LABEL ALIGNMENT DIVERGENCE RERANKING FOR IN-CONTEXT DEMONSTRATION SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Semantic retrieval identifies demonstrations that resemble a query, but input similarity need not capture their relationship to the task's label distinctions. We study how auxiliary task supervision can improve demonstration selection for classification in-context learning. We propose Label-Alignment Divergence Reranking, which combines semantic similarity with Jensen–Shannon compatibility between label distributions estimated by an auxiliary predictor. The method reranks a fixed semantic candidate pool without scoring candidates with the target language model. Because candidate labels are observed whereas the query label is unknown, we also introduce a gold-anchored extension that anchors candidate posteriors to their observed labels and balances distribution compatibility with label matching according to query posterior concentration. In a controlled evaluation across nine classification tasks with cross-fitted posteriors and three selection seeds, the proposed method reaches macro accuracies of 85.69% and 87.49% on Qwen and Mistral, compared with 84.81% and 82.56% for semantic TopK. Comparisons with direct label-probability matching show that the incremental benefit of the proposed method varies across target models. Selection diagnostics characterize changes within semantic neighborhoods, while reliability analyses identify sensitivity to posterior quality and domain shift. These findings support auxiliary label information as a useful complement to semantic relevance in classification ICL. The code is released here: `https://anonymous.4open.science/r/L2D-401B`.

1 INTRODUCTION

In-context learning (ICL) uses a small set of input–label examples to guide a large language model (LLM) at inference time (Brown et al., 2020). Because the selected examples and their order can change predictions (Liu et al., 2022; Rubin et al., 2022; Lu et al., 2022), demonstration selection is an important part of an ICL system. Semantic retrieval provides a strong starting point: it identifies inputs related to the query without requiring the target LLM to evaluate every candidate.

However, input relevance and compatibility with the task's label distinctions are not the same. Negation, contrast, and fine-grained intent differences can change a label while leaving inputs topically similar. The issue is not that examples with different labels are inherently unhelpful, but that semantic similarity alone does not describe their relationship to the query's task decision. We refer to this gap as *output-space incompatibility* (Figure 1). It motivates a simple design principle: retain semantic retrieval to identify relevant candidates, then use task predictions to refine which candidates become demonstrations.

Prior studies show that label mappings, label words, and context label distributions affect ICL behavior (Yoo et al., 2022; Wang et al., 2023; Xu et al., 2024). Their effects depend on the setting: some utility can survive randomized labels (Min et al., 2022), and symbol-tuned models can use arbitrary label symbols (Wei et al., 2023). For demonstration selection, these findings motivate considering how examples relate in the label space, alongside their input similarity. We study a practical way to obtain this signal from an auxiliary predictor rather than from repeated candidate evaluations by the target LLM.

We consider a labeled demonstration memory and a frozen target LLM. The prompt budget limits the examples shown to the target, not the labeled data available to train an auxiliary predictor. The

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.

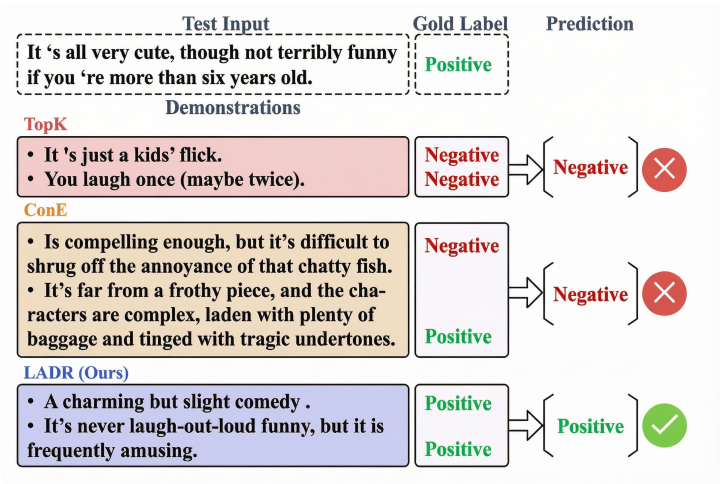


Figure 1: Two-shot illustration of different SST-2 selections. The displayed query label is for interpretation, not an input to selection. Quantitative comparisons use eight demonstrations.

question is how to turn that supervision into a useful selection signal without target-LLM candidate scoring.

We propose **L**abel-**A**lignment **D**ivergence **R**eranking (**LADR**). An auxiliary posterior model (APM) estimates a label distribution for the query and each candidate. LADR combines semantic similarity with Jensen–Shannon (JS) compatibility between these distributions to rerank a fixed semantic pool. Candidate representations and posteriors can be cached, and the target LLM is invoked only after selection. We also introduce **Gold-Anchored LADR (LADR-GA)**[1], which addresses an information asymmetry: the query label is unknown, but candidate labels are observed. It anchors candidate predictions to those labels and mixes direct label matching with distribution compatibility according to query-posterior concentration.

Our evaluation connects effectiveness with selection behavior. A broad benchmark compares LADR with existing selectors across nine classification tasks. A separate controlled study uses cross-fitted candidate posteriors, three selection seeds, and two target LLMs to compare posterior alignment with direct label-based strategies. We organize the evidence around three questions:

RQ1: Does label alignment improve demonstration selection across the evaluated tasks and target LLMs?

RQ2: How does alignment change the selected examples, and how does it compare with simpler uses of auxiliary predictions and candidate labels?

RQ3: How do posterior quality, label mappings, and domain shift affect its effectiveness?

The main contributions are:

1. We introduce LADR, a reranker combining semantic relevance with auxiliary label-distribution compatibility. LADR-GA explicitly incorporates observed candidate labels. Both refine a semantic pool without candidate scoring by the target LLM.
2. We evaluate the approach on nine classification tasks through a broad benchmark and a separate controlled study with cross-fitted posteriors and multiple selection seeds. The results establish aggregate improvements over semantic TopK and characterize the additional value of LADR-GA relative to direct label-based selection across target models.
3. We characterize selection behavior and practical limitations. The analyses describe local changes within semantic neighborhoods and identify posterior quality, label mappings, and domain shift as conditions affecting performance.

[1] "Gold" refers only to an observed memory label; the query's gold label is unavailable during selection.

2 Related Work

Demonstration retrieval and reranking. Beyond similarity retrieval, selectors have used information compression (Wu et al., 2023), cross-entropy differences (Iter et al., 2023), and conditional entropy after semantic TopK filtering (Peng et al., 2024). Other systems learn selection preferences from target LLM feedback (Zhang et al., 2025) or use influence-guided pseudo-labeling (Chen et al., 2025). LADR instead obtains a task-related selection signal from an auxiliary label predictor. It preserves the semantic candidate pool and avoids using the target LLM to score candidates during reranking.

