DRIFT: BREAKING DOMAIN ENTANGLEMENT IN MULTI-SOURCE OUT-OF-DISTRIBUTION IN-CONTEXT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we study multi-source out-of-distribution in-context learning (OOD ICL), where demonstrations from heterogeneous source domains are used to support inference on an unseen target distribution for Large Language Models (LLMs). We reveal that increasing source diversity can unexpectedly degrade ICL performance. To explain this phenomenon, we establish a Bayesian framework that decouples demonstration retrieval from LLM inference, and theoretically characterize how domain information entangled in retrieval representations propagates source-dependent bias into the predictive distribution. Our analysis further shows that suppressing such domain information yields a tighter bound on the divergence from the target predictive distribution. Motivated by this result, we propose a domain-disentangled retrieval framework that preserves transferable semantics while minimizing domain-specific information during demonstration selection. Experiments on multiple LLMs and OOD benchmarks consistently validate our theoretical findings and demonstrate improved robustness under multi-source shifts, establishing a new perspective for multi-source OOD ICL[1].

1 INTRODUCTION

With the rapid development of large language models (LLMs), in-context learning (ICL) has emerged as an important paradigm for adapting LLMs to downstream tasks without parameter updates (Brown et al., 2020). Despite its effectiveness, ICL is highly sensitive to the selection and distribution of demonstrations Liu et al. (2022); Rubin et al. (2022). Recent studies further reveal that when the retrieval domain differs substantially from the target domain, such distribution shifts can significantly degrade ICL performance, posing a critical challenge for robust ICL under out-of-distribution (OOD) scenarios Song et al. (2026b); Xu et al. (2026).

To address the challenges of OOD ICL, existing research has largely focused on improving the quality of in-context demonstrations. One line of work enhances demonstration retrieval by learning more discriminative retrieval representations or re-ranking retrieved candidates (Liu et al., 2025; Xu et al., 2026; Byun et al., 2026). Another direction reduces source–target discrepancies through demonstration adaptation, including semantic rewriting (Madine, 2024) and LLM-based generation of transferable exemplars (Song et al., 2026a). More recently, feedback-driven approaches exploit LLM reasoning signals or distribution proxies to estimate demonstration utility and guide context construction under target-inaccessible settings (Song et al., 2026b). Despite their differences, these methods share a common objective: constructing demonstrations that better approximate the target distribution and thereby improving OOD generalization.

However, existing OOD ICL studies have predominantly focused on a single source domain, leaving the multi-source setting largely underexplored. In practice, demonstrations are often collected from multiple heterogeneous sources with distinct distributions (Mansour et al., 2008; Peng et al., 2019a). Our preliminary study in Fig. 1 reveals a counter-intuitive phenomenon: directly retrieving demonstrations from a pooled multi-source space can underperform retrieval from an appropriate single source, despite having access to a larger and more diverse candidate pool. This observation

[1] https://anonymous.4open.science/r/DRIFT-C1EB

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.

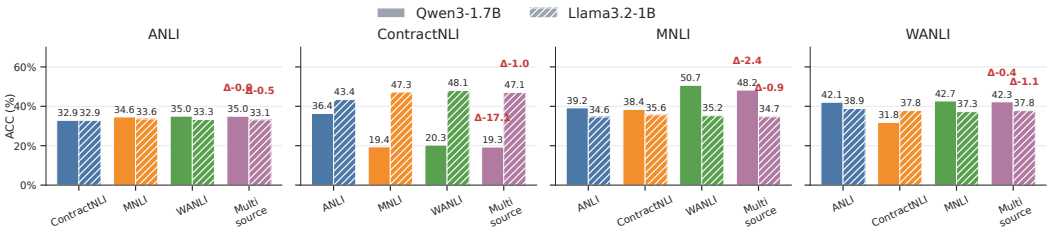


Figure 1: Preliminary study of single-source and multi-source demonstration retrieval for OOD ICL across four NLI target domains and two LLMs. Δ denotes the performance difference between multi-source retrieval and the best-performing single source, revealing the potential negative effect of source heterogeneity.

suggests that increasing source diversity does not necessarily yield more effective demonstrations and may instead introduce additional interference into the retrieval process.

