CROSS: CONSERVATIVE ROUTING OVER COMPLEMENTARY DEMONSTRATION SELECTORS FOR IN-CONTEXT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In-context learning (ICL) enables large language models (LLMs) to adapt to novel tasks from a small number of demonstrations, but its performance depends heavily on which demonstrations are selected. Although existing approaches often produce query-specific rankings or subsets, they typically apply a fixed selection strategy, with a predetermined bias toward relevance, diversity, or their trade-off, across different queries. This formulation overlooks query-level heterogeneity: some queries may benefit more from tightly localized relevance, whereas others may benefit from broader diversity at the demonstration-set level. To address this challenge, we propose a novel framework named Conservative Routing over Complementary Demonstration Selectors (CROSS) for demonstration selection in in-context learning. At its core, CROSS treats the prediction pathway associated with each demonstration selector as an expert and learns from calibration data which expert is more reliable for a given query. Specifically, one branch of CROSS learns a label-free projection into a Poincaré ball that preserves local neighborhood structures defined in the target LLM's representation space, and then retrieves demonstrations based on hyperbolic distance. The other branch employs a determinantal point process (DPP) to select a demonstration set that balances query-aware relevance with set-level diversity. Building on these complementary branches, CROSS trains a lightweight, prediction-aware router using calibration instances where the two branches disagree and exactly one prediction is correct, thereby learning which selector is more reliable for a given query. Additionally, we conduct extensive experiments across multiple benchmarks and LLM backbones, where CROSS consistently outperforms diverse demonstration-selection baselines. The source code for CROSS is available at https://anonymous.4open.science/r/CROSS-3776.

1 INTRODUCTION

The emergence of large language models (LLMs) has driven remarkable progress (Chang et al., 2024; Minaee et al., 2024; Naveed et al., 2025) across a wide range of tasks, including question answering (Zhuang et al., 2023; Molfese et al., 2025), code generation (Wang et al., 2025a; Ma et al., 2026), and knowledge-intensive reasoning (Li et al., 2025; Jiang et al., 2026; Xiao et al., 2026; Fu et al., 2025). Among various adaptation paradigms, in-context learning (ICL) provides a lightweight approach that steers model behavior through task demonstrations without updating model parameters (Dong et al., 2024; De Wynter, 2026; Sinha et al., 2024). However, the performance of ICL is highly sensitive to the choice of demonstrations included in the prompt (Wang et al., 2024b). Since the context budget typically permits only a small subset of the available demonstrations to be contained, selecting the most appropriate demonstrations for each query (Wang et al., 2023; Purohit et al., 2025) becomes critical to effective ICL.

Existing approaches to demonstration selection typically optimize predefined criteria, most notably query-dependent relevance (Levy et al., 2023; Peng et al., 2024) and set-level diversity (Wang et al., 2025b; Xin et al., 2026), with some methods additionally considering model-specific preferences (Wang et al., 2024a). However, these criteria impose different retrieval preferences, and no single preference is uniformly preferable across queries (Wu et al., 2023; Qi et al.). Notably, some queries may benefit from a concentrated neighborhood of highly relevant demonstrations, whereas others prefer broader coverage provided by a more diverse set.

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.

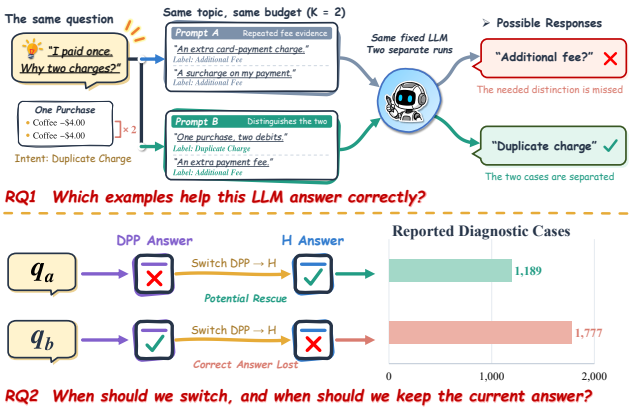


Figure 1: Motivation of CROSS. The **top** panel reveals query-dependent demonstration utility, while the **bottom** panel shows that switching between selectors can be either beneficial or harmful, motivating complementary selection with conservative routing.

As illustrated in Figure 1, many existing demonstration selection approaches rely on representations produced by external semantic encoders. However, the resulting representation geometry primarily reflects generic semantic similarity rather than demonstration utility for the target LLM, potentially leading to a representation-selection mismatch. This motivates our first research question: ❶ ***How can we learn a retrieval geometry that is better aligned with the target LLM for demonstration selection?*** Beyond representation alignment, a fixed selector typically encodes a predefined inductive bias, whereas different queries may favor different criteria, such as query-local relevance or set-level diversity. This leads to our second research question: ❷ ***How can we adaptively leverage complementary selectors while avoiding harmful routing decisions?***

To address these challenges, we propose a novel framework, termed Conservative Routing over Complementary Demonstration Selectors (CROSS), for demonstration selection in ICL. At its core, CROSS conservatively routes between two complementary selectors: a hyperbolic selector emphasizing query-local relevance and a DPP-based selector promoting set-level diversity. On the one hand, CROSS learns a target-aligned projection that distills local neighborhood relations from frozen LLM representations into hyperbolic space and retrieves demonstrations using geodesic distances. On the other hand, CROSS constructs a query-conditioned DPP kernel to select demonstration sets that balance query relevance and set-level diversity. To adaptively choose between the two selectors, we employ a logistic router with a calibrated switching threshold that maximizes routing utility, thereby allowing CROSS to switch selectors only when the predicted benefit is sufficiently large. Extensive experiments on benchmark datasets validate the effectiveness of the proposed CROSS. The main contributions of CROSS are summarized as follows:

❶ ***Problem Reframing.*** We identify two key limitations of existing demonstration selection methods: generic embedding geometries may not align with the target LLM's selection preferences, while a single fixed selector may fail to accommodate heterogeneous query-level retrieval needs.

❷ ***Novel Framework with Theoretical Analysis.*** We propose CROSS, a conservative routing framework that adaptively routes between the hyperbolic and DPP-based selectors for each query. Theoretically, we characterize the complementarity between the two selectors and derive an upper bound on the router's prediction risk, providing theoretical support for CROSS.

❸ ***Comprehensive Evaluation.*** Extensive experiments across multiple backbones and benchmarks demonstrate that CROSS achieves consistent improvements over strong demonstration-selection baselines, validating the generalizability of CROSS.

2 Related Work

In-context Learning. In-context learning (ICL) (Coda-Forno et al., 2023; Krishnamurthy et al., 2024; Wu et al., 2024) adapts large language models (LLMs) (Yin et al., 2023; Liu et al., 2025b) to downstream tasks using a small set of demonstrations without parameter updates (Brown et al., 2020; Zhao et al., 2021; Min et al., 2022). However, its performance is highly sensitive to which demonstrations are included in the limited context (Lu et al., 2022; He et al., 2026; Li et al., 2024), motivating extensive research on query-specific demonstration selection (Wang et al., 2023; Qin et al., 2024). Furthermore, most existing methods adopt a fixed retrieval preference across queries (Xin et al., 2026), implicitly assuming that a single selection bias suits all inputs. We challenge this assumption and ask whether heterogeneous queries may instead favor different demonstration-selection preferences within the same target task.

Demonstration Selection. Prior work on demonstration selection in in-context learning (Wang et al., 2023; Luo et al., 2024) mainly follows three paradigms: similarity-based retrieval, set-level

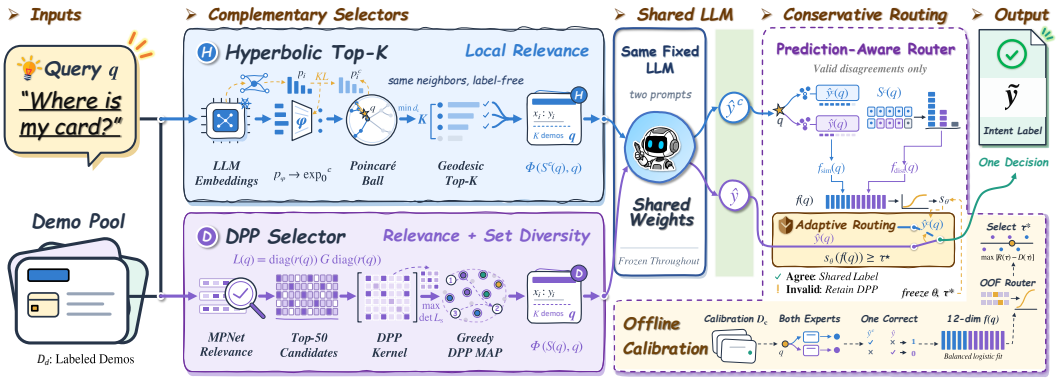


Figure 2: Overview of the proposed CROSS. It integrates a target-aligned hyperbolic selector for query-local relevance, a query-conditioned DPP selector for relevance–diversity balance, and a calibrated logistic router that conservatively switches between them.

diversity or coverage modeling, and learning-based selection. Representative methods include EPR (Rubin et al., 2022) and MDR (Wang et al., 2024a) for relevance-based retrieval, CEIL (Ye et al., 2023) and UCS (Xin et al., 2026) for set-level selection, and GENICL (Zhang et al., 2025c) and AutoSelect (Qi et al.) for learned selection. In contrast, our work does not commit to a single selection principle. Instead, we introduce a conservative routing mechanism that adaptively chooses between complementary selectors according to the query.

3 The Proposed CROSS

Problem Formulation. Given a labeled demonstration pool $\mathcal{D}_d = \{(x_i, y_i)\}_{i=1}^{N_d}$ and a calibration dataset $\mathcal{D}_c = \{(q_j, y)j)\}_{j=1}^{N_c}$, where x_i denotes a candidate demonstration and q_j denotes a calibration query, and y_i represents its corresponding ground-truth target. $\mathbf{M}$ stands for a fixed language model. For a generic query q, our objective is to learn a query-dependent demonstration selection selector π:

$$\mathcal{S}_\pi(q) = \pi(q; \mathcal{D}_d), \qquad \mathcal{S}_\pi(q) \subseteq \mathcal{D}_d, \quad |\mathcal{S}_\pi(q)| \le K, \tag{1}$$

where K is the demonstration budget and $\mathcal{S}_\pi(q)$ denotes the subset of demonstrations selected from $\mathcal{D}_d$ for the query q. Subsequently, the selected demonstrations are combined with the query to construct an in-context prompt $\Phi(\mathcal{S}(q), q)$, which is fed into the fixed LLM to produce the prediction:

$$\widetilde{y}(q) = \mathbf{M}\big(\Phi(\mathcal{S}_\pi(q), q)\big). \tag{2}$$

3.1 Framework Overview

In this work, we address the problem of demonstration selection for in-context learning through conservative routing between complementary demonstration selectors. Specifically, CROSS learns an unsupervised mapping into hyperbolic space by distilling local neighborhood structures from the target LLM's representation space, enabling structure-aware demonstration retrieval based on geodesic distance. To balance relevance and diversity, CROSS constructs a query-conditioned DPP kernel and performs determinant-based subset selection. Since the two selectors exhibit complementary strengths across queries, we introduce a conservative routing mechanism that switches between them only when sufficient evidence favors one selector, yielding more query-adaptive demonstration selection. Additionally, an overview of CROSS is illustrated in Figure 2.

3.2 Target-aligned Hyperbolic Neighborhood Distillation

Although frozen LLM embeddings provide rich semantic representations, their Euclidean geometry may be suboptimal for demonstration selection. To better capture multi-scale and branching semantic structures, we map the frozen LLM embeddings into the Poincaré ball model while preserving local semantic neighborhood structures from the original space (Liu et al., 2025c). We then use hyperbolic geodesic distances to guide demonstration selection.

Target-aligned Hyperbolic Projection. To align semantic representations with demonstration selection, we first obtain demonstration embeddings via masked mean pooling and then project

them into hyperbolic space with neighborhood regularization. Given a demonstration (x_i, y_i), we feed the raw text sequence into the frozen language model $\mathbf{M}$ and extract the final-layer hidden states $\mathbf{H}_i = [\mathbf{h}_1, \ldots, \mathbf{h}_{T_i}]$, where $\mathbf{h}_t \in \mathbb{R}^d$ and T_i denotes sequence length. The demonstration is calculated as:

$$\mathbf{e}_i = \frac{\sum_{t=1}^{T_i} m_t \mathbf{h}_t}{\sum_{t=1}^{T_i} m_t}, \tag{3}$$

where $m_t \in \{0, 1\}$ denotes the attention mask at position t. Subsequently, we apply a learnable projection p_ϕ to map each embedding into the tangent space at the origin of the Poincaré ball:

$$\mathbf{u}_i = p_\phi(\mathbf{e}_i) \in T_{\mathbf{0}}\mathbb{D}_c^d. \tag{4}$$

We bound the norm of $\mathbf{u}_i$ to prevent excessively large tangent-space representations. Following Eq. 25, we map $\mathbf{u}_i$ onto the Poincaré ball through the exponential map at the origin, yielding $\mathbf{z}_i^c \in \mathbb{D}_c^d$. Throughout this process, we keep $\mathbf{M}$ frozen and optimize only the projection parameters p_ϕ.