Label information in demonstration selection. Context label distributions can guide ordering (Xu et al., 2024), Bayesian inverse inference can score demonstrations (Wang et al., 2024), and learned selectors can encode task-defined output similarity (Liu et al., 2025). TopK sampling with Synthetic Data (TopK-SD) (Chen et al., 2026) uses semantic and label information to construct a different retrieval space. Thus, using task information for selection is not itself new. Our focus is the specific combination of auxiliary distribution compatibility, candidate-label anchoring, and semantic reranking within an existing neighborhood. Comparing LADR and LADR-GA with direct candidate-label strategies helps characterize this choice under a shared pool and prompt budget. TopK-SD remains a separate system comparison because it changes the retrieval neighborhood.

3 Method

The method separates candidate relevance from task compatibility. A semantic retriever supplies a small pool, an auxiliary predictor represents its examples in the label space, and a reranker combines the two signals. LADR uses the predictive distributions directly; LADR-GA additionally incorporates the observed candidate labels. The underlying retrieval pool and prompt budget remain unchanged.

3.1 Preliminaries

Let $\mathcal{X}$ be the input space and $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ a labeled demonstration memory of N examples, where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y} = \{1, \ldots, C\}$ for $C \geq 2$ classes. Given a query $x \in \mathcal{X}$, LADR first uses a frozen encoder $f_{\text{emb}} : \mathcal{X} \to \mathbb{R}^d$ to compute normalized d-dimensional embeddings. Their dot product is cosine similarity:

$$S_{\text{sem}}(x, x_i) = f_{\text{emb}}(x)^\top f_{\text{emb}}(x_i). \tag{1}$$

The $K = 30$ memory items with the highest scores form the candidate pool $\mathcal{C}_K(x)$. A reranker assigns each candidate a score $R(x, x_i)$ and retains the highest $L = 8$. Exact ties are resolved by position in the saved semantic pool. Selected pairs are kept in the reranker's order and formatted with their observed labels y_i; the target LLM then predicts the query label. The query's gold label is unavailable to the deployable selectors and prompt formatter. Figure 2 gives an overview.

3.2 Cross-Fitted Posterior Estimation

Evaluation rows are held out from APM fitting; exact-text overlaps are audited separately. A single APM trained on all of $\mathcal{D}$ would still score memory items that occurred in its own training data. This asymmetry can make candidate posteriors more confident than query posteriors and confound the interpretation of distribution alignment. We therefore use row-wise cross-fitting in the controlled experiments.

The memory indices are partitioned into $F = 3$ stratified folds. Let $\mathcal{D}_{-f}$ be all memory rows outside fold f, and let $P_{\text{Aux}}^{(f)}(y \mid z)$ be the probability that the APM trained on $\mathcal{D}_{-f}$ assigns label $y \in \mathcal{Y}$ to input $z \in \mathcal{X}$. Each fold model is fitted by cross-entropy minimization (Appendix F). Its output belongs to the probability simplex $\Delta^{C-1} = \{\mathbf{p} \in \mathbb{R}^C_{\geq 0} : \sum_{y=1}^{C} \mathbf{p}(y) = 1\}$.

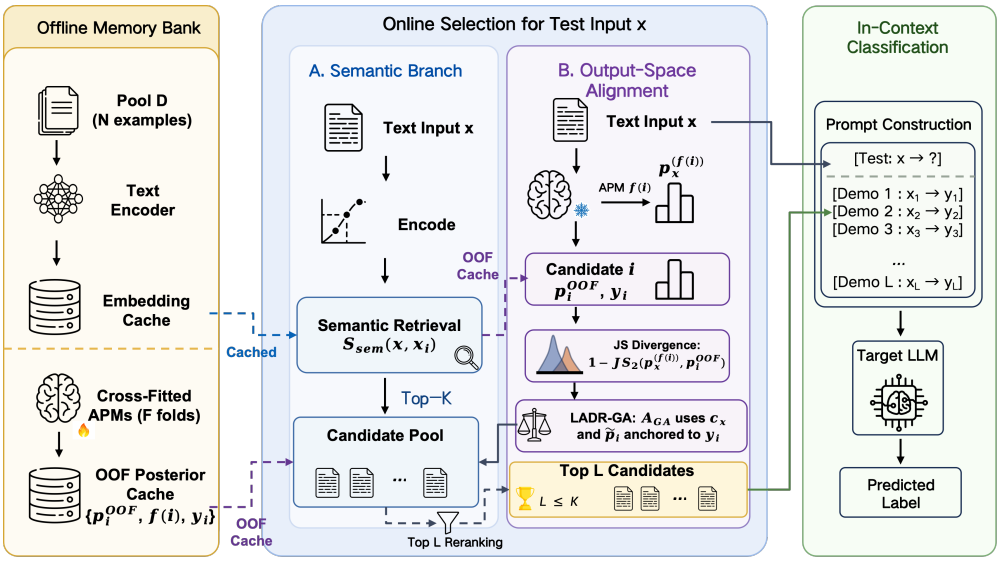


Figure 2: LADR and LADR-GA under the controlled cross-fitting protocol. Retrieval supplies candidates; the APM supplies task predictions; reranking selects demonstrations. In the diagram, JS_2 denotes D_{JS} with base-2 logarithms, so $1 - \mathrm{JS}_2$ is compatibility, not divergence. Candidate labels, but not the query gold label, are observed.

If $f(i)$ denotes the fold holding out memory row i, define its OOF posterior and the corresponding query posterior as

$$\mathbf{p}_i^{\mathrm{OOF}} = P_{\mathrm{Aux}}^{(f(i))}(\cdot \mid x_i), \qquad \mathbf{p}_x^{(f)} = P_{\mathrm{Aux}}^{(f)}(\cdot \mid x). \tag{2}$$

Thus row i is absent from the training data used to produce $\mathbf{p}_i^{\mathrm{OOF}}$. We compare candidate i with $\mathbf{p}_x^{(f(i))}$, the query posterior from the same APM. This pairing ensures that each JS divergence compares predictions from the same auxiliary model. Candidate posteriors and fold assignments are cached. Candidate duplicates and exposure effects are further audited in Section 6.4.

3.3 LADR: Label-Alignment Divergence Reranking

For any two predictive label distributions $\mathbf{p}, \mathbf{q} \in \Delta^{C-1}$, where $\mathbf{p}(y)$ and $\mathbf{q}(y)$ denote the probabilities assigned to label y, let $\mathbf{m} = (\mathbf{p} + \mathbf{q})/2$. We compute Jensen–Shannon divergence (D_{JS}; Lin, 1991) using base-2 logarithms:

$$D_{\mathrm{JS}}(\mathbf{p}, \mathbf{q}) = \tfrac{1}{2} D_{\mathrm{KL}}(\mathbf{p} \| \mathbf{m}) + \tfrac{1}{2} D_{\mathrm{KL}}(\mathbf{q} \| \mathbf{m}). \tag{3}$$

Here,

$$D_{\mathrm{KL}}(\mathbf{p} \| \mathbf{m}) = \sum_{y=1}^{C} \mathbf{p}(y) \log_2 \frac{\mathbf{p}(y)}{\mathbf{m}(y)} \tag{4}$$

is the Kullback–Leibler (KL) divergence; $D_{\mathrm{KL}}(\mathbf{q} \| \mathbf{m})$ is defined analogously. We use the convention that terms with $\mathbf{p}(y) = 0$ contribute zero.