To bridge this gap, we study Multi-Source Out-of-Distribution In-Context Learning (MS-OOD ICL), where demonstrations are drawn from multiple heterogeneous source domains while the target domain remains strictly inaccessible (Song et al., 2024). Building on a Bayesian view of ICL, we establish a theoretical framework that decouples demonstration retrieval from LLM inference and characterizes how multi-source heterogeneity affects the predictive distribution. Our analysis shows that domain information entangled in retrieval representations can distort demonstration selection and propagate source-dependent bias into subsequent inference, leading to a looser bound on the divergence from the target predictive distribution. Motivated by this analysis, we propose **DRIFT**, a **D**omain-disentangled **R**etrieval framework for **I**n-context learning under multi-source distribution shi**FT**, which explicitly separates transferable semantic information from domain-specific variations and performs demonstration selection in the resulting semantic space. We theoretically show that suppressing domain-specific information yields a tighter Bayesian divergence bound. To further exploit the shared semantic structure across heterogeneous sources, we introduce class-conditioned cross-domain alignment, which encourages same-class examples from different domains to form consistent semantic neighborhoods, thereby facilitating domain-invariant demonstration retrieval. Extensive experiments across 4 natural language understanding tasks, 7 LLMs, and 16 datasets consistently validate our theoretical findings and demonstrate the broad effectiveness of DRIFT under diverse multi-source distribution shifts. Our main contributions are summarized as follows:

- We systematically investigate multi-source OOD ICL and reveal that increasing source diversity can unexpectedly degrade ICL performance.
- We establish a Bayesian retrieval–inference framework and propose **DRIFT**, a domain-disentangled retrieval method with class-conditioned cross-domain alignment, theoretically yielding a tighter divergence bound.
- Extensive experiments across 4 tasks, 7 LLMs, and 16 datasets validate our theoretical findings and demonstrate the consistent effectiveness of DRIFT.

2 Related Work

OOD Generalization. OOD generalization aims to learn models from source distributions that remain robust to unseen target domains (Wang et al., 2022). Existing approaches mainly improve cross-domain robustness through data augmentation (Volpi et al., 2018; Zhou et al., 2021), invariant learning (Arjovsky et al., 2019; Krueger et al., 2021), and domain-invariant representation learning (Ganin et al., 2016). Particularly relevant to our work, representation disentanglement explicitly separates transferable semantics from domain-specific factors to mitigate domain interference (Bousmalis et al., 2016; Peng et al., 2019b). Different from these studies that improve OOD generalization through parameter learning, we introduce disentanglement into the retrieval process of ICL, separating semantic and domain information to mitigate source-specific interference during multi-source demonstration selection.

OOD ICL. Recent studies have revealed that ICL remains vulnerable to distribution shifts, despite the strong generalization capabilities of LLMs (Yuan et al., 2023; Mueller et al., 2024; Wang et al., 2025). Accordingly, increasing attention has been devoted to constructing robust demonstrations for OOD inference, including contrastive demonstration construction (Kwon et al., 2025), retrieval augmentation (Song et al., 2026a), and causal-invariant demonstration selection (Byun et al., 2026). More recently, LPL mitigates OOD bias through iterative pseudo-demonstration generation (Song et al., 2026b), while DOPA leverages OOD proxies to guide demonstration retrieval when the target distribution is inaccessible (Xu et al., 2026). Nevertheless, these studies predominantly consider demonstrations from a single source distribution, leaving the influence of heterogeneous multi-source demonstrations on the retrieval–inference process largely unexplored.

3 Definition and Theory

Problem Setup and Notation. Consider M source domains $\{\mathcal{D}_m\}_{m=1}^{M}$, whose demonstrations form a pooled candidate set $\mathcal{B} = \bigcup_{m=1}^{M} \mathcal{D}_m$. For an OOD query x_q from target domain, a retriever R selects K demonstrations from $\mathcal{B}$, denoted by $\mathcal{C}_R$. We decompose the representation of each input into a task-relevant semantic component z^s and a domain-specific component z^d. Let $d_s(x_q, x_i)$ denote the desired semantic distance between x_q and a candidate x_i, and let $\hat{d}_s(x_q, x_i)$ denote its learned approximation. The oracle semantic retriever is defined as