Local Semantic Neighborhood Distillation. Here, we characterize the local geometry of the frozen LLM embeddings using pairwise similarity, treating the original embedding space as the teacher and distilling its local neighborhood structure into hyperbolic space. For each anchor demonstration i, we select its k_{nbr} most similar neighbors, where $k_{\text{nbr}} = 8$ is the distillation neighborhood size. This parameter is independent of the demonstration budget K, which controls the number of demonstrations included in an ICL prompt. We represent the teacher affinities as a soft distribution:

$$p_{i,j} = \frac{\exp(s_{i,j}/\tau_1)}{\sum_{l \in \mathcal{N}_i} \exp(s_{i,l}/\tau_1)}, \quad j \in \mathcal{N}_i = \text{Top}_{k_{\text{nbr}},\, j \neq i}(s_{i,j}). \tag{5}$$

Here, $s_{i,j} = \text{sim}(\mathbf{e}_i, \mathbf{e}_j)$ denotes the cosine similarity between demonstrations i and j, and $\tau_1 > 0$ is the teacher temperature. Using the same teacher-defined neighborhood $\mathcal{N}_i$, we define the hyperbolic affinity as $s_{i,j}^c = -d_c(\mathbf{z}_i^c, \mathbf{z}_j^c)$ and construct the corresponding soft distribution with student temperature $\tau_2 > 0$:

$$p_{i,j}^c = \frac{\exp(s_{i,j}^c/\tau_2)}{\sum_{l \in \mathcal{N}_i} \exp\left(s_{i,l}^c/\tau_2\right)}, \quad j \in \mathcal{N}_i. \tag{6}$$

To learn a target-aligned hyperbolic projection, we optimize p_ϕ by minimizing the KL divergence between the teacher and hyperbolic neighborhood distributions:

$$\mathcal{L}_{\text{nbr}} = \frac{1}{B} \sum_{i=1}^{B} D_{\text{KL}}(\mathbf{p}_i \| \mathbf{p}_i^c), \tag{7}$$

where $\mathbf{p}_i = \{p_{i,j}\}_{j \in \mathcal{N}_i}$ and $\mathbf{p}_i^c = \{p_{i,j}^c\}_{j \in \mathcal{N}_i}$, and B stands for the minibatch size. This regularization encourages the hyperbolic projection to preserve local semantic neighborhoods induced by the frozen LLM, thereby better capturing their complex structure. Given a query embedding $\mathbf{e}(q)$, we obtain its hyperbolic representation $\mathbf{z}^c(q) = \exp_{\mathbf{0}}^c(p_\phi(\mathbf{e}(q)))$, and then retrieve demonstrations according to their geodesic distances to the query:

$$\mathcal{S}^c(q) = \underset{\mathcal{S} \subseteq \mathcal{D}_d, |\mathcal{S}| = K}{\arg\min} \sum_{i \in \mathcal{S}} d_c(\mathbf{z}^c(q), \mathbf{z}_i^c). \tag{8}$$

Equivalently, $\mathcal{S}^c(q)$ consists of the K demonstrations with the smallest Poincaré geodesic distances to the query. In this way, semantic neighborhoods induced by the target LLM are transferred into a learnable hyperbolic space and directly used for query-dependent demonstration selection.

3.3 Complementary Diversity-oriented Demonstration Selection

The hyperbolic selector emphasizes pointwise proximity between the query and individual demonstrations. To complement this local relevance focus, we construct a DPP-based selector that jointly accounts for query relevance and set-level diversity (Xin et al., 2026).

Query-conditioned Kernel Construction. Given a query q, let $\mathbf{e}(q)$ and $\mathbf{e}_i$ denote the normalized embeddings of the query and demonstration $i \in \mathcal{D}_d$, respectively. To balance effectiveness and

efficiency, we first retrieve the top 50 demonstrations according to query-candidate relevance, forming a candidate set $\mathcal{C}(q) \subseteq \mathcal{D}_d$, and define the query-dependent quality score as:

$$r_i(q) = \exp\left(\frac{\mathrm{sim}(\mathbf{e}(q), \mathbf{e}_i) - \max_j \mathrm{sim}(\mathbf{e}(q), \mathbf{e}_j)}{2\tau}\right), \tag{9}$$

where $j \in \mathcal{C}(q)$. Accordingly, the pairwise similarity between candidates is defined as $G_{i,j} = \frac{\mathrm{sim}(\mathbf{e}_i, \mathbf{e}_j)+1}{2}$. Similar to Eq. 28, we then construct the query-conditioned DPP kernel as $\mathbf{L}(q) = \mathrm{diag}(\mathbf{r}(q))\mathbf{G}\,\mathrm{diag}(\mathbf{r}(q))$. The resulting kernel jointly promotes query relevance and pairwise diversity among selected demonstrations.

DPP-based Subset Selection. Given a demonstration budget K, the diversity-aware selector chooses a subset from $\mathcal{C}(q)$ for query q by solving:

$$\mathcal{S}(q) \in \underset{\mathcal{S}\subseteq\mathcal{C}(q),\,|\mathcal{S}|=K}{\arg\max} \det\left((\mathbf{L}(q))_{\mathcal{S}}\right). \tag{10}$$

We approximately solve this fixed-cardinality DPP MAP problem using greedy selection. Unlike hyperbolic nearest-neighbor retrieval, which focuses on individual query-demonstration proximity, this selector operates at the set level and explicitly balances relevance with intra-set diversity.

3.4 Conservative Routing over Complementary Selectors

The hyperbolic and DPP selectors exhibit complementary retrieval preferences: the former emphasizes structure-aware semantic proximity to the query, whereas the latter jointly accounts for query relevance and set-level diversity. Since neither preference is uniformly optimal across queries, we hereby introduce a conservative router that adaptively chooses between their predictions.

Prediction-relative Routing Evidence. Given a query q, let $\mathcal{S}^c(q)$ and $\mathcal{S}(q)$ denote the demonstration sets returned by the hyperbolic and DPP selectors, respectively. We then obtain the corresponding predictions for both sets using the same fixed language model and in-context inference pipeline:

$$\widehat{y}^c(q) = \mathbf{M}\big(\Phi(\mathcal{S}^c(q), q)\big), \qquad \widehat{y}(q) = \mathbf{M}\big(\Phi(\mathcal{S}(q), q)\big), \tag{11}$$

where $\mathbf{M}$ denotes the fixed target language model, and $\Phi(\mathcal{S}(q), q)$ constructs an ICL prompt by combining the selected demonstrations with query q using a fixed template. Due to their complementary retrieval biases, the two selectors may produce different predictions for the same query. For queries on which the two selectors disagree, we construct a compact routing representation relative to the two competing predictions:

$$\mathbf{f}(q) = [\mathbf{f}_{\mathrm{sim}}(q); \mathbf{f}_{\mathrm{dist}}(q)], \tag{12}$$

where $[\cdot;\cdot]$ denotes the concatenation for the two vectors. $\mathbf{f}_{\mathrm{sim}}(q)$ measures the relative semantic compatibility of the query with the two competing predictions, while $\mathbf{f}_{\mathrm{dist}}(q)$ characterizes how strongly the selected demonstrations support each prediction.

Utility-calibrated Selective Routing. Using the calibration set $\mathcal{D}_c$, we train a lightweight logistic router $s_\theta(f(q)) \in [0, 1]$ with ℓ_2 regularization. We denote the inverse regularization strength by $C > 0$, with $C = 0.1$ by default; larger values of C correspond to weaker regularization. A larger router score indicates stronger evidence that the hyperbolic selector should be preferred over the DPP selector. We then introduce a calibrated threshold $\tau^\star$ to determine whether a disagreement warrants switching from the DPP prediction to the hyperbolic prediction:

$$\widetilde{y}(q) = \begin{cases} \widehat{y}^c(q), & \widehat{y}^c(q) \neq \widehat{y}(q) \ \wedge \ s_\theta(\mathbf{f}(q)) \geq \tau^\star, \\ \widehat{y}(q), & \text{otherwise.} \end{cases} \tag{13}$$

Here, $\tau^\star$ is a calibrated gating threshold given by:

$$\tau^\star = \underset{\tau}{\arg\max}\,[R(\tau) - D(\tau)], \tag{14}$$

where

$$R(\tau) = \sum_{(q_j, y_j)\in\mathcal{D}_c} \mathbb{I}\left[s_\theta(\mathbf{f}(q_j)) \geq \tau\right] \mathbb{I}\left[\widehat{y}^c(q_j) = y_j\right] \mathbb{I}\left[\widehat{y}(q_j) \neq y_j\right], \tag{15}$$

and

$$D(\tau) = \sum_{(q_j, y_j) \in \mathcal{D}_c} \mathbb{I}\left[s_\theta(\mathbf{f}(q_j)) \geq \tau\right] \mathbb{I}\left[\widehat{y}^c(q_j) \neq y_j\right] \mathbb{I}\left[\widehat{y}(q_j) = y_j\right] \tag{16}$$

count the rescued and damaged predictions induced by switching from the DPP selector to the hyperbolic selector, respectively. This asymmetric decision rule retains the diversity-aware DPP selector as the default while selectively exploiting the complementary strengths of hyperbolic retrieval.

3.5 Summarization of CROSS

In this work, CROSS integrates a hyperbolic selector and a DPP selector with conservative routing. More specifically, we map the frozen LLM embeddings into the hyperbolic space with target-aligned hyperbolic neighborhood distillation, and then retrieve demonstrations based on the geodesic distances. To complement the local relevance focus, we then construct a DPP-based selector to balance relevance and set-level diversity. Subsequently, we maintain a compact routing representation relative to the two prections, respectively, and then employs a lightweight router decides whether to override the default DPP prediction. Furthermore, the switching threshold is calibrated to maximize the net gain between rescued and damaged predictions. A detailed pipeline is summarized in Algorithm 1.

3.6 Theoretical Analysis

We provide theoretical support for the three main components of CROSS: target-aligned hyperbolic neighborhood distillation, DPP-based diversity selection, and conservative routing. Additionally, detailed proofs of the following reults are deferred to Appendix B.

Theorem 3.1 (Proof in Appendix B.1)**.** *For every minibatch and any $\tau_1, \tau_2 > 0$, we have*

$$\frac{1}{B}\sum_{i=1}^{B} \sup_{A \subseteq \mathcal{N}_i} \left| \sum_{j \in A} (p_{i,j} - p^c_{i,j}) \right| \leq \sqrt{\frac{\mathcal{L}_{\text{nbr}}}{2}}. \tag{17}$$

In particular, the same-label neighborhood mass satisfies

$$\frac{1}{B}\sum_{i=1}^{B} \sum_{j \in \mathcal{N}_i, y_j = y_i} p^c_{i,j} \geq \frac{1}{B}\sum_{i=1}^{B} \sum_{j \in \mathcal{N}_i, y_j = y_i} p_{i,j} - \sqrt{\frac{\mathcal{L}_{\text{nbr}}}{2}}. \tag{18}$$

Remark 3.1. *Theorem 3.1 shows that minimizing $\mathcal{L}_{nbr}$ tightens a uniform bound on local neighborhood distortion, helping preserve the teacher's local label-consistent mass without supervision.*

Theorem 3.2 (Proof in Appendix B.2)**.** *For any $\tau > 0$, the DPP matrices satisfy $\mathbf{G} \succeq 0$ and $\mathbf{L}(q) \succeq 0$. For every $S \subseteq \mathcal{C}(q)$ with $|S| = K$,*

$$\det\big(\mathbf{L}(q)_S\big) = \left(\prod_{i \in S} r_i(q)^2\right) \det(\mathbf{G}_S). \tag{19}$$

For an intermediate set S with nonsingular $\mathbf{G}_S$ and $i \in \mathcal{C}(q) \setminus S$,

$$\frac{\det\big(\mathbf{L}(q)_{S \cup \{i\}}\big)}{\det\big(\mathbf{L}(q)_S\big)} = r_i(q)^2\big(1 - \mathbf{G}_{i,S}\mathbf{G}_S^{-1}\mathbf{G}_{S,i}\big) \in [0, r_i(q)^2]. \tag{20}$$

Remark 3.2. *Theorem 3.2 explains why the DPP selector complements hyperbolic nearest-neighbor retrieval. A candidate with slightly lower query relevance can still be selected if it contributes sufficiently more novelty relative to the demonstrations already chosen. Appendix B.2 gives the explicit comparison condition.*

Theorem 3.3 (Proof in Appendix B.3)**.** *Let $m = \dim(\mathbf{f}(q))$. Conditional on $\mathcal{D}_d$ and the upstream feature maps, assume $N_c \geq m + 1$ independent calibration records drawn from the test-query law. For an affine logistic router s_θ and $0 < \tau^\star < 1$, with probability at least $1 - \delta$, $\delta \in (0, 1)$,*

$$\mathbb{P}\{\widetilde{y}(q) \neq y\} \leq \frac{\sum_{j=1}^{N_c} \mathbb{I}\{\widehat{y}(q_j) \neq y_j\} - R(\tau^\star) + D(\tau^\star)}{N_c} + \varepsilon_{N_c}(\delta), \tag{21}$$

where

$$\varepsilon_{N_c}(\delta) = \sqrt{\frac{2(m+1)\log\{eN_c/(m+1)\}}{N_c}} + \sqrt{\frac{\log(1/\delta)}{2N_c}}. \tag{22}$$

Remark 3.3. *Theorem 3.3 bounds the prediction risk of the router, implying that maximizing $R(\tau) - D(\tau)$ reduces the empirical error while the calibration penalty $\varepsilon_{N_c}(\delta)$ vanishes as N_c grows.*

Table 1: Classification accuracy across five benchmarks. Best and second-best results are shown in **bold** and underlined, respectively. Values denote mean ± standard deviation over three seeds.