Here, $\mathbf{p}$ and $\mathbf{q}$ are generic label distributions. In the LADR score, they are instantiated as the fold-matched query posterior $\mathbf{p}_x^{(f(i))}$ and the candidate posterior $\mathbf{p}_i^{\mathrm{OOF}}$, respectively. Because $D_{\mathrm{JS}}(\mathbf{p}, \mathbf{q}) \in [0, 1]$ under base-2 logarithms, LADR defines label-distribution compatibility as $1 - D_{\mathrm{JS}}(\mathbf{p}, \mathbf{q})$. Its score is

$$\begin{aligned} R_{\mathrm{LADR}}(x, x_i) = \alpha S_{\mathrm{sem}}(x, x_i) \\ + (1 - \alpha) \left[1 - D_{\mathrm{JS}} \left(\mathbf{p}_x^{(f(i))}, \mathbf{p}_i^{\mathrm{OOF}} \right) \right], \end{aligned} \tag{5}$$

where $\alpha \in [0, 1]$ balances input-space relevance and output-space compatibility. We use raw cosine similarity without score normalization inside the Top-30 pool, so α is a reranking hyperparameter rather than an exact equal-scale mixture coefficient. Empirically, we use the frozen value $\alpha = 0.5$. Compatibility is a surrogate: it represents the auxiliary model's view of the task relationship, not the change in target-LLM accuracy caused by including a candidate. Its utility is therefore assessed empirically.

Equation (5) defines LADR independently of the posterior estimation protocol. In the main benchmark, the query and candidate posteriors are produced by a single APM; in the controlled evaluation, they are replaced by fold-matched query and OOF candidate posteriors. In either case, LADR does not directly use the observed candidate label y_i in its reranking score. Memory labels supervise the APM and are included in the final demonstrations, whereas their direct use during reranking is reserved for LADR-GA. Thus, cross-fitting changes how the posteriors are estimated, not the definition of LADR.

3.4 LADR-GA: Gold Anchoring with Candidate Labels

The query label must be inferred, whereas a candidate's label is already observed. LADR-GA uses this asymmetry rather than treating both predictive distributions as the only available evidence. It moves each candidate posterior toward the observed label y_i while retaining part of its predicted probability distribution:

$$\widetilde{\mathbf{p}}_i = (1 - a_i)\mathbf{p}_i^{\mathrm{OOF}} + a_i \mathbf{e}_{y_i}, \tag{6}$$

where $a_i = 1 - \mathbf{p}_i^{\mathrm{OOF}}(y_i)$ and $\mathbf{e}_{y_i}$ is the one-hot vector of the observed candidate label. Writing $t_i = \mathbf{p}_i^{\mathrm{OOF}}(y_i)$ gives $\widetilde{\mathbf{p}}_i(y_i) = t_i^2 + 1 - t_i$: anchoring increases the observed-label mass by $(1 - t_i)^2$ and preserves a normalized distribution. It is a deliberate reweighting toward known information, not a claim of probability calibration.

For query x, we aggregate fold-specific posteriors as $\bar{\mathbf{p}}_x = F^{-1} \sum_{f=1}^{F} \mathbf{p}_x^{(f)}$ and define an entropy-based concentration score:

$$c_x = 1 - \frac{H(\bar{\mathbf{p}}_x)}{\log_2 C}, \qquad H(\mathbf{p}) = -\sum_{y=1}^{C} \mathbf{p}(y) \log_2 \mathbf{p}(y). \tag{7}$$

The LADR-GA alignment combines candidate-label agreement and posterior compatibility:

$$\begin{aligned} A_{\mathrm{GA}}(x, x_i) = {} & c_x \, \bar{\mathbf{p}}_x(y_i) \\ & + (1 - c_x) \left[1 - D_{\mathrm{JS}} \left(\mathbf{p}_x^{(f(i))}, \widetilde{\mathbf{p}}_i \right) \right]. \end{aligned} \tag{8}$$

The final reranking score is:

$$R_{\mathrm{GA}}(x, x_i) = 0.5 S_{\mathrm{sem}}(x, x_i) + 0.5 A_{\mathrm{GA}}(x, x_i). \tag{9}$$

For a concentrated query posterior, the gate emphasizes direct matching to an observed label; for a diffuse posterior, it emphasizes anchored distribution comparison. Concentration is not the probability of being correct, and the gate is a mixing heuristic rather than an optimality result. The formula and its fixed coefficients were chosen during development. Comparisons with Gold-Soft and Gold-Hard evaluate the complete strategy, not separate causal effects of anchoring and gating.

Computation and scope. Candidate embeddings and posteriors are cached. After retrieval, scoring costs $O(KC)$ arithmetic operations and sorting costs $O(K \log K)$. Controlled selection requires F auxiliary query evaluations, but no target-LLM candidate evaluation. Training the APMs and storing $O(NC)$ candidate probabilities remain real costs. The computational claim is avoidance of target-model reranking, not a measured end-to-end speedup or a training-free system.

4 Experimental Setup

Data. We evaluate SST-2 and SST-5 (Socher et al., 2013), CR (Hu & Liu, 2004), Subj (Pang & Lee, 2004), AG News (Zhang et al., 2015), MNLI (Williams et al., 2018), QNLI (Wang et al., 2018),

Table 1: Accuracy (%) on nine tasks under the original benchmark protocol: Qwen2.5-7B-Instruct, eight demonstrations, and one APM per task. LADR reranks the Top-30 pool produced by semantic TopK. Bold and underlined values denote the best and second-best results.

Task	Random	BM25	TopK	MDL	ConE	MAPLE	GenICL	**LADR**
AG News	68.12	75.77	75.71	77.83	80.95	79.44	**86.66**	<u>85.76</u>
CR	92.82	93.62	93.62	<u>94.41</u>	93.88	94.22	93.35	**94.68**
SST-2	95.22	95.00	<u>96.05</u>	<u>96.05</u>	95.61	95.00	94.15	**96.49**
SST-5	44.93	46.38	50.23	50.05	48.91	50.56	<u>51.63</u>	**54.30**
Subj	73.95	90.85	92.45	92.35	90.75	69.11	<u>93.35</u>	**95.15**
MNLI	76.41	79.08	80.01	79.90	78.61	83.67	**85.95**	<u>84.31</u>
QNLI	82.87	82.67	82.76	82.76	84.26	<u>84.44</u>	83.35	**85.45**
Banking77	42.85	89.64	<u>94.29</u>	93.77	91.82	61.33	84.40	**94.42**
MASSIVE	41.32	83.36	<u>87.93</u>	87.49	86.18	77.67	82.20	**89.61**
Avg. Acc.	68.72	81.82	83.67	83.85	83.44	77.27	<u>83.89</u>	**86.68**

Banking77 (Casanueva et al., 2020), and MASSIVE (FitzGerald et al., 2023). They cover sentiment, subjectivity, topic, natural language inference, and intent classification with up to 77 labels. MNLI and QNLI use their labeled validation sets; MASSIVE uses the English en-US split. SST-2 uses the 6,920-example sentence-level training split from *SetFit/sst2* (SetFit, 2022).