$$R^\star(x_q) = \arg\min_{|R|=K} \sum_{x_i \in R} d_s(x_q, x_i), \tag{1}$$

and, for any retriever R, we define its total semantic retrieval cost as

$$S_R(x_q) = \sum_{x_i \in R(x_q)} d_s(x_q, x_i). \tag{2}$$

For conventional entangled retrieval, we model the retrieval distance as

$$\hat{d}_e(x_q, x_i) = \hat{d}_s(x_q, x_i) + \lambda b_d(x_q, x_i), \tag{3}$$

where b_d denotes domain-induced perturbation and λ controls its influence. In the multi-domain case, for $x_i \in \mathcal{D}_m$, we further decompose the perturbation as

$$b_m(x_q, x_i) = b_0(x_q, x_i) + h_m(x_q, x_i), \tag{4}$$

where b_0 denotes the common domain-related perturbation and h_m represents the additional perturbation induced by source heterogeneity. Detailed definitions of the notations are in Tab. 3.

Theoretical Analysis. Building upon the above definitions, we develop the theoretical framework underlying our approach. We first formulate ICL as a Bayesian reasoning process (Panwar et al., 2024; Zhou et al., 2026) (**Def.** 1). We then introduce two mild assumptions that bridge retrieval-space distortion and inference degradation, characterizing the semantic smoothness of predictive likelihood and the bounded perturbations arising from semantic approximation and domain-specific information, respectively (**Asm.** 1 and 2). Under these assumptions, we derive two theoretical results that establish how retrieval deviations propagate through Bayesian inference and isolate the additional penalty induced by domain-entangled representations (**Thm.** 1 and 2). Finally, we demonstrate that removing domain-specific information from retrieval representations yields a strictly tighter bound on Bayesian inference error (**Cor.** 1), thereby providing a principled theoretical justification for our approach. Due to space constraints, all theoretical derivations are deferred to App. B.

Definition 1 (Bayesian ICL)**.** *Let Θ denote a latent task hypothesis. We model ICL as Bayesian updating with the prior distribution $p_0(\theta \mid x_q)$:*

$$p_R(\theta \mid x_q, \mathcal{C}_R) \propto p_0(\theta \mid x_q) \prod_{i=1}^{K} p_\theta(y_i \mid x_i), \tag{5}$$

with predictive distribution

$$P_R(y_q \mid x_q) = \int p_\theta(y_q \mid x_q) p_R(\theta \mid x_q, \mathcal{C}_R) d\theta. \tag{6}$$

Assumption 1 (Semantic Smoothness)**.** *Following the Lipschitz stability principle commonly adopted in domain adaptation (Cai et al., 2021), we assume that the task log-likelihood is locally smooth in the semantic representation space:* $|\log p_\theta(y \mid x) - \log p_\theta(y \mid x')| \le L d_s(x, x')$*, for semantically related samples within the retrieval neighborhood.*

Assumption 2 (Bounded Representation Error)**.** *The learned semantic metric* $\hat{d}_s$ *uniformly approximates the ideal semantic metric* $d^\star$*:* $\left|\hat{d}_s(x_q, x_i) - d^\star(x_q, x_i)\right| \le \delta_s$*. Moreover, the domain-induced perturbation is bounded by* $|b_d(x_q, x_i)| \le B_D$*.*

Theorem 1 (Bayesian Stability of Retrieved Demonstrations)**.** *Let* R *be any retrieval strategy and* $R^\star$ *the oracle semantic retriever. Under Assumption 1, the corresponding Bayesian predictive distributions satisfy*

$$D_{\mathrm{KL}}\left(P_R(Y_q \mid x_q) \| P_{R^\star}(Y_q \mid x_q)\right) \le 2L\left[S_R(x_q) + S_{R^\star}(x_q)\right]. \tag{7}$$

Theorem 2 (Domain-Induced Retrieval Penalty)**.** *Under Assumption 2, semantic retrieval satisfies* $S_s(x_q) \le S_{R^\star}(x_q) + 2K\delta_s$*, whereas domain-entangled retrieval satisfies*

$$S_{\mathrm{ent}}(x_q) \le S_{R^\star}(x_q) + 2K(\delta_s + \lambda B_D). \tag{8}$$