Method	BANKING77	CLINC150	HWU64	MTOP-en	MASSIVE en-US	Avg.
`Qwen2.5-7B Instruct`						
Hyperbolic Top-K	75.88±0.16	81.21±0.22	80.39±0.32	81.45±0.17	76.37±0.14	79.06±0.20
DPP	81.27±0.13	85.56±0.16	80.24±0.67	83.53±0.14	77.81±0.25	81.68±0.27
UCS	81.13±0.10	85.85±0.11	79.83±0.57	83.48±0.19	78.00±0.27	81.66±0.25
AutoSelect	71.28±0.76	86.03±0.84	76.89±0.63	82.74±0.80	75.60±0.49	78.51±0.70
RDES	80.19±0.20	87.36±0.15	80.73±0.23	82.82±0.68	77.26±0.46	81.67±0.34
D.Va	80.42±0.20	88.19±0.03	81.38±0.33	83.00±0.19	78.42±0.31	82.28±0.21
MLSM	76.49±0.28	84.23±0.16	78.35±0.34	82.25±0.53	77.16±0.13	79.69±0.29
Gradient Selection	52.37±2.68	75.85±1.50	69.30±0.92	62.56±2.19	65.68±4.24	65.15±2.31
TopK+ConE	79.75±0.50	87.60±0.04	80.58±0.58	82.28±0.18	76.99±0.54	81.44±0.37
MMR	79.74±0.51	86.01±0.15	80.55±0.30	83.17±0.40	78.09±0.08	81.51±0.29
VoteK	50.87±0.18	71.33±0.07	66.02±0.51	51.04±0.28	63.37±0.10	60.52±0.23
CROSS (Ours)	**84.47**±0.22	**88.76**±0.54	**85.38**±0.84	**87.50**±0.41	**82.87**±0.37	**85.80**±0.48
`Mistral-7B Instruct-v0.3`						
Hyperbolic Top-K	87.11±0.05	94.63±0.03	89.31±0.00	93.43±0.01	85.24±0.09	89.94±0.04
DPP	91.81±0.04	96.80±0.02	90.09±0.04	94.39±0.06	88.40±0.06	92.30±0.04
UCS	91.24±0.02	96.89±0.03	90.09±0.04	**95.62**±0.01	88.43±0.09	92.45±0.04
AutoSelect	82.24±0.23	91.41±0.51	86.03±0.18	91.69±0.25	82.90±0.32	86.85±0.30
RDES	92.11±0.05	95.41±0.06	89.99±0.12	94.48±0.00	87.86±0.21	91.97±0.09
D.Va	92.21±0.05	95.47±0.03	90.27±0.04	94.80±0.00	87.76±0.00	92.10±0.02
MLSM	90.91±0.19	95.36±0.02	90.09±0.00	95.39±0.09	87.33±0.17	91.82±0.09
Gradient Selection	63.30±0.50	72.28±0.01	77.83±1.63	79.88±1.68	74.31±0.05	73.52±0.77
TopK+ConE	92.07±0.03	95.53±0.00	90.64±0.04	95.09±0.01	87.39±0.00	92.14±0.02
MMR	91.39±0.04	96.30±0.02	90.24±0.08	**95.62**±0.00	87.73±0.00	92.26±0.03
VoteK	63.25±0.05	72.53±0.10	78.47±0.22	78.29±0.02	71.79±0.06	72.87±0.09
CROSS (Ours)	**92.31**±0.17	**96.99**±0.03	**90.68**±0.32	95.44±0.02	**89.04**±0.10	**92.89**±0.13
`Llama-3.1-8B Instruct`						
Hyperbolic Top-K	74.91±0.04	77.93±0.14	74.47±0.58	89.93±0.12	71.25±0.10	77.70±0.20
DPP	83.34±0.07	83.12±0.04	80.17±0.09	**93.74**±0.05	77.56±0.18	83.59±0.09
UCS	80.93±0.48	78.65±0.23	69.02±0.43	93.14±0.14	77.28±0.15	79.80±0.29
AutoSelect	72.59±0.71	76.53±0.19	75.03±0.75	86.35±0.97	74.67±0.56	77.03±0.64
RDES	82.99±0.11	81.31±0.06	80.01±0.80	91.47±0.15	78.27±0.18	82.81±0.26
D.Va	82.50±0.12	80.13±0.11	**82.46**±0.51	92.01±0.13	80.55±0.02	83.53±0.18
MLSM	78.98±0.22	81.15±0.21	78.90±0.42	91.47±0.30	76.40±0.24	81.38±0.28
Gradient Selection	58.12±1.06	60.79±2.88	68.12±0.59	73.75±1.22	69.96±0.92	66.15±1.33
TopK+ConE	83.01±0.07	82.14±0.17	77.38±0.22	92.45±0.09	76.51±0.18	82.30±0.15
MMR	83.26±0.10	81.86±0.03	76.45±0.09	92.76±0.10	75.58±0.36	81.98±0.14
VoteK	55.82±0.29	62.22±0.14	67.19±0.40	64.99±0.22	61.75±0.35	62.39±0.28
CROSS (Ours)	**85.04**±0.09	**85.09**±0.25	82.37±0.68	**93.74**±0.05	**80.65**±0.29	**85.38**±0.27

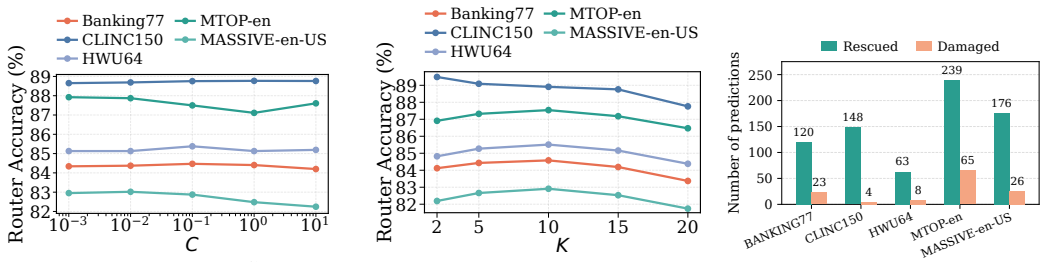
4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmark Datasets. We evaluate on five fine-grained intent classification benchmarks: BANKING77 (Casanueva et al., 2020), CLINC150 (Larson et al., 2019), HWU64 (Liu et al., 2019), MTOP-English (Li et al., 2021), and MASSIVE-en-US (FitzGerald et al., 2023). For each dataset, the official training split is partitioned into a demonstration pool and a calibration set, which are used for demonstration retrieval and router calibration, respectively. Final performance is evaluated on the official test split. More details are illustrated in Appendix C.1.

Compared Baselines. We compare our complete method with its two constituent retrieval branches, Hyperbolic Top-K and DPP (Ye et al., 2023), to isolate the contribution of query-dependent routing. We further compare against a broad set of demonstration selection methods, including UCS and UCS+DPP (Xin et al., 2026), AutoSelect (Qi et al.), RDES (Wang et al., 2025b), D.Va (Zhang et al., 2025b), MLSM (Liu et al., 2025a), Gradient Selection (Zhang et al., 2025a), ConE/TopK+ConE (Peng et al., 2024), MMR (Kapuriya et al., 2025), and VoteK (Su et al., 2023).We use intent classification accuracy as the primary metric and report mean ± standard deviation over three random seeds. Additional implementation details are provided in Appendix C.3.

Implementation Details. We use Qwen2.5-7B-Instruct (Yang et al., 2024), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), and Mistral-7B-Instruct-v0.3 (Mistral AI, 2024) as target LLM backbones,



(a) Routed accuracy vs. C (b) Accuracy vs. budget K (c) Rescued vs. damaged

Figure 3: Parameter sensitivity and router behavior on Qwen2.5-7B-Instruct. **Left (a):** Routed accuracy under different inverse regularization strengths C with five-fold calibration. **Middle (b):** Accuracy of CROSS under different demonstration budgets K on five benchmarks. **Right (c):** Rescued and damaged switches, which respectively correct DPP errors and overturn correct DPP predictions.

with a 10-shot demonstration budget in all main experiments. The hyperbolic projection outputs 64-dimensional and is optimized with AdamW at a learning rate of 1×10^{-3}. For DPP, we employ `all-mpnet-base-v2` as the sentence encoder to retrieve 50 candidates, from which the DPP select 10 with a scaling factor of 0.1.

4.2 Empirical Results

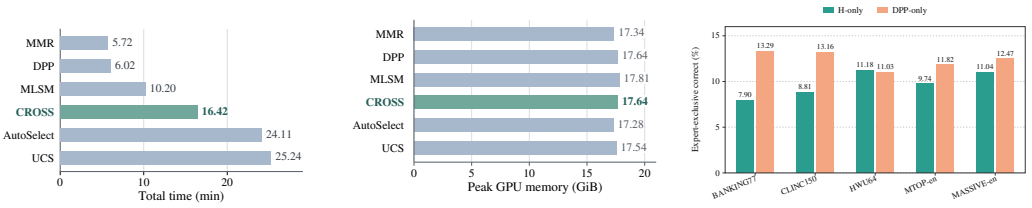
Performance Comparison. We compare CROSS with representative demonstration selection baselines across five fine-grained intent classification benchmarks and three LLM backbones. Table 1 leads to the following insights: ***Obs.❶: The routing gains are consistent across backbones and datasets.*** Our router achieves the best average performance on all three backbones, ranking best or tied for best on 13 of the 15 backbone–dataset combinations and second on the remaining two. ***Obs. ❷: Selective routing improves upon DPP even though the auxiliary retriever is weaker on average.*** Hyperbolic Top-K underperforms DPP on average for every backbone, yet selectively routing queries to this expert enables our CROSSto outperform the DPP default in all three cases. On Qwen2.5-7B-Instruct, the router improves average accuracy over DPP by 4.12 points, with similar gains across the other backbones. These results highlight the hyperbolic expert as a complement to DPP, motivating query-dependent conservative routing.

Ablation Study. To examine the contribution of each component, we construct four variants of CROSS with Qwen2.5-7B-Instruct: V1 replaces the learned switching threshold with a fixed threshold of 0.9, V2 removes retrieved-set support features, V3 removes prototype similarity features, and V4 retains only plurality-based retrieved-set statistics.Table 2 reveals several key findings: ***Obs. ❸: Calibrated switching is essential for conservative routing.*** Replacing the calibrated threshold in V1 causes the largest average accuracy drop of 2.77 percentage points, with particularly pronounced losses on HWU64 and MTOP-English. This result highlights the importance of calibrating the switching criterion: estimating which expert is preferable must be coupled with deciding whether the evidence is strong enough to depart from the DPP default. ***Obs. ❹: Prototype evidence and retrieved-set support provide complementary routing signals.*** Removing retrieved-set support in V2 and prototype similarity in V3 reduces average accuracy by 1.56 and 1.74 percentage points, respectively. Together, these results support combining prototype-level similarity with retrieved-set support to better inform query-dependent expert selection.

Table 2: Ablation results on Qwen2.5-7B-Instruct.