Target LLMs. The main benchmark uses Qwen2.5-7B-Instruct (Qwen Team, 2024b). The controlled evaluation also uses Mistral-7B-Instruct-v0.3 (Mistral AI, 2024). Greedy decoding uses bfloat16 and at most 16 generated tokens. Invalid label outputs are counted as incorrect. We also examine generalizability across LLM families and scales using Gemma2-2B-it (Gemma Team, 2024), Phi3-mini-128k-Instruct (Abdin et al., 2024), Qwen2.5-7B-Instruct (Qwen Team, 2024a), LLaMA3-8B-Instruct (Meta AI, 2024), Gemma2-9B-it (Gemma Team, 2024), and Qwen2.5-14B-Instruct (Qwen Team, 2024a).

Protocols. Semantic retrieval uses normalized embeddings from *gte-base-en-v1.5* (Alibaba-NLP, 2024). The main benchmark follows the original setup: one RoBERTa-base APM, an 8:2 training–validation split, one Qwen run, and nine task configurations. For Banking77 and MASSIVE, the APM is XLM-RoBERTa-base (Conneau et al., 2020). All methods use eight demonstrations. This benchmark provides broad method comparison, but its memory posteriors are not OOF.

The controlled evaluation uses three-fold cross-fitting with RoBERTa-base (Liu et al., 2019). Each fold model is trained for three epochs with learning rate 2×10^{-5}, weight decay 0.01, warmup ratio 0.06, maximum length 256, and effective batch size 32. Three random seeds change APM training and fold assignments. The target LLM uses greedy decoding, so the seed axis measures selection variation. The candidate size, number of demonstrations, and $\alpha = 0.5$ remain fixed. The value of α and the LADR-GA form are empirically chosen during initial development on CR, SST-2, and Subj, then frozen for the controlled runs.

Controlled comparisons. The seven selectors are TopK, Gold-Soft, Gold-Hard, Predicted-Label Filter, Posterior Confidence, LADR, and LADR-GA. Gold-Soft scores candidate i with $0.5S_{\mathrm{sem}}(x, x_i) + 0.5\bar{\mathbf{p}}_x(y_i)$. Gold-Hard first selects candidates whose observed labels match $\arg\max_y \bar{\mathbf{p}}_x(y)$, ranks matching candidates semantically, and fills any remaining positions by semantic rank. Predicted-Label Filter instead matches the most likely query and candidate posterior labels and uses the same semantic fallback. Posterior Confidence ranks candidate i by $\mathbf{p}_i^{\mathrm{OOF}}(\arg\max_y \bar{\mathbf{p}}_x(y))$, with semantic score as the secondary key. Gold-Soft, Gold-Hard, and LADR-GA use candidate labels, which are part of the demonstration memory; no selector observes the query gold label. LADR uses the original distribution-alignment objective with cross-fitted posteriors, whereas LADR-GA additionally uses candidate-label anchoring and entropy-based gating.

We also compare with TopK-SD (Chen et al., 2026) using a fixed $\lambda = 0.7$. Unlike the component controls, TopK-SD reconstructs the neighborhood from the full memory and is therefore treated as

Table 2: Unweighted macro accuracy (%) over the nine controlled tasks. Same-pool selectors report mean ± sample SD over three runs; deterministic TopK and the separate full-memory TopK-SD comparison are fixed. Bold and underlined values are the highest and second-highest means per target LLM. Pred.-Filter and Post.-Conf. denote Predicted-Label Filter and Posterior Confidence.

Selector	Qwen	Mistral
TopK	84.81	82.56
Gold-Soft	<u>85.62 ± 0.10</u>	<u>86.66 ± 0.09</u>
Gold-Hard	85.46 ± 0.09	86.41 ± 0.01
Pred.-Filter	85.36 ± 0.04	84.61 ± 0.02
Post.-Conf.	85.52 ± 0.10	85.26 ± 0.25
TopK-SD	84.99	83.15
LADR	85.47 ± 0.08	84.78 ± 0.19
LADR-GA	$\mathbf{85.69 \pm 0.09}$	$\mathbf{87.49 \pm 0.15}$

a separate system comparison. Its settings were fixed before generating its selections and target outputs.

Statistical analysis. Controlled results report the mean and sample standard deviation (SD) across three selection seeds. Macros weight tasks equally; SDs are computed from the seed-level macros, not averaged task SDs. We report the available paired bootstrap intervals and distinguish them from across-seed variation. Appendix A specifies their interpretation and the original testing procedure. TopK-SD is a separate descriptive system comparison.

5 Results

5.1 Overall Comparison with Existing Selectors

Table 1 compares lexical BM25 (Robertson & Zaragoza, 2009), semantic TopK, and other selection methods. LADR achieves the highest average accuracy, 86.68%. It improves semantic TopK from 83.67% to 86.68%, an increase of 3.01 percentage points (pp) without replacing the underlying semantic retriever. LADR also exceeds MDL, ConE, and GenICL by 2.83, 3.24, and 2.79 pp, respectively.

LADR improves over TopK on all nine tasks. It obtains the highest point estimate on seven tasks and ranks second on AG News and MNLI. The gains over TopK are particularly large on AG News ($+10.05$ pp), SST-5 ($+4.07$ pp), MNLI ($+4.30$ pp), and QNLI ($+2.69$ pp). On Banking77 and MASSIVE, which contain 77 and 60 labels, LADR reaches 94.42% and 89.61%, outperforming TopK by 0.13 and 1.68 pp. These values are point estimates from one run under the main benchmark protocol.

5.2 Controlled Evaluation with Cross-Fitted Posteriors

Table 2 reports the cross-fitted evaluation. LADR raises the TopK macro from 84.81% to 85.47% on Qwen and from 82.56% to 84.78% on Mistral. LADR-GA reaches 85.69% and 87.49%, giving improvements of 0.88 and 4.93 percentage points (pp) over TopK. The recorded Qwen interval for the latter comparison is $[+0.51, +1.25]$ pp.

Gold-Soft is a strong comparator at 85.62% and 86.66%, exceeding the base LADR means on both target models. LADR-GA exceeds Gold-Soft by 0.07 pp on Qwen, with a paired 95% interval of $[-0.14, +0.27]$, and by 0.83 pp on Mistral, with an interval of $[+0.62, +1.03]$. The Qwen difference is not clearly resolved by this interval; the Mistral interval lies above zero. These results support a target-dependent incremental benefit rather than a universal advantage for the more elaborate score. The gains of LADR-GA over LADR are 0.22 and 2.71 pp, respectively. Against the separate TopK-SD system comparison, its point-estimate margins are 0.70 and 4.34 pp.

The controlled results include losses on individual tasks: LADR-GA is 2.56 pp below TopK on Qwen MNLI and 1.00 pp below TopK on Mistral Banking77. Appendix E gives the task-level means and SDs. The aggregate advantage should not be read as improvement on every task.

5.3 Performance across Target Models

Figure 3 (Appendix B) compares six target models from the Gemma 2, Phi-3, Qwen2.5, and Llama 3 families, spanning 2B–14B parameters. Under this original-protocol seven-task panel, LADR has the highest average accuracy among the plotted methods for every target model. This broadens the evidence beyond a single architecture, without implying a constant gain across target models or evaluation protocols.

