04	91.66	90.93	90.66
MASSIVE	84.03	85.56	85.65	84.57	85.66	86.71
SemEval-D	59.07	77.24	77.73	58.63	73.35	73.33
MAMS	52.25	73.73	73.25	52.32	59.78	59.76
FEVER-NLI	66.64	68.64	68.49	63.63	66.90	66.76
Macro	73.85	79.21	79.21	71.29	77.45	77.90

The macro differences between FPA and TopK were approximately $+6.61$ percentage points for Mistral and $+5.37$ for Qwen. The corresponding differences from LABEL-MASS were approximately $+0.45$ and 0.00 points. The Mistral comparisons are point estimates from the task-level means; for Qwen, the paired 95% confidence intervals were $[4.93, 5.81]$ against TopK and $[-0.20, +0.21]$ against LABEL-MASS. Thus, FPA had higher macro accuracy than TopK on both models, a positive mean difference from LABEL-MASS on Mistral, and a similar mean to LABEL-MASS on Qwen. On Mistral, FPA exceeded LABEL-MASS on five of the six shared tasks, but not on Banking77 or the three additional tasks.

A separate five-task panel compared FPA with TopK+ConE (Peng et al., 2024) and D.Va (Zhang et al., 2025). ConE scored the semantic Top-30 using the target model's conditional input likelihood. The D.Va implementation reserved a nearest labeled memory example for validation and combined its validation loss with the query-to-validation likelihood discrepancy at a mixing weight of 0.6. Table 7 reports this panel separately.

Table 7: Target-aware comparison on five fixed 512-query panels. Each target block reports accuracy (%) and the paired FPA–comparator contrast in percentage points. Brackets denote 95% confidence intervals. These panels are separate from suites A and B.

Selector	Accuracy (%)	FPA–comparator [95% CI]
Qwen2.5-7B		
FPA	75.23	—
TopK+ConE	71.17	$+4.06$ $[+2.63, +5.49]$
D.Va	71.05	$+4.18$ $[+2.72, +5.63]$
Mistral-7B		
FPA	74.51	—
TopK+ConE	67.30	$+7.20$ $[+5.70, +8.71]$
D.Va	67.58	$+6.93$ $[+5.39, +8.48]$

C.4 CERTIFICATE AUDIT AND APM STRESS

The five-task audit applied Theorem 3 to 100,455 query–seed rows. Averaging coverage within each task–seed cell and then equally across cells gave 32.71%; query-row weighting gave 58.71%. The lower-endpoint implementation check evaluated one configuration per set and detected no membership violations, consistent with the guarantee over all admissible tails in the theorem.

The stress experiments used separate APM seeds, with undertraining and symmetric corruption of 20% of APM training labels while preserving clean demonstration-memory labels. Set changes were grouped by the original certificate; intervals used 100,000 hierarchical bootstrap draws with equal task and seed weights. Table 8 reports selection changes under APM retraining, which jointly varies query posteriors and anchored masses.

Table 8: Set changes after APM retraining, grouped by the original certificate. Rates (%) are equal-task/equal-seed means; Δ is uncertified minus certified, in percentage points. Retraining also changes query posteriors and label masses, so these interventions lie outside the theorem's uncertainty set.

APM intervention	Certified	Uncertified	Δ [95% CI]
Undertraining	67.52	88.80	21.28 [19.17, 23.34]
20% label noise	21.13	81.92	60.78 [57.20, 64.05]