Variant	BANK77	CLINC150	HWU64	MTOP-en	MASSIVE en-US
V1	83.01$^{\downarrow 1.46}$	86.81$^{\downarrow 1.95}$	81.23$^{\downarrow 4.15}$	83.53$^{\downarrow 3.97}$	80.55$^{\downarrow 2.32}$
V2	82.07$^{\downarrow 2.40}$	87.06$^{\downarrow 1.70}$	83.58$^{\downarrow 1.80}$	87.30$^{\downarrow 0.20}$	81.18$^{\downarrow 1.69}$
V3	81.90$^{\downarrow 2.57}$	86.94$^{\downarrow 1.82}$	83.40$^{\downarrow 1.98}$	87.10$^{\downarrow 0.40}$	80.95$^{\downarrow 1.92}$
V4	81.79$^{\downarrow 2.68}$	87.10$^{\downarrow 1.66}$	83.36$^{\downarrow 2.02}$	85.93$^{\downarrow 1.57}$	80.69$^{\downarrow 2.18}$
CROSS	**84.47**	**88.76**	**85.38**	**87.50**	**82.87**

Parameter Sensitivity. We investigate how the router's inverse regularization strength C, the number of calibration folds, and the demonstration budget K affect routing performance. Figure 3a reports routed prediction accuracy as C varies under five-fold calibration. Figure 3b shows the accuracy of CROSS across demonstration budgets on five benchmarks. We observe that: ***Obs. ❺: CROSS maintains its advantage over DPP across the evaluated regularization and calibration settings and exhibits modest sensitivity to demonstration budget.*** Accuracy remains similar when varying C from 10^{-3} to 10^{1} and using 3-, 5-, or 10-fold calibration. Across $K \in \{2, 5, 10, 15, 20\}$, accuracy



(a) Full workflow runtime (b) Peak GPU memory (c) evidence complementarity

Figure 4: Computational cost and expert complementarity. **Left (a):** Full workflow runtime. **Middle (b):** Peak GPU memory during test generation. **Right (c):** Expert-exclusive correct predictions.

varies by at most 1.73 percentage points within each benchmark. BANKING77, HWU64, MTOP-en, and MASSIVE-en-US achieve their highest accuracy at $K = 10$, while CLINC150 peaks at $K = 2$. All five benchmarks show a modest decline from $K = 10$ to $K = 20$, indicating that additional demonstrations do not necessarily improve prediction accuracy. These results support $K = 10$ as an effective default across benchmarks. Complete sensitivity results are provided in Appendix E.

Router Behavior. To understand how CROSS benefits from the two retrieval experts, we examine their prediction complementarity and routed accuracy in Figure 4c and the router's switching outcomes in Figure 3c. We observe that: ***Obs. 6: CROSS exploits complementary expert predictions while limiting harmful switches.*** Hyperbolic Top-K uniquely answers 7.90%–11.18% of test evaluations correctly across the five datasets, providing opportunities to recover DPP errors. On Qwen2.5-7B-Instruct, beneficial switches substantially outnumber harmful ones across all five datasets. For example, on CLINC150, CROSS corrects an average of 148 DPP errors while introducing only 4 per run, improving accuracy from 85.56% to 88.76%. This indicates that selective switching translates expert complementarity into accuracy gains over the DPP default. Complete task-label composition and prediction complementarity statistics are provided in Appendix E.3. Additional router statistics and local-ambiguity analyses are provided in Appendix E.4.

Efficiency Analysis. As shown in Figure 4a, CROSS completes the full BANKING77 workflow in 16.42 minutes, requiring 2.73× the DPP runtime while remaining faster than AutoSelect and UCS. Its peak GPU memory during test generation is 17.64 GiB, matching DPP (Figure 4b). Together with the accuracy gains, these results show that CROSS exploits complementary experts at a moderate additional runtime cost while maintaining the same measured peak GPU memory as DPP. Detailed timing results are provided in Appendix E.5.



Figure 5: Case study. An HWU64 case in which label-consistent hyperbolic demonstrations allow the router to correct the DPP prediction.

Case Study. Figure 5 illustrates a successful routing case on HWU64 with Qwen2.5-7B-Instruct. Hyperbolic Top-K retrieves label-consistent demonstrations, whereas DPP introduces competing label evidence. The router selects the correct hyperbolic prediction, showing how CROSS exploits expert complementarity to recover errors made by the DPP default in this representative classification example.

5 Conclusion

This paper investigates how demonstration selection can accommodate heterogeneous query-level needs while limiting harmful switches between complementary selectors. Towards this end, we propose CROSS, a conservative routing framework that combines target-aligned hyperbolic retrieval with DPP-based selection to capture complementary preferences for local relevance and set-level diversity. By learning prediction-relative routing evidence and calibrating the switching threshold to balance rescued and damaged predictions, CROSS selectively exploits the strengths of each selector to provide complementary semantics for demonstration selection. Experiments across five intent classification benchmarks and three LLM backbones demonstrate that the effectiveness of CROSS is consistent across diverse settings, highlighting its contribution to in-context learning.

AI USE STATEMENT

This draft was prepared with generative AI assistance for manuscript organization and language editing. All technical claims, experimental results, citations, and final wording should be verified by the authors before submission.

ETHICS STATEMENT

This work studies demonstration selection on established intent classification benchmarks and does not introduce new human subject data collection. The present experiments are intended for benchmark analysis rather than direct deployment in user facing systems. Any deployment of intent classification systems should separately evaluate privacy, domain shift, calibration, and potential performance disparities across user populations.

REPRODUCIBILITY STATEMENT

The main text specifies the data split, representation learning objective, retrieval procedures, router training protocol, decision rule, main results, ablations, and sensitivity analyses. Appendices C and D.3 record the implementation-level hyperparameters, prompt construction, output parsing, and threshold selection procedure. Appendix E.1 records the complete CLINC150 sensitivity grids available in the experiment document.

REFERENCES

James Anderson. *Hyperbolic geometry*. Springer Science & Business Media, 2005.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of machine learning research*, 3(Nov):463–482, 2002. URL https://jmlr.org/papers/v3/bartlett02a.html.

Tom 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.

Clément L. Canonne. A short note on an inequality between kl and tv, 2023. URL https://arxiv.org/abs/2202.07198.

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. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.nlp4convai-1.5.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024.

Laming Chen, Guoxin Zhang, and Eric Zhou. Fast greedy map inference for determinantal point process to improve recommendation diversity. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/dbbf603ff0e99629dda5d75b6f75f966-Paper.pdf.

Julian Coda-Forno, Marcel Binz, Zeynep Akata, Matt Botvinick, Jane Wang, and Eric Schulz. Meta-in-context learning in large language models. *Advances in Neural Information Processing Systems*, 36:65189–65201, 2023.

Adrian De Wynter. Is in-context learning learning? In *International Conference on Learning Representations*, volume 2026, pp. 101317–101355, 2026.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 1107–1128, 2024.

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. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.235.

Yiwei Fu, Yuxing Zhang, Chunchun Chen, Jianwen Ma, Quan Yuan, Rong-Cheng Tu, Xinli Huang, Wei Ye, Xiao Luo, and Minghua Deng. Mark: Multi-agent collaboration with ranking guidance for text-attributed graph clustering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6057–6072, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yinhan He, Wendy Zheng, Song Wang, Zaiyi Zheng, Yushun Dong, Yaochen Zhu, and Jundong Li. Hierarchical demonstration order optimization for many-shot in-context learning. *Advances in Neural Information Processing Systems*, 38:92325–92356, 2026.

Pengcheng Jiang, Lang Cao, Ruike Zhu, Minhao Jiang, Yunyi Zhang, Jiaming Shen, Jimeng Sun, and Jiawei Han. Ras: Retrieval-and-structuring for knowledge-intensive llm generation. In *International Conference on Learning Representations*, volume 2026, pp. 20867–20904, 2026.

Janak Kapuriya, Manit Kaushik, Debasis Ganguly, and Sumit Bhatia. Exploring the role of diversity in example selection for in-context learning. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2962–2966, 2025. doi: 10.1145/3726302.3730194.

Akshay Krishnamurthy, Keegan Harris, Dylan J Foster, Cyril Zhang, and Aleksandrs Slivkins. Can large language models explore in-context? *Advances in Neural Information Processing Systems*, 37:120124–120158, 2024.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2-3):123–286, 2012.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 1311–1316. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1131.

Frédéric Lavancier, Jesper Møller, and Ege Rubak. Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(4):853–877, 2015.

Itay Levy, Ben Bogin, and Jonathan Berant. Diverse demonstrations improve in-context compositional generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1401–1422, 2023.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2950–2962. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.257.

Rui Li, Guoyin Wang, and Jiwei Li. Are human-generated demonstrations necessary for in-context learning? In *International Conference on Learning Representations*, volume 2024, pp. 6227–6263, 2024.

Zhuoqun Li, Xuanang Chen, Haiyang Yu, Hongyu Lin, Yaojie Lu, Qiaoyu Tang, Fei Huang, Xianpei Han, Le Sun, and Yongbin Li. Structrag: Boosting knowledge intensive reasoning of llms via inference-time hybrid information structurization. In *International Conference on Learning Representations*, volume 2025, pp. 36107–36124, 2025.

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, 2025a. doi: 10.18653/v1/2025.acl-long.132.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. Benchmarking natural language understanding services for building conversational agents. In *Proceedings of the Tenth International Workshop on Spoken Dialogue Systems Technology*, pp. 165–183, 2019. doi: 10.1007/978-981-15-9323-9_15.

Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. Datasets for large language models: A comprehensive survey. *Artificial Intelligence Review*, 58(12):403, 2025b.

Yuanpei Liu, Zhenqi He, and Kai Han. Hyperbolic category discovery. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9891–9900. IEEE, 2025c.

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.

Man Luo, Xin Xu, Yue Liu, Panupong Pasupat, and Mehran Kazemi. In-context learning with retrieved demonstrations for language models: A survey. *arXiv preprint arXiv:2401.11624*, 2024.

Zihan Ma, Taolin Zhang, Junnan Liu, Wenwei Zhang, Minnan Luo, Songyang Zhang, and Kai Chen. Rethinking verification for llm code generation: From generation to testing. *Advances in Neural Information Processing Systems*, 38:137549–137582, 2026.

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.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

Mistral AI. Mistral-7b-instruct-v0.3. Hugging Face Model Card, 2024. URL https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3.

Francesco Maria Molfese, Luca Moroni, Luca Gioffré, Alessandro Scirè, Simone Conia, and Roberto Navigli. Right answer, wrong score: Uncovering the inconsistencies of llm evaluation in multiple-choice question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18477–18494, 2025.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72, 2025.

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.

Kiran Purohit, Venktesh V, Sourangshu Bhattacharya, and Avishek Anand. Sample efficient demonstration selection for in-context learning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49959–49982. PMLR, 13–19 Jul 2025.

Yunzhe Qi, Sirui Chen, Jiaru Zou, and Jingrui He. Auto-regressive in-context demonstration selection. In *Forty-third International Conference on Machine Learning*.

Chengwei Qin, Aston Zhang, Chen Chen, Anirudh Dagar, and Wenming Ye. In-context learning with iterative demonstration selection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7441–7455, 2024.

John G Ratcliffe. *Foundations of hyperbolic manifolds*. Springer, 2006.

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.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Sanchit Sinha, Yuguang Yue, Victor Soto, Mayank Kulkarni, Jianhua Lu, and Aidong Zhang. Maml-en-llm: Model agnostic meta-training of llms for improved in-context learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2711–2720, 2024.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. Selective annotation makes language models better few-shot learners. In *International Conference on Learning Representations*, 2023.

Evan Wang, Federico Cassano, Catherine Wu, Yunfeng Bai, William Song, Vaskar Nath, Ziwen Han, Sean Hendryx, Summer Yue, and Hugh Zhang. Planning in natural language improves llm search for code generation. In *International Conference on Learning Representations*, volume 2025, pp. 2432–2478, 2025a.

Huazheng Wang, Jinming Wu, Haifeng Sun, Zixuan Xia, Daixuan Cheng, Jingyu Wang, Qi Qi, and Jianxin Liao. Mdr: Model-specific demonstration retrieval at inference time for in-context learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4189–4204, 2024a.

Song Wang, Zihan Chen, Chengshuai Shi, Cong Shen, and Jundong Li. Mixture of demonstrations for in-context learning. *Advances in Neural Information Processing Systems*, 37:88091–88116, 2024b.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. *Advances in Neural Information Processing Systems*, 36:15614–15638, 2023.

Xubin Wang, Jianfei Wu, Yuan Yichen, Deyu Cai, Mingzhe Li, and Weijia Jia. Demonstration selection for in-context learning via reinforcement learning. In *Forty-second International Conference on Machine Learning*, 2025b.

Tong Wu, Ashwinee Panda, Jiachen Tianhao Wang, and Prateek Mittal. Privacy-preserving in-context learning for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 20005–20040, 2024.

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.

Yilin Xiao, Chuang Zhou, Qinggang Zhang, Bo Li, Qing Li, and Xiao Huang. Reliable reasoning path: Distilling effective guidance for llm reasoning with knowledge graphs. *IEEE Transactions on Knowledge and Data Engineering*, 2026.