Furthermore, for multi-domain entangled retrieval, let the domain-specific perturbation be $b_m = b_0 + h_m$*, where* $|b_0| \le B_D$ *and* $|h_m| \le H_M$*. Then multi-domain entangled retrieval satisfies*

$$S_{\mathrm{multient}}(x_q) \le S_{R^\star}(x_q) + 2K\left[\delta_s + \lambda(B_D + H_M)\right]. \tag{9}$$

Corollary 1 (Tighter Disentangled Bayesian Bound)**.** *Combining Theorems 1 and 2, the disentangled semantic retrieval satisfies*

$$D_{\mathrm{KL}}\left(P_s \| P_\star\right) \le 4L\left[S_{R^\star}(x_q) + K\delta_s\right]. \tag{10}$$

For multi-domain entangled retrieval, we have

$$D_{\mathrm{KL}}\left(P_{\mathrm{multient}} \| P_\star\right) \le 4L\left[S_{R^\star}(x_q) + K\left(\delta_s + \lambda(B_D + H_M)\right)\right]. \tag{11}$$

Therefore, the gap between the two derived upper bounds is

$$\boxed{\mathrm{UB}_{\mathrm{multient}} - \mathrm{UB}_s = 4LK\lambda(B_D + H_M).} \tag{12}$$

4 Model: Class-Conditioned Domain-Disentangled Retrieval

Guided by the above theoretical analysis, our goal is to learn sufficiently disentangled semantic and domain representations, and to perform domain-invariant demonstration retrieval exclusively in the semantic space, thereby minimizing the influence of domain-specific variations on downstream ICL inference. Beyond representation disentanglement, we introduce class-conditioned cross-domain alignment to encourage semantically consistent representations of the same class across heterogeneous domains. Finally, we develop prototype-guided demonstration retrieval, which leverages class-level semantic prototypes to improve the semantic consistency and stability of retrieved demonstrations. The overall framework of DRIFT is illustrated in Fig. 2.

Domain-disentangled representations. Rather than introducing a new disentanglement mechanism, our contribution lies in establishing a principled integration of multi-domain disentanglement and in-context learning within a unified retrieval–inference framework. Accordingly, to keep the instantiation deliberately simple and broadly applicable, we adopt two canonical disentanglement mechanisms: domain-adversarial learning, which suppresses domain-discriminative information from semantic representations (Ganin et al., 2016), and representation orthogonality regularization, which explicitly encourages the semantic and domain-specific subspaces to encode complementary factors (Bousmalis et al., 2016).

Given an input x_i from source domain d_i, a shared encoder produces a representation that is projected into a semantic embedding z_i^s and a domain embedding z_i^d. For domain-adversarial learning, we attach a domain classifier C_{adv} to the semantic embedding through a gradient reversal layer (GRL): $\mathcal{L}_{\mathrm{adv}} = \frac{1}{B}\sum_{i=1}^{B} \mathrm{CE}(C_{\mathrm{adv}}(\mathrm{GRL}(z_i^s)), d_i)$. The classifier learns to predict the source domain, while the reversed gradient encourages the semantic encoder to remove domain-discriminative