6 Selection Behavior and Reliability

The two accuracy panels answer **RQ1** within their own protocols. The broad benchmark provides comparison breadth; the controlled study tests the selection rules with cross-fitted predictions and repeated runs. Their absolute accuracies are not interchangeable. We now examine which information the rules use and what changes in the selected examples.

6.1 Alternative Uses of Auxiliary Predictions

The controlled comparisons address **RQ2** without changing the candidate pool or prompt budget. Predicted-Label Filter discards probability magnitudes and matches predicted classes; Posterior Confidence ranks candidates by support for the query's most likely class. Gold-Hard and Gold-Soft additionally use observed candidate labels. LADR and LADR-GA instead compare distributions, with the latter also incorporating those labels.

Gold-Soft has a particularly direct interpretation: $\bar{\mathbf{p}}_x(y_i) = \mathbb{E}_{Y\sim\bar{\mathbf{p}}_x}[\mathbb{1}\{Y = y_i\}]$, where Y is a label drawn from the query posterior and $\mathbb{1}$ is the indicator function. Its label term is the auxiliary model's estimated probability of matching the candidate label, not a target-LLM utility estimate. This is a meaningful comparator, rather than an intentionally weak alternative. The controlled results show that explicit candidate labels capture an important part of the available benefit. LADR-GA's value is the additional performance of the complete combination, especially on Mistral, not evidence that every posterior coordinate is always useful.

6.2 Changes within Semantic Neighborhoods

We also compare the TopK and LADR lists query by query on CR, SST-2, and Subj in panel (a) of Table 3 (Appendix C), using the original Qwen single-APM protocol. Label α-nDCG treats label identities as intents, whereas semantic α-nDCG uses embedding-based clusters. LADR retains 6.20 TopK demonstrations (77.52%) with a Jaccard similarity of 0.683; label α-nDCG changes by -10.07, but semantic α-nDCG changes by only -0.21. These statistics describe changes to label composition within a largely preserved semantic neighborhood, consistent with refinement after retrieval.

For panel (b), we repeat the analysis on the seven listed tasks. Pred-Match@8 is the fraction of selected demonstrations whose most likely APM label matches the query's; Majority@8 is the largest predicted-label share; Gold-Match@8 uses query and candidate gold labels only after selection. Pred-Match@8 is 93.38%, while Gold-Match@8 increases from 67.7% for TopK to 76.5% for LADR. The selected lists therefore contain a larger share of matching gold labels in this panel. This is an association between the selection rule and prompt composition, not a causal explanation of the accuracy gain. Moreover, examples are presented in reranker order, so membership and ordering effects are not separately identified by these comparisons.

6.3 Incorporating Observed Candidate Labels

The controlled label-based comparisons extend **RQ2** to information already available in memory. Gold-Soft combines the query's label probabilities with observed candidate labels, whereas LADR-GA also uses anchored candidate distributions and a query concentration gate. LADR-GA has the

highest macro mean in the controlled evaluation, with a larger margin over Gold-Soft on Mistral than on Qwen.

Panel (c) of Table 3 describes the corresponding selections, averaged over nine tasks and three seeds. LADR-GA shares 7.27 of eight demonstrations with Gold-Soft, compared with 6.67 with LADR. The largest observed-label fractions in the prompts are 0.970 for LADR-GA and 0.980 for Gold-Soft. The methods therefore often select similar label configurations, even though their scoring rules differ.

Taken together, the analyses describe a refinement of semantic retrieval: posterior alignment changes part of the selected neighborhood, and explicit use of memory labels can further improve the resulting strategy. The small Qwen margin and high selection overlap are consistent with direct label matching capturing much of the benefit in that setting. They do not establish statistical equivalence between the methods.

6.4 Reliability and Boundary Conditions

RQ3 concerns the quality of the compatibility signal. The development weight-sensitivity curves favor combining semantic relevance and posterior alignment on the three plotted tasks, although their best weights differ (Appendix D). On CR, SST-2, and Subj, a full RoBERTa-base APM raises the average from 94.04% for TopK to 95.10%, whereas an APM trained on a label-balanced eight-shot subset reaches 93.57% and falls to 92.75% when semantic scores are removed from reranking. The benefit therefore depends on useful auxiliary predictions, not merely on adding a second score.

Label mappings and domain shift also constrain the gains. Replacing standard verbalizers with letters reduces the seven-task average by 5.43 pp; reversing demonstration labels reduces it by 37.84 pp. The recorded candidate-exposure audit reports no statistically distinguished macro difference, and its overlap-removal check preserves the aggregate conclusions. Neither finding proves the absence of exposure effects in general. However, source-only transfer between SST-2 and CR does not yield consistent gains across target models. These findings support label alignment as a task-dependent complement to semantic retrieval. Appendix D reports the full diagnostics, including weight sensitivity, label mappings, and exposure and overlap audits.

7 Conclusion

We introduced LADR to use auxiliary task predictions for refining semantic demonstration retrieval. LADR-GA extends this approach by incorporating labels already available in memory. The contribution is a concrete selection strategy and its evaluation: task supervision informs reranking without updating the target LLM or asking it to score candidate demonstrations.

The original benchmark and the separate controlled evaluation show aggregate improvements over semantic TopK. Direct label-probability matching is already competitive, and the additional gain of LADR-GA depends on the target model. Selection diagnostics describe a local refinement of semantic neighborhoods, while weak predictors and domain shift expose important limitations. These findings support auxiliary label information as a practical complement to semantic relevance, rather than a universal preference for full-distribution scoring. The evidence concerns finite-label classification with labeled task data; extension to generation and independent attribution of the anchoring, gating, membership, and ordering effects remain open directions. Practically, cached candidate posteriors can support several frozen target LLMs without learning a new target-specific reranker. This is an architectural reuse property, not a guarantee of identical gains: the controlled comparison shows why each target model must still be evaluated. LADR therefore offers a modular use of task supervision alongside an existing semantic retriever.

ETHIC STATEMENT

This work evaluates LADR and LADR-GA on existing text classification benchmarks using pretrained language models. Demonstration selection may propagate biases from auxiliary predictions and dataset annotations, and benchmark accuracy alone does not establish fairness or suitability for consequential decisions. Downstream applications should assess data representativeness, evaluate performance across relevant groups and domains, and maintain appropriate human oversight. Users should respect dataset and model licenses and protect personal or confidential information in demonstration memories and prompts. Although the method avoids target-model reranking, auxiliary training and inference still incur computational costs; environmental benefits require end-to-end assessment.

AI USE STATEMENT

Generative AI tools assisted with manuscript organization, English-language polishing, citation checking, and LaTeX formatting. Responsibility for the scientific claims, reported results, citations, and final wording rests with the authors.