Jiayi Xin, Xiang Li, Evan Qiang, Weiqing He, Tianqi Shang, Weijie J Su, and Qi Long. Ucs: Estimating unseen coverage for improved in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 10965–10981, 2026.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. Compositional exemplars for in-context learning. In *International Conference on Machine Learning*, pp. 39818–39833. PMLR, 2023.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuan-Jing Huang. Do large language models know what they don't know? In *Findings of the association for Computational Linguistics: ACL 2023*, pp. 8653–8665, 2023.

Jianfei Zhang, Bei Li, Jun Bai, Rumei Li, Yanmeng Wang, Chenghua Lin, and Wenge Rong. Selecting demonstrations for many-shot in-context learning via gradient matching. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11686–11704. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.findings-acl.608.

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. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.acl-long.129.

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, 2025c.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pp. 12697–12706. Pmlr, 2021.

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. Toolqa: A dataset for llm question answering with external tools. *Advances in Neural Information Processing Systems*, 36: 50117–50143, 2023.

Supplementary Materials

A Mathematical Background . . . 16
A.1 Hyperbolic Geometry . . . 16
A.2 Determinantal Point Processes . . . 16
B Proofs of Theorems . . . 16
B.1 Proofs of Theorem 3.1 . . . 16
B.2 Proofs of Theorem 3.2 . . . 17
B.3 Proof of Theorem 3.3 . . . 18
C Details of the Experimental Setup . . . 20
C.1 More Details about Datasets . . . 20
C.2 More Details about Evaluation Metrics . . . 20
C.3 More Details about Baseline Methods . . . 21
C.4 Optimization Procedure . . . 22
D Case Studies and Prompt Templates . . . 23
D.1 Beneficial Routing Cases . . . 23
D.2 Analysis of Retrieval and Routing Evidence . . . 24
D.3 Prompt Template . . . 26
E Additional Experimental Results . . . 27
E.1 Sensitivity Grids . . . 27
E.2 Demonstration-Budget Sensitivity . . . 28
E.3 Task-Label Composition and Prediction Complementarity . . . 28
E.4 Local Ambiguity and Router Behavior . . . 29
E.5 Efficiency Analysis . . . 29

A Mathematical Background

We review the hyperbolic geometry and determinantal point processes used in CROSS.

A.1 Hyperbolic Geometry

Unlike Euclidean geometry, hyperbolic geometry (Anderson, 2005; Ratcliffe, 2006) provides a natural inductive bias for representing hierarchical data: the volume of a hyperbolic ball grows exponentially with its radius, mirroring the expansion of hierarchical structures. For $c > 0$, consider the d-dimensional Poincaré ball with constant sectional curvature $-c$, whose underlying domain is

$$\mathbb{D}_c^d = \left\{ \mathbf{x} \in \mathbb{R}^d : c\|\mathbf{x}\|_2^2 < 1 \right\}. \tag{23}$$

We equip $\mathbb{D}_c^d$ with the Riemannian metric

$$\mathbf{g}_{\mathbf{x}}^c = (\lambda_{\mathbf{x}}^c)^2 \mathbf{I}_d, \qquad \lambda_{\mathbf{x}}^c = \frac{2}{1 - c\|\mathbf{x}\|_2^2}. \tag{24}$$

Given a tangent vector $\mathbf{z} \in T_{\mathbf{0}}\mathbb{D}_c^d \simeq \mathbb{R}^d$, we obtain its hyperbolic representation $\mathbf{z}^c \in \mathbb{D}_c^d$ through the exponential map at the origin:

$$\mathbf{z}^c = \exp_{\mathbf{0}}^c(\mathbf{z}) = \tanh\left(\sqrt{c}\|\mathbf{z}\|_2\right) \frac{\mathbf{z}}{\sqrt{c}\|\mathbf{z}\|_2}. \tag{25}$$

At $\mathbf{z} = \mathbf{0}$, the map is defined by continuity as $\exp_{\mathbf{0}}^c(\mathbf{0}) = \mathbf{0}$.

For two hyperbolic embeddings $\mathbf{z}_i^c, \mathbf{z}_j^c \in \mathbb{D}_c^d$, the geodesic distance induced by the Poincaré metric is

$$d_c(\mathbf{z}_i^c, \mathbf{z}_j^c) = \frac{2}{\sqrt{c}} \operatorname{artanh}\left(\sqrt{c}\left\|(-\mathbf{z}_i^c) \oplus_c \mathbf{z}_j^c\right\|_2\right), \tag{26}$$

where $\oplus_c$ denotes Möbius addition, defined as

$$\mathbf{z}_i^c \oplus_c \mathbf{z}_j^c = \frac{\left(1 + 2c\langle \mathbf{z}_i^c, \mathbf{z}_j^c \rangle + c\|\mathbf{z}_j^c\|_2^2\right)\mathbf{z}_i^c + \left(1 - c\|\mathbf{z}_i^c\|_2^2\right)\mathbf{z}_j^c}{1 + 2c\langle \mathbf{z}_i^c, \mathbf{z}_j^c \rangle + c^2\|\mathbf{z}_i^c\|_2^2\|\mathbf{z}_j^c\|_2^2}. \tag{27}$$

A.2 Determinantal Point Processes

An $\mathbf{L}$-ensemble determinantal point process (DPP) (Kulesza & Taskar, 2012; Lavancier et al., 2015) defines a probability distribution over subsets parameterized by a positive semidefinite kernel $\mathbf{L}$, which jointly encodes item-wise relevance and pairwise similarity. A common factorization of $\mathbf{L}$ is

$$\mathbf{L} = \operatorname{diag}(\mathbf{r})\mathbf{G}\operatorname{diag}(\mathbf{r}), \tag{28}$$

where r_i denotes the relevance score of demonstration i, and $\mathbf{G} \succeq \mathbf{0}$ is a similarity matrix whose entry $\mathbf{G}_{i,j}$ measures the similarity between demonstrations i and j. Given a target subset size K, DPP-based selection identifies a subset $\mathcal{S}$ by maximizing the determinant of the corresponding principal submatrix:

$$\mathcal{S}^\star = \underset{\mathcal{S}:\,|\mathcal{S}|=K}{\arg\max} \det(\mathbf{L}_{\mathcal{S}}). \tag{29}$$

Consequently, DPP-based selection favors subsets that balance relevance and diversity.

B Proofs of Theorems

B.1 Proofs of Theorem 3.1

The hyperbolic representation setting follows Liu et al. (2025c), the proof uses a Pinsker-type KL–TV argument (Canonne, 2023).

Proof. Fix i and denote $p = \mathbf{p}_i$ and $p' = \mathbf{p}_i^c$. If $p = p'$, the claim is immediate. Otherwise, let

$$A = \{j \in \mathcal{N}_i : p_j > p'_j\}, \qquad v = p(A) - p'(A) = \frac{1}{2}\sum_{j \in \mathcal{N}_i} |p_j - p'_j|. \tag{30}$$

The positive and negative parts of $p - p'$ have the same total mass, and therefore

$$\sup_{S \subseteq \mathcal{N}_i} |p(S) - p'(S)| = v. \tag{31}$$

Writing $a = p(A)$ and $b = p'(A)$, the log-sum inequality yields the lower bound

$$D_{\mathrm{KL}}(p\|p') \geq a \log \frac{a}{b} + (1-a) \log \frac{1-a}{1-b} =: F_b(a). \tag{32}$$

Because the two neighborhood distributions are produced by softmax transformations, all their entries are positive. Hence, when $p \neq p'$, the set A satisfies $0 < b < a < 1$. Since $F_b(b) = F_b'(b) = 0$ and

$$F_b''(t) = \frac{1}{t(1-t)} \geq 4, \qquad 0 < t < 1, \tag{33}$$

we obtain

$$F_b(a) = \int_b^a (a-t) F_b''(t)\, dt \geq 2(a-b)^2 = 2v^2. \tag{34}$$

Thus, for each i,

$$\sup_{S \subseteq \mathcal{N}_i} |p_i(S) - p_i^c(S)| \leq \sqrt{D_{\mathrm{KL}}(\mathbf{p}_i \| \mathbf{p}_i^c)/2}. \tag{35}$$

Averaging over i and applying the concavity of the square-root function gives

$$\frac{1}{B} \sum_{i=1}^{B} \sup_{S \subseteq \mathcal{N}_i} |p_i(S) - p_i^c(S)| \leq \sqrt{\frac{1}{2B} \sum_{i=1}^{B} D_{\mathrm{KL}}(\mathbf{p}_i \| \mathbf{p}_i^c)} = \sqrt{\frac{\mathcal{L}_{\mathrm{nbr}}}{2}}, \tag{36}$$

which proves Eq. 17. Taking $S = \{j \in \mathcal{N}_i : y_j = y_i\}$ in the preceding bound and rearranging proves Eq. 18. □

B.2 Proofs of Theorem 3.2

The proof follows the determinant factorization and greedy DPP MAP analysis of Chen et al. (2018).

Proof. For each candidate i, we set $v_i = (\mathbf{e}_i, 1)/\sqrt{2}$. Since the embeddings are normalized, $\|v_i\|_2 = 1$ and

$$v_i^\top v_j = \frac{\mathrm{sim}(\mathbf{e}_i, \mathbf{e}_j) + 1}{2} = \mathbf{G}_{i,j}. \tag{37}$$

Thus $\mathbf{G}$ is a Gram matrix and is positive semidefinite. Since $r_i(q) > 0$ for all i, the query-dependent DPP kernel $\mathbf{L}(q) = \mathrm{diag}(\mathbf{r}(q))\mathbf{G}\,\mathrm{diag}(\mathbf{r}(q))$ is a positive diagonal congruence of $\mathbf{G}$, and hence $\mathbf{L}(q) \succeq 0$.

For any $S \subseteq \mathcal{C}(q)$, multiplicativity of determinants gives

$$\det\{\mathbf{L}(q)_S\} = \left\{ \prod_{i \in S} r_i(q)^2 \right\} \det(\mathbf{G}_S). \tag{38}$$

This proves Eq. 19. Substituting the definition of $r_i(q)$ and using $|S| = K$ gives the explicit factorization

$$\det\big(\mathbf{L}(q)_S\big) = \exp\left\{ \frac{\sum_{i \in S} \mathrm{sim}(\mathbf{e}(q), \mathbf{e}_i) - K \max_{j \in \mathcal{C}(q)} \mathrm{sim}(\mathbf{e}(q), \mathbf{e}_j)}{\tau} \right\} \det(\mathbf{G}_S). \tag{39}$$

Now suppose that $\mathbf{G}_S$ is nonsingular, let V_S be the matrix whose columns are v_j, $j \in S$, and set

$$P_S = V_S \mathbf{G}_S^{-1} V_S^\top. \tag{40}$$

This is the orthogonal projector onto the span of $\{v_j : j \in S\}$. Consequently,

$$1 - \mathbf{G}_{i,S} \mathbf{G}_S^{-1} \mathbf{G}_{S,i} = v_i^\top (I - P_S) v_i = \|(I - P_S) v_i\|_2^2 \in [0, 1]. \tag{41}$$

The block determinant formula further gives

$$\det(\mathbf{G}_{S \cup \{i\}}) = \det(\mathbf{G}_S) \left(1 - \mathbf{G}_{i,S} \mathbf{G}_S^{-1} \mathbf{G}_{S,i}\right). \tag{42}$$

Combining this identity with the determinant factorization above proves Eq. 20.