REPRODUCIBILITY STATEMENT

Section 3 specifies the selection scores, posterior estimation, candidate pool, and prompt budget. Section 4 describes the datasets, target models, auxiliary-model training, and comparator definitions. Appendices A and E describe the statistical summaries and provide the existing task-level controlled results. Section 6 describes the selection and reliability diagnostics. The accompanying source distinguishes recorded summaries from raw predictions; this manuscript revision did not rerun model training, inference, or hypothesis tests. The anonymous code link is provided on the first page.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. URL https://arxiv.org/abs/2404.14219.

Alibaba-NLP. gte-base-en-v1.5: Model card and repository. Hugging Face repository, 2024. URL https://huggingface.co/Alibaba-NLP/gte-base-en-v1.5/tree/a829fd0e060bb84554da0dfd354d0de0f7712b7f.

Anonymous Authors. When does posterior geometry matter for in-context demonstration selection? Concurrent manuscript; author identities withheld for double-blind review, 2026.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 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/.

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, 2026. doi: 10.1609/aaai.v40i36.40273. URL https://ojs.aaai.org/index.php/AAAI/article/view/40273.

Zihan Chen, Song Wang, Zhen Tan, Jundong Li, and Cong Shen. MAPLE: Many-shot adaptive pseudo-labeling for in-context learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9277–9294. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/chen25bo.html.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, 2020. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747/.

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/.

Gemma Team. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

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/FBS/sentiment-analysis.html.

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/.

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.

Hui Liu, Wenya Wang, Hao Sun, Chris Xing Tian, Chenqi Kong, Xin Dong, and Haoliang Li. Unraveling the mechanics of learning-based demonstration selection for In-Context Learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2623–2641. Association for Computational Linguistics, jul 2025. doi: 10.18653/v1/2025.acl-long.132. URL https://aclanthology.org/2025.acl-long.132/.

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. *CoRR*, abs/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/.

Meta AI. Llama 3 model card, 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, 2022. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759/.

Mistral AI. Mistral-7B-Instruct-v0.3: Model card and repository. Hugging Face repository, May 2024. URL https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3/tree/c170c708c41dac9275d15a8fff4eca08d52bab71.

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, 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/.

Qwen Team. Qwen2.5: A party of foundation models, September 2024a. URL https://qwenlm.github.io/blog/qwen2.5/.

Qwen Team. Qwen2.5-7B-Instruct: Model card and repository. Hugging Face repository, September 2024b. URL https://huggingface.co/Qwen/Qwen2.5-7B-Instruct/tree/a09a35458c702b33eeacc393d103063234e8bc28.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019. URL https://www.staff.city.ac.uk/~sbrp622/papers/foundations_bm25_review.pdf.

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/.

SetFit. SST-2: Dataset card and repository. Hugging Face repository, 2022. URL https://huggingface.co/datasets/SetFit/sst2.

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/.

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 *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355. Association for Computational Linguistics, 2018. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446/.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9840–9855, 2023. doi: 10.18653/v1/2023.emnlp-main.609. URL https://aclanthology.org/2023.emnlp-main.609/.

Siyin Wang, Chao-Han Huck Yang, Ji Wu, and Chao Zhang. Bayesian example selection improves In-Context Learning for speech, text and visual modalities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20812–20828. Association for Computational Linguistics, nov 2024. doi: 10.18653/v1/2024.emnlp-main.1158. URL https://aclanthology.org/2024.emnlp-main.1158/.

Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, and Quoc V. Le. Symbol tuning improves in-context learning in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 968–979, 2023. doi: 10.18653/v1/2023.emnlp-main.61. URL https://aclanthology.org/2023.emnlp-main.61/.

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/.

Zhichao Xu, Daniel Cohen, Bei Wang, and Vivek Srikumar. In-Context example ordering guided by label distributions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2623–2640. Association for Computational Linguistics, jun 2024. doi: 10.18653/v1/2024.findings-naacl.167. URL https://aclanthology.org/2024.findings-naacl.167/.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. Ground-truth labels matter: A deeper look into input-label demonstrations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2422–2437, 2022. doi: 10.18653/v1/2022.emnlp-main.155. URL https://aclanthology.org/2022.emnlp-main.155/.

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.

Zheng Zhang, Shaocheng Lan, Lei Song, Jiang Bian, Yexin Li, and Kan Ren. Learning to select in-context demonstration preferred by large language model. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11345–11360. Association for Computational Linguistics, July 2025. doi: 10.18653/v1/2025.findings-acl.592. URL https://aclanthology.org/2025.findings-acl.592/.

A Statistical Analysis

Controlled accuracies are unweighted means across the nine tasks. For each selection seed, the task accuracies are first averaged; the reported macro SD is the sample SD of these three seed-level macros. Averaging task SDs would not yield this quantity because task-level results can covary across seeds. The tables retain recorded macros rather than recomputing them from rounded task entries. A deterministic TopK run has no selection-seed variation. The TopK-SD values are from a separate fixed system comparison, not three repeated controlled runs.

The original procedure uses 100,000 paired bootstrap draws over seeds and test instances. The recorded Qwen LADR-GA–TopK interval is $[+0.51, +1.25]$ pp. The LADR-GA–Gold-Soft intervals are $[-0.14, +0.27]$ pp for Qwen and $[+0.62, +1.03]$ pp for Mistral. The latter accompanies the corrected Mistral summary. These intervals are reported paired summaries; they are not calculated from the rounded task means or from the between-seed SDs in the tables.

The original randomization procedure averages paired differences over seeds. Within each target model, Holm correction covers 54 task-level comparisons (nine tasks against six controlled comparators); the six macro comparisons form a separate family. The comparators are TopK, Gold-Soft, Gold-Hard, Predicted-Label Filter, Posterior Confidence, and LADR. TopK-SD is outside this family. We do not carry over unreported adjusted p-values to the corrected Mistral results or infer familywise significance from an individual bootstrap interval. Target models and evaluation panels are analyzed separately.

B Performance across Target Models

Figure 3 supplements the comparison in Section 5. It uses the original benchmark protocol, rather than the cross-fitted controlled protocol, and averages the seven tasks listed in the caption.

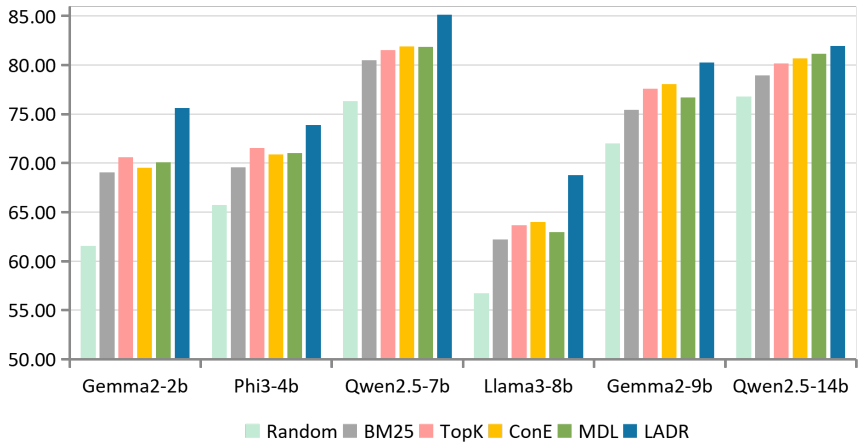


Figure 3: Average accuracy on the seven tasks (AG News, CR, SST-2, SST-5, Subj, MNLI, and QNLI) under the original benchmark protocol for comparison breadth. Target LLMs span four families and 2B–14B parameters. The plot's "Phi3-4b" label refers to Phi-3-mini (3.8B parameters); model labels are abbreviated.

C SELECTION DIAGNOSTICS

Table 3 reports the statistics discussed in Section 6. The panels distinguish changes to the semantic selection, agreement with labels measured after selection, and the observed-label composition of LADR-GA prompts. These summaries characterize the selected examples; query gold labels enter only the post-selection Gold-Match diagnostic. Mean retention, Jaccard, and label-composition statistics are averaged separately; one aggregate is not obtained by applying a nonlinear transformation to another aggregate. The α in the metric name α-nDCG is its novelty discount, distinct from the reranking weight in Equation (5). Label identities and embedding clusters define the two diversity views, respectively.

Table 3: Selection diagnostics. Panel (a) uses Qwen single-APM averages on CR, SST-2, and Subj; nDCG differences also average BERT-, RoBERTa-, and DeBERTa-based diagnostics. Panel (b) uses Qwen single-APM averages on AG News, CR, SST-2, SST-5, Subj, MNLI, and QNLI. Panel (c) averages nine tasks and three controlled seeds. Δ is LADR minus TopK; Gold-Match is computed only after selection. nDCG denotes normalized discounted cumulative gain. GA and GS denote LADR-GA and Gold-Soft.

(a) Changes to semantic TopK

Statistic	**Value**
Retained TopK items	6.20/8
Retention@8	77.52%
Jaccard similarity	0.683
Δ label α-nDCG	-10.07
Δ semantic α-nDCG	-0.21

(b) Posterior to gold label

Statistic	**Value**
Pred-Match@8	93.38%
Majority@8	95.02%
TopK Gold-Match@8	67.7%
LADR Gold-Match@8	76.5%
Δ Gold-Match@8	$+8.8$ pp

(c) LADR-GA selection

Statistic	**Value**
Query concentration	0.736
Anchored label mass	0.923
Overlap with Gold-Soft	7.27/8
Overlap with LADR	6.67/8
Dominant share (GA/GS/LADR)	.970/.980/.902

D Reliability and Boundary Conditions

The following details supplement Section 6.4 without changing the evaluation protocols. Table 4 collects the posterior, label-mapping, exposure, transfer, and overlap diagnostics.

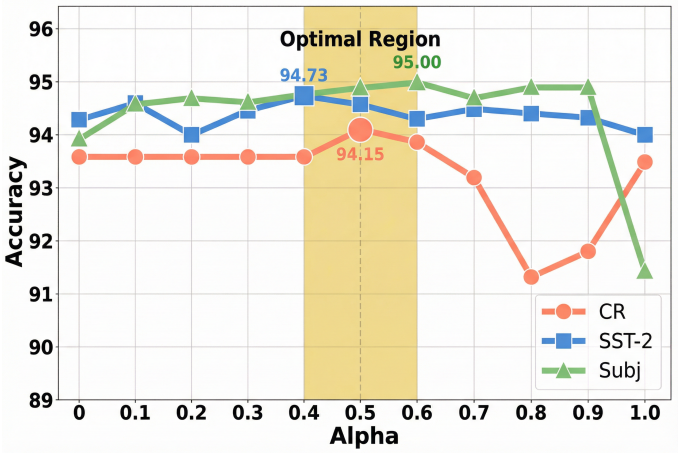


Figure 4: Per-task accuracy on CR, SST-2, and Subj in the development weight-sensitivity diagnostic. Each curve is a separate task, not a task average. The highlighted band summarizes this diagnostic; it is not a universally optimal range.

The weight-sensitivity diagnostic examines the interpolation coefficient α in Equation (5). Both endpoints retain the initial semantic candidate pool: $\alpha = 0$ removes semantic scores only from reranking, whereas $\alpha = 1$ ranks that pool only by semantic similarity. Figure 4 favors intermediate weights on the plotted tasks, with task-dependent maxima. The fixed $\alpha = 0.5$ is a development choice, not a universally optimal coefficient. This diagnostic is a separate evaluation configuration and is not pooled with the broad or controlled benchmark.

Panel (a) of Table 4 combines two Qwen single-APM diagnostics. For posterior quality, Full APM is RoBERTa-base trained on the full task training split; Weak APM uses the same architecture with a label-balanced eight-shot subset. Both use $\alpha = 0.5$ unless semantic guidance is removed with $\alpha = 0$. Full APM raises the three-task average from 94.04% to 95.10%, whereas Weak APM reaches 93.57% and falls to 92.75% when the semantic term is removed from reranking.

For label mapping, Original uses each dataset's standard verbalizers; Letters replaces them with arbitrary alphabetic symbols while preserving the class mapping; Reversed changes the labels attached to demonstrations while retaining the evaluation mapping. Letters reduces the seven-task average by 5.43 pp, whereas Reversed reduces it by 37.84 pp. The Letters condition preserves the class mapping, whereas Reversed changes the example–label relation. Their effects therefore concern different prompt interventions; neither isolates posterior geometry.

A controlled exposure audit compares cross-fitted LADR with settings in which APM training includes part or all of the candidate memory. The recorded macro comparisons are not statistically distinguished; normalized cross-fold duplicates account for 0.07785% under the recorded normalization. A null result is not proof that exposure never matters. A source-only transfer analysis between SST-2 and CR, with aligned label semantics and exact text overlap removed, finds no consistent gain across target LLMs. Removing overlapping evaluation rows leaves the conclusions unchanged. For **RQ3**, these checks provide no detected macro effect of candidate exposure and show that the aggregate conclusions persist after overlap removal. Together with the weak-posterior and transfer results, they place the method's benefits in settings where the auxiliary predictor supplies useful task information; broad cross-domain transfer remains unsupported.

D.1 Summary of Diagnostic Results

Table 4: Boundary diagnostics. Panel (a): Qwen single-APM averages on CR, SST-2, and Subj for posterior quality, and on the seven tasks in panel (b) of Table 3 for label mapping. Δ is relative to TopK or Original; accuracy and Δ are in percent and percentage points. Panel (b): exposure uses Qwen; transfer averages SST-2→CR and CR→SST-2 and is reported as TopK→LADR-GA. The final rows are controlled LADR-GA–TopK differences. "n.s." means not statistically distinguished.

(a) Posterior and label sensitivity

Diagnostic	**Condition**	**Acc.**	Δ
Posterior quality	TopK	94.04	ref.
	Full APM	95.10	$+1.06$
	Weak APM	93.57	-0.47
	Weak APM, $\alpha = 0$	92.75	-1.29
Label mapping	Original	85.16	ref.
	Letters	79.73	-5.43
	Reversed	47.32	-37.84

(b) Exposure, transfer, and overlap audits

Audit	**Outcome**
Candidate exposure	max. $+0.27$ pp; n.s.
Cross-fold duplicates	0.07785%
Transfer, Qwen	$92.42 \rightarrow 92.06$; n.s.
Transfer, Mistral	$90.05 \rightarrow 88.41$; lower mean
Evaluation overlap	42/45,171; conclusions unchanged
Qwen MNLI	-2.56 pp vs. TopK
Mistral Banking77	-1.00 pp vs. TopK

E Complete Controlled Results

The following tables expand the existing nine-task panel in Table 2; they add no new experiments. They use the same task summaries also reported as Suite A in Anonymous Authors (2026). Here, LADR is the method called LADR-Base in the archived results, and LADR-GA is the FPA rule in the concurrent analysis. TopK-SD remains a separate full-memory comparison and is not included in these same-pool task tables.