The following argument explains the relevance and diversity trade-off between candidates. For two candidates $i, j \in \mathcal{C}(q) \setminus S$ with positive residual terms in Eq. 20, greedy DPP selection prefers j to i if and only if

$$\mathrm{sim}(\mathbf{e}(q), \mathbf{e}_i) - \mathrm{sim}(\mathbf{e}(q), \mathbf{e}_j) \le \tau \log\left\{ \frac{1 - \mathbf{G}_{j,S}\mathbf{G}_S^{-1}\mathbf{G}_{S,j}}{1 - \mathbf{G}_{i,S}\mathbf{G}_S^{-1}\mathbf{G}_{S,i}} \right\}. \tag{43}$$

In particular, suppose that, for some $0 < \epsilon < \eta \le 1$,

$$1 - \mathbf{G}_{i,S}\mathbf{G}_S^{-1}\mathbf{G}_{S,i} \le \epsilon < \eta \le 1 - \mathbf{G}_{j,S}\mathbf{G}_S^{-1}\mathbf{G}_{S,j}. \tag{44}$$

Then j is preferred to an at-least-as-relevant candidate i whenever

$$0 \le \mathrm{sim}(\mathbf{e}(q), \mathbf{e}_i) - \mathrm{sim}(\mathbf{e}(q), \mathbf{e}_j) < \tau \log(\eta/\epsilon). \tag{45}$$

To verify these consequences, write

$$\nu_i = 1 - \mathbf{G}_{i,S}\mathbf{G}_S^{-1}\mathbf{G}_{S,i}, \qquad a_i = \mathrm{sim}(\mathbf{e}(q), \mathbf{e}_i). \tag{46}$$

When $\nu_i, \nu_j > 0$, Eq. 20 and the definition of $r_i(q)$ imply

$$\log \frac{\det\{\mathbf{L}(q)_{S\cup\{j\}}\}}{\det\{\mathbf{L}(q)_{S\cup\{i\}}\}} = \frac{a_j - a_i}{\tau} + \log \frac{\nu_j}{\nu_i}. \tag{47}$$

The greedy step prefers j to i exactly when this log-ratio is nonnegative, which is equivalent to Eq. 43. Under Eq. 44, if Eq. 45 holds, then

$$\frac{r_i(q)^2}{r_j(q)^2} = \exp\{(a_i - a_j)/\tau\} < \eta/\epsilon. \tag{48}$$

Therefore,

$$r_i(q)^2\nu_i \le r_i(q)^2\epsilon < r_j(q)^2\eta \le r_j(q)^2\nu_j, \tag{49}$$

and Eq. 20 shows that the marginal determinant gain of j is larger than that of i, which completes the proof. □

B.3 Proof of Theorem 3.3

Proof. Write $n = N_c$ and $v = m + 1$. Let $\boldsymbol{W} = (W_1, \ldots, W_n)$ consist of independent calibration records $W_j = (q_j, y_j, \xi_j)$, where ξ_j includes any independent per-query decoding randomness. The predictions and routing features are then fixed functions of (q_j, ξ_j).

For an affine logistic score $s_\theta(x) = (1 + \exp\{-a^\top x - b\})^{-1}$ and any $\tau \in (0, 1)$, we notice

$$s_\theta(x) \ge \tau \quad \Longleftrightarrow \quad a^\top x + b - \log \frac{\tau}{1 - \tau} \ge 0. \tag{50}$$

Thus every thresholded router belongs to the affine halfspace class

$$\mathcal{G} = \big\{ x \mapsto \mathbb{I}\{a^\top x + c \ge 0\} : a \in \mathbb{R}^m,\ c \in \mathbb{R} \big\}. \tag{51}$$

For $g \in \mathcal{G}$, the prediction loss induced by Eq. 13 is

$$\ell_g(W) = \{1 - g(\mathbf{f}(q))\}\mathbb{I}\{\widehat{y}(q) \ne y\} + g(\mathbf{f}(q))\mathbb{I}\{\widehat{y}^c(q) \ne y\} \in \{0, 1\}. \tag{52}$$

Since affine halfspaces in $\mathbb{R}^m$ have VC dimension $v = m + 1$, Sauer's lemma gives

$$\Big| \big\{ (\ell_g(W_j))_{j=1}^n : g \in \mathcal{G} \big\} \Big| \le M_n := \sum_{a=0}^{v} \binom{n}{a} \le \left(\frac{en}{v} \right)^v, \qquad n \ge v, \tag{53}$$

see Lemma 6.10 and Theorem 9.3 of Shalev-Shwartz & Ben-David (2014). For a gate g, write

$$L(g) = \mathbb{E}_W \ell_g(W), \qquad \widehat{L}_n(g) = \frac{1}{n} \sum_{j=1}^{n} \ell_g(W_j), \tag{54}$$

where W is an independent fresh test record. Define

$$\Psi(\boldsymbol{W}) = \sup_{g\in\mathcal{G}}\{L(g) - \widehat{L}_n(g)\}. \tag{55}$$

Let $\boldsymbol{W}'$ be an independent copy of $\boldsymbol{W}$, and let $\sigma_1, \ldots, \sigma_n$ be independent Rademacher variables, independent of both samples. Using symmetrization, exchangeability of (W_j, W_j'), and subadditivity of the supremum, we have

$$\mathbb{E}\Psi \le \mathbb{E}_{\boldsymbol{W},\boldsymbol{W}'} \sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{j=1}^{n}\{\ell_g(W_j') - \ell_g(W_j)\} \tag{56}$$

$$= \mathbb{E}_{\boldsymbol{W},\boldsymbol{W}',\boldsymbol{\sigma}} \sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{j=1}^{n}\sigma_j\{\ell_g(W_j') - \ell_g(W_j)\} \tag{57}$$

$$\le 2\mathbb{E}_{\boldsymbol{W}}\mathbb{E}_{\boldsymbol{\sigma}} \sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{j=1}^{n}\sigma_j\ell_g(W_j). \tag{58}$$

For fixed $\boldsymbol{W}$, put $t_{g,j} = \ell_g(W_j) - 1/2$ and $\mathcal{T} = \{(t_{g,j})_{j=1}^n : g \in \mathcal{G}\}$. Then $|\mathcal{T}| \le M_n$ and $\|t\|_2^2 = n/4$ for every $t \in \mathcal{T}$. Subtracting $1/2$ only removes the common term $(2n)^{-1}\sum_j \sigma_j$, whose expectation is zero. Therefore,

$$\mathbb{E}_{\boldsymbol{\sigma}} \sup_{g\in\mathcal{G}} \frac{1}{n}\sum_j \sigma_j\ell_g(W_j) = \mathbb{E}_{\boldsymbol{\sigma}} \max_{t\in\mathcal{T}} \frac{1}{n}\sum_j \sigma_j t_j. \tag{59}$$

Independence of the signs and the inequality $\cosh(u) \le e^{u^2/2}$ imply, for every $\lambda > 0$,

$$\mathbb{E}_{\boldsymbol{\sigma}} \exp\Big(\lambda\sum_j \sigma_j t_j\Big) = \prod_j \cosh(\lambda t_j) \le \exp(\lambda^2\|t\|_2^2/2) = \exp(\lambda^2 n/8). \tag{60}$$

The log-sum-exp bound then gives

$$\mathbb{E}_{\boldsymbol{\sigma}} \max_{t\in\mathcal{T}} \frac{1}{n}\sum_j \sigma_j t_j \le \frac{1}{n\lambda}\log\sum_{t\in\mathcal{T}} \mathbb{E}_{\boldsymbol{\sigma}} \exp\Big(\lambda\sum_j \sigma_j t_j\Big) \le \frac{\log M_n}{n\lambda} + \frac{\lambda}{8}. \tag{61}$$

Choosing $\lambda = \sqrt{8\log M_n/n}$ and applying Eq. 53, we obtain

$$\mathbb{E}\Psi \le \sqrt{\frac{2\log M_n}{n}} \le \sqrt{\frac{2v\log(en/v)}{n}}. \tag{62}$$

If $\boldsymbol{W}^{(i)}$ differs from $\boldsymbol{W}$ only in record i, then the boundedness of ℓ_g and the inequality $|\sup_g A_g - \sup_g B_g| \le \sup_g |A_g - B_g|$ give

$$|\Psi(\boldsymbol{W}) - \Psi(\boldsymbol{W}^{(i)})| \le \sup_{g\in\mathcal{G}} \frac{|\ell_g(W_i) - \ell_g(W_i^{(i)})|}{n} \le \frac{1}{n}. \tag{63}$$

McDiarmid's bounded-differences inequality (Bartlett & Mendelson, 2002, Theorem 9) therefore yields

$$\mathbb{P}\{\Psi - \mathbb{E}\Psi \ge t\} \le \exp\{-2nt^2\}. \tag{64}$$

Taking $t = \sqrt{\log(1/\delta)/(2n)}$ and using Eq. 62, we obtain, on an event $\mathcal{E}_\delta$ with probability at least $1-\delta$,

$$L(g) \le \widehat{L}_n(g) + \varepsilon_n(\delta), \qquad \text{for every } g \in \mathcal{G}. \tag{65}$$

Apply Eq. 65 to the gate $\widehat{g}(x) = \mathbb{I}\{s_\theta(x) \ge \tau^\star\}$. Then $L(\widehat{g}) = \mathbb{P}\{\widetilde{y}(q) \ne y\}$ is the conditional fresh-query risk of the routed predictor. Relative to the default DPP prediction, a switch decreases the error by one on rescued cases, increases it by one on damaged cases. Hence,

$$n\widehat{L}_n(\widehat{g}) = \sum_{j=1}^{n}\mathbb{I}\{\widehat{y}(q_j) \ne y_j\} - R(\tau^\star) + D(\tau^\star). \tag{66}$$

Combining this identity with Eq. 65 proves Eq. 21. This completes the proof. □

C Details of the Experimental Setup

C.1 More Details about Datasets

We evaluate CROSS on five intent classification benchmarks: BANKING77, CLINC150, HWU64, MTOP-English, and MASSIVE-en-US. These datasets cover both single-domain and multi-domain intent classification, allowing us to assess demonstration selection across different application settings. Below, we describe each dataset.

- **BANKING77.** BANKING77 (Casanueva et al., 2020) contains user queries from the banking domain, covering 77 fine-grained intent categories. The semantic similarity between closely related banking intents makes this dataset suitable for evaluating whether selected demonstrations help distinguish subtle differences in user intent.
- **CLINC150.** CLINC150 (Larson et al., 2019) contains 150 in-scope intents across ten domains. We use its in-scope intent classification setting to evaluate demonstration selection with a large label space and queries spanning multiple domains.
- **HWU64.** HWU64 is derived from the task-oriented natural language understanding benchmark introduced by Liu et al. (2019). It contains user utterances for conversational systems. We use its intent classification setting to examine the effectiveness of demonstration selection across different types of user requests.
- **MTOP-English.** MTOP (Li et al., 2021) is a multilingual task-oriented semantic parsing benchmark. We use its English subset and formulate the task as intent classification using the intent label associated with each query. Our evaluation focuses on intent prediction rather than slot prediction or the generation of complete semantic parsing structures.
- **MASSIVE-en-US.** MASSIVE (FitzGerald et al., 2023) is a multilingual natural language understanding benchmark. We use its English-US subset, treating each user utterance and its associated intent label as the input and prediction target, respectively.

For each dataset, we partition the training inputs into a demonstration pool $\mathcal{D}_d$ and a calibration set $\mathcal{D}_c$. Calibration examples are sampled in an approximately class-stratified manner, and the remaining training examples form the demonstration pool. Specifically, examples are shuffled within each class, and classes are visited in ascending label-ID order to draw examples in a round-robin manner until the required calibration size is reached. We fix the split seed to 42 for all datasets. Table 3 summarizes the actual dataset sizes used in our experiments.

Table 3: Dataset splits used in our experiments. Training inputs are partitioned into a demonstration pool and a calibration set.

Dataset	Training	Pool	Calibration	Test
BANKING77	10,003	9,843	160	3,080
CLINC150	15,000	14,700	300	4,500
HWU64	8,954	8,826	128	1,076
MTOP-English	15,667	15,367	300	4,386
MASSIVE-en-US	11,514	11,214	300	2,974

The demonstration pool provides candidate examples for retrieval, while the calibration set is used to fit the router and select the switching threshold. The hyperbolic projection is trained on text representations from the demonstration pool without using their labels. Final performance is evaluated on the test set. Test labels are never used for projection training, demonstration selection, router fitting, or threshold selection.

C.2 More Details about Evaluation Metrics

We use intent classification accuracy as the primary evaluation metric. Given a test set containing N_{test} examples, let y_i and $\hat{y}_i$ denote the ground-truth intent and the final prediction for the i-th query, respectively. Accuracy is defined as

$$\text{Accuracy} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \mathbb{I}[\hat{y}_i = y_i] \times 100\%, \tag{67}$$

where $\mathbb{I}[\cdot]$ is the indicator function. For CROSS, $\hat{y}_i$ is the final prediction determined by the conservative routing rule. For standalone retrieval branches, it is the prediction produced by the corresponding branch.

We conduct experiments using Qwen2.5-7B-Instruct (Yang et al., 2024), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), and Mistral-7B-Instruct-v0.3 (Mistral AI, 2024). All main experiments use a demonstration budget of $K = 10$. The compared methods share the same data splits, prompt template, generation settings, and output parsing rules.

To account for downstream randomness, we use three random seeds, 42, 43, and 44, and report the mean and standard deviation of accuracy. Across the three CROSS runs, we keep the data split, frozen encoder features, trained hyperbolic projection, and demonstration selections from both branches fixed. For each downstream seed, we regenerate branch predictions and refit and calibrate the router using the corresponding calibration predictions. Thus, the reported standard deviations capture variation in downstream prediction and router fitting and calibration, rather than variation from repeating data partitioning or hyperbolic projection training.

In the main results table, Avg. denotes the unweighted mean of the five dataset-level mean accuracies. Its accompanying standard deviation entry is the unweighted mean of the five dataset-level standard deviations.

C.3 More Details about Baseline Methods

We compare CROSS with its two constituent retrieval branches and a range of demonstration selection methods. The standalone branches isolate the contribution of query-dependent routing, while the remaining baselines cover relevance, coverage, diversity, and learning-based selection strategies. We describe each method below.

- **Hyperbolic Top-K.** This baseline uses the hyperbolic retrieval branch of CROSS independently. It extracts frozen representations from the corresponding target language model, learns a projection into the Poincaré ball through neighborhood distillation, and selects the K demonstrations with the smallest hyperbolic distances to the query. Predictions are taken directly from this branch without routing. Each of the three target models uses its own hidden representations.
- **DPP.** This baseline independently uses the DPP retrieval branch of CROSS. Following the determinantal point process formulation (Kulesza & Taskar, 2012), it constructs a query-conditioned kernel to jointly account for relevance and diversity within the selected demonstration set. Specifically, we encode text with `all-mpnet-base-v2`, retrieve 50 candidates, and select 10 demonstrations using a greedy approximation, with the scale factor set to 0.1. DPP also serves as the default prediction branch in CROSS.
- **UCS.** UCS (Xin et al., 2026) evaluates demonstration sets from the perspective of latent cluster coverage. It uses cluster frequency information to estimate uncovered structure and incorporates a coverage signal into the selection objective. In our experiments, UCS refers to its combination with DPP, namely DPP+UCS. We consistently denote this variant as UCS throughout the paper.
- **AutoSelect.** AutoSelect (Qi et al.) performs demonstration selection in an autoregressive manner, progressively constructing the demonstration sequence used for prediction. We include it as a learning-based demonstration selection baseline.
- **RDES.** RDES (Wang et al., 2025b) formulates demonstration selection as a reinforcement learning problem and uses Q-learning to jointly consider query relevance and demonstration diversity. Its diversity signal incorporates the label distribution of selected demonstrations to encourage more balanced demonstration sets.
- **D.Va.** D.Va (Zhang et al., 2025b) evaluates candidate demonstrations through a validation procedure. It calibrates the validation loss according to the model's preferences between validation and test inputs, and uses the resulting scores to rank candidate demonstrations.
- **MLSM.** MLSM (Liu et al., 2025a) combines task-agnostic input similarities at different levels for demonstration selection. These multilevel matching signals characterize the relationship between candidate demonstrations and the query.
- **Gradient Selection (CLG).** This method (Zhang et al., 2025a) selects demonstrations through gradient matching, aiming to align the fine-tuning gradients of the selected subset with those of the full training set. We include it to evaluate a selection strategy based on task-gradient information.

Algorithm 1 CROSS: Conservative Routing over Complementary Demonstration Selectors

Require: Demonstration pool $\mathcal{D}_d = \{(x_i, y_i)\}_{i=1}^{N_d}$, calibration set $\mathcal{D}_c = \{(q_j, y_j)\}_{j=1}^{N_c}$, frozen language model $\mathbf{M}$, prompt function Φ, and demonstration budget K

Encode each demonstration x_i with the frozen model $\mathbf{M}$ to obtain $\mathbf{e}_i$ by Eq. 3
for each minibatch $\mathcal{B} \subseteq \mathcal{D}_d$ **do** // Target-aligned neighborhood distillation
    Construct target-aware neighborhood distributions $\mathbf{p}_i$ by Eq. 5
    Project $\mathbf{e}_i$ to the Poincaré ball by Eq. 4 and construct $\mathbf{p}_i^c$ from hyperbolic distances by Eq. 6
    Optimize the projection p_ϕ by Eq. 7 // Neighborhood-preserving hyperbolic projection
end for
for each calibration instance $(q_j, y_j) \in \mathcal{D}_c$ **do** // Conservative routing over selectors
    Retrieve $\mathcal{S}^c(q_j)$ via hyperbolic nearest-neighbor retrieval (Eq. 8) and $\mathcal{S}(q_j)$ via DPP selection (Eq. 10)
    Obtain the predictions $\widehat{y}^c(q_j)$ and $\widehat{y}_D(q_j)$ by Eq. 11
    Construct a compact routing representation for the two predictions $\mathbf{f}(q_j)$ by Eq. 12
    Optimize a lightweight router $s_\theta(\mathbf{f}(q))$ by Eq. 13 on $\mathcal{D}_c$// Calibrated selective routing
end for
return Optimized projection p_ϕ and logistic router $s_\theta(\mathbf{f}(q))$

- **ConE / TopK+ConE.** ConE (Peng et al., 2024) selects demonstrations according to how effectively they help the language model understand the test input, using model scores for the input conditioned on candidate demonstrations. TopK+ConE first retrieves a candidate set based on similarity and then applies ConE to select demonstrations.
- **MMR.** Maximal Marginal Relevance (Kapuriya et al., 2025) iteratively selects demonstrations by considering both their relevance to the query and their redundancy with previously selected examples, thereby balancing relevance and diversity.
- **VoteK.** VoteK (Su et al., 2023) is a graph-based selection method that uses neighborhood relationships among examples to identify diverse and representative samples. We include it as a representative example selection baseline.

Hyperbolic Top-K and the complete CROSS framework share the same encoder and projection configurations. We obtain text representations through masked mean pooling of the target model's final-layer hidden states, using an encoding batch size of 32. The projection has a hidden dimension of 256 and an output dimension of 64, and is trained with AdamW at a learning rate of 10^{-3}. The neighborhood size is 8, and both teacher and student temperatures are set to 0.1. We fix the curvature to $c = 1$ and the projection training seed to 42. Table 4 reports the dataset-specific encoding lengths and training schedules.

Table 4: Encoding lengths and projection training schedules for the hyperbolic branch. The encoding batch size is 32 for all datasets. Both the data split seed and the projection training seed are fixed to 42.

Dataset	Max. tokens	Projection batch size	Epochs
BANKING77	192	64	50
CLINC150	256	256	30
HWU64	256	256	30
MTOP-English	256	256	30
MASSIVE-en-US	256	256	30

C.4 Optimization Procedure

Algorithm 1 summarizes the optimization procedure of CROSS. We first train the hyperbolic projection through target-aligned neighborhood distillation on the demonstration pool. We then obtain predictions from Hyperbolic Top-K and DPP for the calibration queries and construct the corresponding routing features. These features are used to fit the lightweight logistic router, with the switching threshold selected using out-of-fold scores to maximize the net gain over the DPP default.

D Case Studies and Prompt Templates

D.1 Beneficial Routing Cases

In this section, we present three cases illustrating how CROSS exploits complementary predictions from its two retrieval branches to correct classification errors. All cases are drawn from HWU64 and CLINC150 using Qwen2.5-7B-Instruct with downstream seed 42. For each case, we report the query, ground-truth label, branch predictions, and routing decision. Here, $s_\theta(f(q))$ denotes the router score, and $\tau^\star$ is the threshold selected on the calibration set for the corresponding run. When the branches disagree and the router score meets the threshold, CROSS adopts the Hyperbolic Top-K prediction. The router score is used as a switching criterion and is not interpreted as a verified probability of prediction correctness.

Example 1: Query

mute speaker

Dataset: HWU64
Query ID: `hwu64:test:61`
Ground truth: `audio_volume_mute`

Example 1: DPP Prediction

Predicted label: `audio_volume_down`

Outcome: Incorrect. The prediction corresponds to reducing the volume rather than muting the speaker.

Example 1: Hyperbolic Prediction and Routing Decision

Hyperbolic Top-K prediction: `audio_volume_mute`

Router score: 0.718873
Switching threshold: 0.698954
Action: Switch to Hyperbolic Top-K.
Final prediction: `audio_volume_mute` **(Correct)**

Analysis. The query requests muting, whereas lowering the volume does not necessarily silence the speaker. Both branch predictions concern volume control but represent different operations. The router score exceeds the calibrated threshold, and switching to Hyperbolic Top-K corrects this fine-grained intent confusion.

Example 2: Query

put milk on my shopping list

Dataset: CLINC150
Query ID: `clinc150:test:606`
Ground truth: `shopping_list_update`

Example 2: DPP Prediction

Predicted label: `shopping_list`

Outcome: Incorrect. The prediction identifies the shopping-list query intent rather than the requested update.

Example 2: Hyperbolic Prediction and Routing Decision

Hyperbolic Top-K prediction: `shopping_list_update`

Router score: 0.850963
Switching threshold: 0.699305
Action: Switch to Hyperbolic Top-K.
Final prediction: `shopping_list_update` **(Correct)**

Analysis. The phrase "put milk on" specifies adding an item to the list, rather than inspecting its existing contents. Hyperbolic Top-K correctly identifies the update operation, and `CROSS` selects this prediction because the router score exceeds the threshold. This case illustrates disagreement between intents involving the same object but different operations. We examine its retrieved demonstrations and recoverable routing evidence in Appendix D.2.

Example 3: Query

please speak in tagalog

Dataset: CLINC150
Query ID: `clinc150:test:330`
Ground truth: `change_language`

Example 3: DPP Prediction

Predicted label: `change_accent`

Outcome: Incorrect. The prediction corresponds to changing the accent rather than the language.

Example 3: Hyperbolic Prediction and Routing Decision

Hyperbolic Top-K prediction: `change_language`

Router score: 0.853851
Switching threshold: 0.699305
Action: Switch to Hyperbolic Top-K.
Final prediction: `change_language` **(Correct)**

Analysis. The request specifies Tagalog as the desired language. DPP confuses language switching with accent switching, while Hyperbolic Top-K predicts the correct intent. The routing decision retains this distinction and corrects the default prediction.

These cases illustrate specific queries for which Hyperbolic Top-K complements DPP and the router successfully selects the correct alternative. They provide qualitative examples of beneficial switches; aggregate gains and harmful switching behavior are evaluated in the quantitative experiments.

D.2 Analysis of Retrieval and Routing Evidence

We further examine Example 2, whose query is "put milk on my shopping list." The branch predictions, router score, and final decision are taken from saved per-query records. The demonstrations are reconstructed from the original selection indices, archived data, and data-splitting code, preserving the order in the selection files. Consistency checks matched the reconstructed calibration IDs and their order, as well as all query IDs and their order, against the saved records. The original pool-ID cache was unavailable; the pool order was reconstructed from the archived training data and splitting code. The demonstrations and prompts below are therefore source-based reconstructions rather than directly saved inference input dumps.

Retrieved demonstrations. The following boxes show all ten demonstrations selected by each branch for this query.

Example 2: Hyperbolic Top-K Demonstrations

All ten demonstrations have the label `shopping_list_update`. The original selection order is retained.

1. *put bread on my shopping list*
2. *put carrots on my shopping list*
3. *put soda on the shopping list and add carrots*
4. *remove carrots from my shopping list and add soda*
5. *add soda on my shopping list and remove carrots*
6. *put rice and pudding on my shopping list cause i need them*
7. *take frozen food off my shopping list*
8. *take fresh carrots off my shopping list*
9. *take bread off my shopping list and add milk*
10. *get rid of butter on my shopping list*

Example 2: DPP Demonstrations

The original selection order is retained. Each demonstration is followed by its label.

1. *put milk on my shopping list cause i'm out of it*
   `shopping_list_update`
2. *did i put milk on the shopping list*
   `shopping_list`
3. *please put milk on my list i just used the last of it*
   `shopping_list_update`
4. *i want you to add milk on my shopping list unless it is already on it*
   `shopping_list_update`
5. *i no longer need milk, so forget that*
   `shopping_list_update`
6. *take bread off my shopping list and add milk*
   `shopping_list_update`
7. *is the milk on the shopping list if not, please add*
   `shopping_list_update`
8. *do i need to get milk*
   `shopping_list`
9. *add a reminder to pick milk up at the store*
   `reminder_update`
10. *put carrots on my shopping list*
    `shopping_list_update`

The Hyperbolic Top-K demonstrations describe additions, removals, and combinations of these operations involving different products. All ten belong to the shopping-list update intent. The DPP demonstrations frequently mention "milk," while covering shopping-list updates, shopping-list queries, and reminder updates. Their label composition is summarized below.

Example 2: Label Composition

Label	Hyperbolic	DPP
`shopping_list_update`	10	7
`shopping_list`	0	2
`reminder_update`	0	1

For this query, Hyperbolic Top-K provides a more label-concentrated demonstration set. However, the correct update label also accounts for the majority of the DPP demonstrations. The branch predictions therefore cannot be explained simply by majority voting over demonstration labels. Likewise, lexical overlap alone does not establish why DPP produced an incorrect prediction.

Recoverable support features. Let c_H denote `shopping_list_update` and c_D denote `shopping_list`. The router computes its seven support features from the Hyperbolic demonstration set. The DPP label counts above are included for comparison and do not directly enter these seven features.

Let $n(c)$ be the number of Hyperbolic demonstrations with label c, and let M_1 and M_2 be the largest and second-largest label counts. When only one label is present, $M_2 = 0$. The unstandardized support features for this case are as follows.