Table 5: Qwen task-level accuracy (%). Entries are recorded means ± sample SD over three selection seeds, except deterministic TopK. The macro SD is taken from seed-level macros, not averaged task SDs. Bold marks the largest deployable mean in each row.

Task	TopK	Gold-Soft	Gold-Hard	Pred.-Filter	Post.-Conf.	LADR	LADR-GA
SST-2	95.72	96.34±0.14	96.30±0.11	96.27±0.15	95.94±0.40	96.30±0.27	**96.43±0.20**
SST-5	56.83	60.36±0.39	59.70±0.45	57.90±0.13	58.79±0.32	58.34±0.59	**60.94±0.51**
CR	94.95	94.41±0.46	94.33±0.41	**95.12±0.15**	93.88±0.46	94.68±0.53	94.86±0.41
Subj	85.95	86.55±0.05	86.55±0.05	86.45±0.18	**88.73±0.19**	86.57±0.29	86.48±0.16
AG News	91.89	93.91±0.06	93.85±0.06	93.11±0.09	93.90±0.04	93.36±0.05	**93.93±0.05**
MNLI	**80.41**	77.62±0.04	77.63±0.03	79.82±0.05	78.83±0.07	79.28±0.02	77.85±0.10
QNLI	84.53	87.06±0.08	87.04±0.05	86.58±0.14	**87.17±0.16**	86.72±0.16	87.02±0.16
Banking77	88.96	88.74±0.25	88.21±0.06	88.48±0.38	87.98±0.46	**89.07±0.18**	88.04±0.33
MASSIVE	84.03	85.56±0.22	85.57±0.06	84.49±0.27	84.47±0.43	84.94±0.12	**85.65±0.45**
Macro	84.81	85.62±0.10	85.46±0.09	85.36±0.04	85.52±0.10	85.47±0.08	**85.69±0.09**

Table 6: Mistral task-level accuracy (%). Entries are recorded means ± sample SD over three selection seeds, except deterministic TopK. The macro SD is taken from seed-level macros, not averaged task SDs. Bold marks the largest deployable mean in each row.

Task	TopK	Gold-Soft	Gold-Hard	Pred.-Filter	Post.-Conf.	LADR	LADR-GA
SST-2	95.17	96.27±0.24	96.17±0.27	95.79±0.28	95.99±0.10	95.97±0.14	**97.25±0.12**
SST-5	51.40	59.58±0.22	58.55±0.36	52.35±0.82	52.67±1.50	51.42±0.61	**60.43±0.10**
CR	94.68	93.79±0.15	93.71±0.15	94.59±0.41	93.53±0.31	93.71±0.31	**94.88±0.30**
Subj	86.20	92.83±0.03	92.80±0.05	91.13±0.18	92.68±0.45	91.72±0.24	**93.97±0.21**
AG News	90.61	94.00±0.04	94.03±0.05	92.07±0.05	93.18±0.06	92.32±0.11	**94.90±0.12**
MNLI	73.91	81.89±0.03	81.86±0.03	78.70±0.10	80.73±0.08	79.71±0.13	**82.88±0.32**
QNLI	74.89	85.00±0.12	84.99±0.11	79.96±0.35	84.27±0.27	80.87±0.28	**85.71±0.21**
Banking77	91.66	90.93±0.15	89.70±0.21	90.92±0.34	89.10±0.59	**91.71±0.07**	90.66±0.02
MASSIVE	84.57	85.66±0.19	85.84±0.07	85.94±0.09	85.16±0.20	85.63±0.35	**86.71±0.05**
Macro	82.56	86.66±0.09	86.41±0.01	84.61±0.02	85.26±0.25	84.78±0.19	**87.49±0.15**

F Implementation and Interpretation Details

Auxiliary objective. The fold-specific predictor is trained with

$$\mathcal{L}_{\mathrm{Aux}}^{(f)} = -\frac{1}{|\mathcal{D}_{-f}|} \sum_{(x_j, y_j) \in \mathcal{D}_{-f}} \log P_{\mathrm{Aux}}^{(f)}(y_j \mid x_j), \tag{10}$$

where j indexes training rows and the log in this training objective is the natural logarithm. The JS divergence and entropy used for selection instead use base-2 logarithms, with zero-mass terms contributing zero. Cross-fitting excludes a candidate's row from its own predictor's fitting data; it does not by itself remove duplicate texts or related examples.

Information used by the scores. All methods in the controlled panel use the same semantic candidate pool and eight-example prompt budget. LADR scores use predictive distributions, whereas Gold-Soft, Gold-Hard, and LADR-GA also directly access the observed memory labels. These are permissible selection inputs, not query labels. An oracle selector that observes a query label would be a non-deployable comparator, not a guaranteed upper bound: matching a known label does not ensure an optimal demonstration set for the target LLM.

Score ranges and anchoring. Normalized embeddings give $S_{\mathrm{sem}} \in [-1, 1]$; base-2 JS compatibility lies in $[0, 1]$. Thus the fixed $0.5/0.5$ coefficients do not standardize score variation. For $t_i = \mathbf{p}_i^{\mathrm{OOF}}(y_i)$, anchoring yields $\widetilde{\mathbf{p}}_i(y_i) = 1 - t_i + t_i^2$ and $\widetilde{\mathbf{p}}_i(y) = t_i \mathbf{p}_i^{\mathrm{OOF}}(y)$ for $y \neq y_i$. All entries remain nonnegative and sum to one. These are algebraic properties of the rule, not calibration or target-accuracy guarantees. The entropy gate lies in $[0, 1]$ for $C \geq 2$ and mixes two bounded compatibility signals.

Benchmark adaptation. The broad comparison uses eight demonstrations for every method. MAPLE was introduced for many-shot ICL (Chen et al., 2025); its result here is an eight-shot comparison, not a reproduction of its original many-shot configuration. Likewise, the TopK-SD comparison fixes $\lambda = 0.7$ for this evaluation rather than claiming that it is optimal for every dataset or target model.

Resource and evaluation scope. Caching candidate posteriors costs $O(NC)$ storage in addition to the semantic embedding index. After the F auxiliary query predictions, scoring the retrieved pool costs $O(KC)$ and sorting it costs $O(K \log K)$; target-model inference is performed only after selection. These operation counts exclude auxiliary training, retrieval, model loading, and target decoding. They do not establish a measured latency or energy advantage. The experiments evaluate classification with a labeled memory, not adaptation with only eight labels available to the entire system. No target-independent optimality or extension to open-ended generation is claimed.