Example 2: Reconstructed Support Features

Feature	Definition	Value
Support for the Hyperbolic prediction	$n(c_H)/10$	1
Support for the DPP prediction	$n(c_D)/10$	0
Support difference	$[n(c_H) - n(c_D)]/10$	1
Largest label proportion	$M_1/10$	1
Gap between the two largest counts	$(M_1 - M_2)/10$	1
Hyperbolic prediction is a plurality label	$\mathbb{I}[n(c_H) = M_1]$	1
DPP prediction is a plurality label	$\mathbb{I}[n(c_D) = M_1]$	0

Support feature vector: $[1, 0, 1, 1, 1, 1, 0]$. These values are computed from the reconstructed demonstration set.

The router additionally uses five similarity features. Four compare class-specific nearest-neighbor similarity statistics over the entire demonstration pool: the maximum similarity and the mean of the top 3, 5, and 10 similarities. The fifth compares query-to-label-name similarities. All differences are computed in the direction $c_H - c_D$. Here, prototype statistics refer to within-class nearest-neighbor similarity statistics, rather than class-centroid similarities or statistics restricted to the ten retrieved demonstrations.

The five similarity features, standardization parameters, and regression coefficients were not retained in the available records. We therefore do not decompose the numerical contribution of each feature to this decision. The available evidence shows that more concentrated label support accompanies a successful switch, but does not establish that the support features alone determine it. The router uses twelve features jointly, and its saved score of 0.850963 exceeds the threshold of 0.699305, leading to the correct Hyperbolic Top-K prediction.

D.3 Prompt Template

For the CLINC150 case analyzed above, the two retrieval branches share the same task instructions, legal label list, query, and generation settings; only the retrieved demonstrations differ. The following template is reconstructed from the prompt-building code. Placeholders represent the dataset's legal labels, the ten selected demonstrations, and the query. In the actual CLINC150 input, the label list is expanded to all 150 legal labels in label-ID order.

Intent Classification Prompt Template

```
You are an assistant for intent classification.

You are an assistant for intent classification.
Your task is to assign EXACTLY ONE label to each query.
Answer STRICTLY in the format: 'Label: <intent_name>'.

Available labels: <label_1>, <label_2>, ..., <label_C>

Here are some labeled examples:

Example:
Text: <demo_text_1>
Label: <demo_label_1>

...

Example:
Text: <demo_text_10>
Label: <demo_label_10>


Now answer the following question.
Text: <query_text>
Label:
```

The repeated opening instruction results from concatenating the generation function's instruction prefix with the prompt body. We preserve this repetition and the single quotation marks around the required output format to reflect the implemented input construction. For Example 2, substituting the ten demonstrations from the corresponding branch in their listed order, expanding the legal label list, and setting the query to "put milk on my shopping list" reconstructs the input text before tokenization.

Inference uses plain-text continuation without an additional chat template. The generation batch size is 16, the maximum input length is 16,384 tokens, and the maximum number of new tokens is 16. Batches use left padding, and inputs exceeding the length limit retain the rightmost 16,384 tokens. Sampling is enabled by default with temperature 0.1 and top-p 1.0. The implementation falls back to greedy decoding for an entire batch if a specified numerical error occurs in the sampling probabilities. The available records do not establish whether this fallback was triggered for the batch containing Example 2.

The prompt reconstruction is based on the source code and recovered demonstrations; the original token sequence has not been verified by rerunning the tokenizer.

E Additional Experimental Results

This appendix reports the complete sensitivity grids and additional analyses of retrieval complementarity and router behavior. The main text presents the corresponding trends and summarizes the principal observations; the tables below provide the full per-setting and per-dataset results.

E.1 Sensitivity Grids

We first report the complete CLINC150 sensitivity grids used to examine the regularization strength C and the number of calibration folds. The three DPP baselines are 0.8564, 0.8538, and 0.8567 for downstream seeds 42, 43, and 44, respectively. Across all reported combinations, the router remains above the corresponding DPP baseline. The threshold is the calibrated switching threshold, "switches" counts router overrides, "rescued" counts corrected DPP errors, and "damaged" counts correct DPP predictions replaced by incorrect Hyperbolic predictions.

Table 5: CLINC150 sensitivity to logistic-regression regularization C and calibration folds for downstream seed 42.

C	Folds	Threshold	Switches	Rescued	Damaged	Net gain	Router acc.
0.001	3	0.5140	167	148	10	138	0.8871
0.001	5	0.5124	173	153	11	142	0.8880
0.001	10	0.5127	173	153	11	142	0.8880
0.01	3	0.5675	176	154	11	143	0.8882
0.01	5	0.5773	171	152	10	142	0.8880
0.01	10	0.5784	170	152	9	143	0.8882
0.1	3	0.7181	156	146	3	143	0.8882
0.1	5	0.6993	162	150	3	147	0.8891
0.1	10	0.7031	162	150	3	147	0.8891
1	3	0.8148	162	146	4	142	0.8880
1	5	0.7726	175	155	5	150	0.8898
1	10	0.7956	166	150	4	146	0.8889
10	3	0.6417	212	164	16	148	0.8893
10	5	0.6823	211	164	16	148	0.8893
10	10	0.7220	208	163	16	147	0.8891

Table 6: CLINC150 sensitivity to logistic-regression regularization C and calibration folds for downstream seed 43.

C	Folds	Threshold	Switches	Rescued	Damaged	Net gain	Router acc.
0.001	3	0.5087	151	139	8	131	0.8829
0.001	5	0.5082	153	140	9	131	0.8829
0.001	10	0.5120	133	125	6	119	0.8802
0.01	3	0.5589	149	137	8	129	0.8824
0.01	5	0.5518	154	142	8	134	0.8836
0.01	10	0.5770	128	121	5	116	0.8796
0.1	3	0.6460	163	153	5	148	0.8867
0.1	5	0.7088	132	127	2	125	0.8816
0.1	10	0.6769	150	143	4	139	0.8847
1	3	0.7711	160	148	2	146	0.8862
1	5	0.8882	119	114	1	113	0.8789
1	10	0.8217	148	137	2	135	0.8838
10	3	0.7818	202	166	13	153	0.8878
10	5	0.9497	159	141	2	139	0.8847
10	10	0.7274	209	169	16	153	0.8878

E.2 Demonstration-Budget Sensitivity

We next examine whether the routing gain depends on the number of in-context exemplars. Table 8 reports the CLINC150 results for $K \in \{2, 5, 10, 15, 20\}$ using downstream seed 42. The table reports the standalone Hyperbolic accuracy relative to DPP, the number of switching actions, their rescued and damaged outcomes, and the final router accuracy. The router remains effective across the evaluated budgets, although the number of damaged switches increases at the largest budget.

E.3 Task-Label Composition and Prediction Complementarity

Table 9 reports the complete task-label composition and prediction complementarity statistics of Hyperbolic Top-K and DPP, extending the compact results in Table 4c. Hyperbolic Top-K covers more unique task labels on four of the five datasets, indicating that the two selectors differ in the label composition of their retrieved demonstrations. Their expert-specific correct predictions further reveal complementary strengths: even when Hyperbolic Top-K is less accurate overall, it correctly answers queries on which DPP fails. Recovering all Hyperbolic-only correct cases would improve DPP accuracy by 7.90–11.18 percentage points across the five datasets, motivating the selective routing mechanism of `CROSS`.

Table 7: CLINC150 sensitivity to logistic-regression regularization C and calibration folds for downstream seed 44.

C	Folds	Threshold	Switches	Rescued	Damaged	Net gain	Router acc.
0.001	3	0.4979	272	194	47	147	0.8893
0.001	5	0.5101	175	155	11	144	0.8887
0.001	10	0.5125	162	148	8	140	0.8878
0.01	3	0.5626	179	157	11	146	0.8891
0.01	5	0.5655	176	157	11	146	0.8891
0.01	10	0.5823	160	147	7	140	0.8878
0.1	3	0.6603	184	165	7	158	0.8918
0.1	5	0.6552	185	166	7	159	0.8920
0.1	10	0.6936	167	157	3	154	0.8909
1	3	0.6906	203	178	7	171	0.8947
1	5	0.6671	209	179	9	170	0.8944
1	10	0.7507	191	171	5	166	0.8936
10	3	0.2890	268	191	43	148	0.8896
10	5	0.2649	273	192	47	145	0.8889
10	10	0.8343	198	167	9	158	0.8918

Table 8: CLINC150 demonstration-budget sensitivity for downstream seed 42. The DPP baseline accuracy is 0.8564.

Budget	H acc.	vs. DPP	Switches	Rescued	Damaged	Net gain	Router acc.
$K = 2$	0.8571	0.0007	178	173	0	173	0.8949
$K = 5$	0.8016	-0.0549	189	166	11	155	0.8909
$K = 10$	0.8138	-0.0427	162	150	3	147	0.8891
$K = 15$	0.8113	-0.0451	153	142	2	140	0.8876
$K = 20$	0.7884	-0.0680	314	183	88	95	0.8776

E.4 Local Ambiguity and Router Behavior

Finally, we examine whether expert preference varies systematically with local ambiguity and whether the router converts this structure into targeted gains. We define local ambiguity as the label entropy of the ten nearest examples retrieved from the demonstration pool. Table 10 reports ambiguity slopes under MPNet and an independent BERT probe together with the router's action statistics. A negative slope indicates that the relative frequency of Hyperbolic-only wins decreases as local ambiguity increases. Rescued cases substantially outnumber damaged cases across all five datasets, showing that the router predominantly switches when the alternative expert is useful.

E.5 Efficiency Analysis

We measure runtime on BANKING77 using Qwen2.5-7B-Instruct in BF16 on a single RTX 5090 with seed 42. All methods use the same 10-shot setting, generation batch size, and final classification configuration.

As shown in Table 11, `CROSS` requires 647.87 seconds of test-generation compute, approximately 2.03× that of DPP, consistent with its two-branch prediction design. The complete workflow, including representation training, calibration and testing of both branches, and routing, takes 985.46 seconds (16.42 minutes), approximately 2.73× the DPP runtime, while remaining below the total runtimes of AutoSelect, UCS, RDES, and CLG. Together with the accuracy improvements in the main experiments, these results indicate that `CROSS` makes effective use of complementary experts at a moderate additional computational cost in the evaluated setting.

Table 9: Task-label composition and prediction complementarity with Qwen2.5-7B-Instruct. H denotes Hyperbolic Top-K; uniq. and ent. denote the average number of unique task labels and label entropy within selected demonstration sets. Disagr. denotes prediction disagreement. H only and DPP only count expert-exclusive correct predictions pooled over three seeds. Oracle denotes the potential accuracy gain over DPP from recovering all H-only cases, in percentage points.

Dataset	H uniq.	DPP uniq.	H ent.	DPP ent.	Disagr.	H only	DPP only	Oracle
BANKING77	5.43	2.98	2.074	1.017	31.20%	730	1228	**7.90**
CLINC150	4.01	2.54	1.469	0.765	24.82%	1189	1777	**8.81**
HWU64	3.76	3.00	1.377	1.006	26.12%	361	356	**11.18**
MTOP-en	2.65	2.66	0.845	0.841	24.71%	1281	1555	**9.74**
MASSIVE-en-US	3.89	3.18	1.428	1.096	28.92%	985	1113	**11.04**

Table 10: Local-ambiguity slopes and router action statistics for Qwen2.5-7B-Instruct. Router accuracy and rescued cases are highlighted in light teal.

Dataset	MPNet	BERT	DPP acc.	Router acc.	Switches	Rescued	Damaged	Both wrong
BANKING77	-0.0905	-0.1800	81.27	84.42	184	120	23	41
CLINC150	-0.2169	-0.1091	85.56	88.76	160	148	4	8
HWU64	-0.1173	-0.1425	80.24	85.38	78	63	8	7
MTOP-en	-0.2324	-0.2008	83.53	87.50	329	239	65	25
MASSIVE-en-US	-0.0151	0.0382	77.81	82.87	233	176	26	31

Table 11: Single-run runtime on BANKING77, in seconds. Test gen. compute measures actual model generation time; the remaining stage times and the total are process wall-clock times.

Method	Test gen. compute	Prep./train/ select/other	Calib. gen. process	Test gen. process	Total
`Qwen2.5-7B Instruct  BANKING77  Seed 42`					
Top-K	342.74	47.40	0.00	952.60	1000.01
DPP	319.37	31.72	0.00	329.43	361.15
UCS (DPP+UCS)	308.93	1196.24	0.00	318.32	1514.56
AutoSelect	316.30	1120.54	0.00	326.10	1446.64
RDES	318.76	3044.35	0.00	327.95	3372.29
MLSM	343.64	258.86	0.00	352.96	611.82
Gradient Selection (CLG)	271.87	5937.69	0.00	281.04	6218.73
MMR	318.82	14.08	0.00	329.01	343.08
Vote-K	250.61	108.54	0.00	260.30	368.84
CROSS (Ours)	647.87	267.06	52.25	666.15	985.46