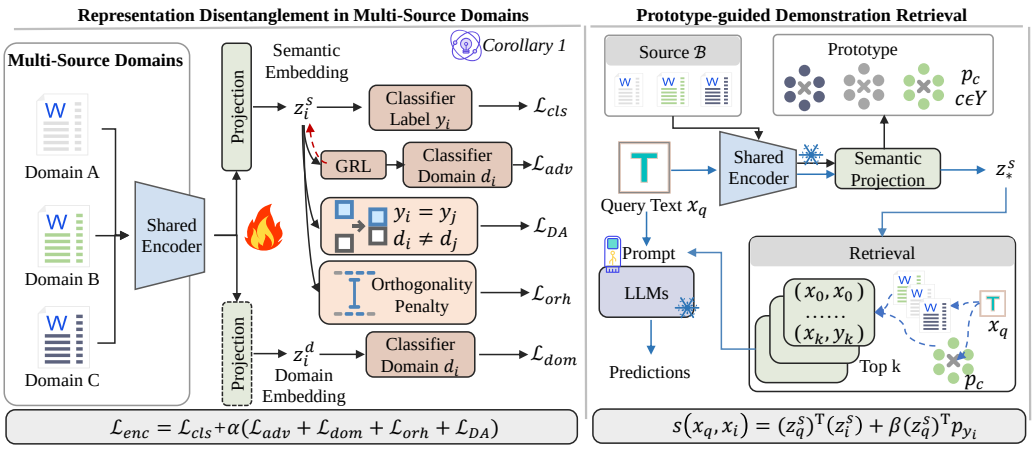


Figure 2: The overall framework of DRIFT.

information. In parallel, a domain classifier trained on z_i^d encourages the domain branch to preserve domain information, which is readily available in the multi-source setting since each demonstration is naturally associated with its source domain. We further discourage the two embeddings from encoding overlapping factors using an orthogonality penalty: $\mathcal{L}_{\text{orth}} = \frac{1}{B}\sum_{i=1}^{B}\left[(z_i^s)^\top z_i^d\right]^2$. Together, these objectives encourage a semantic space suitable for cross-domain retrieval while retaining domain-specific variation in a separate representation.

Class-conditioned cross-domain alignment. Domain invariance alone does not guarantee a retrieval-compatible semantic space: suppressing domain-specific information may remove distributional discrepancies, yet semantically equivalent examples from different domains can remain scattered in the representation space. This is particularly detrimental to ICL, where demonstration selection directly depends on neighborhood structure. We therefore explicitly align same-label representations across domains, encouraging the semantic space to preserve class-consistent neighborhoods while remaining insensitive to domain variations.

For each anchor i, we define its positive set as $\mathcal{P}(i) = \{j \neq i \mid y_j = y_i,\ d_j \neq d_i\}$. We apply a supervised contrastive objective that pulls these cross-domain, same-class examples together in the semantic space:

$$\mathcal{L}_{\text{DA}} = -\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\frac{1}{|\mathcal{P}(i)|}\sum_{p\in\mathcal{P}(i)}\log\frac{\exp\left((z_i^s)^\top z_p^s/\tau\right)}{\sum_{a\neq i}\exp((z_i^s)^\top z_a^s/\tau)}, \tag{13}$$

Examples from the same domain are not selected as positives, so the alignment objective explicitly focuses on semantic consistency across source domains. The disentangled encoder is trained by jointly optimizing task prediction, class-conditioned cross-domain alignment, domain-adversarial learning, domain prediction, and representation orthogonality: $\mathcal{L}_{\text{enc}} = \mathcal{L}_{\text{cls}} + \alpha(\mathcal{L}_{\text{DA}} + \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{dom}} + \mathcal{L}_{\text{orth}})$. Here, $\mathcal{L}_{\text{cls}}$ is the task-label classification[2] loss on z^s, and $\mathcal{L}_d$ is the source-domain classification loss on z^d. For simplicity, we set $\alpha = 0.1$ uniformly across all experiments without further tuning.

Prototype-guided demonstration retrieval. Although disentanglement suppresses domain-specific variations, instance-level retrieval can still be sensitive to sample-specific noise and intra-class variations. Inspired by prototype-based metric learning, where class prototypes provide compact and representative semantic anchors (Snell et al., 2017), we introduce prototype-guided retrieval to complement instance-level similarity with class-level semantic consistency.

After training, we encode all source examples and the target query using the semantic branch. For each class c, we construct a source prototype $p_c = \text{norm}\left(1/|\mathcal{S}_c|\sum_{i:y_i=c} z_i^s\right)$, where $\mathcal{S}_c$ contains source examples labeled c. The default retrieval score for source example i is $s(x_q, x_i) =$

[2]For NER, we adopt a token-level classification head rather than a sentence-level one.

$(z_q^s)^\top z_i^s + \beta(z_q^s)^\top p_{y_i}$. We select the k highest-scoring examples from the pooled source domains and concatenate their input–label pairs with x_q to form the in-context learning prompt. The language model then predicts the target label. To avoid redundant hyperparameter tuning, we set β to a mild, general-purpose value of 0.1.

5 EXPERIMENTS

5.1 DATASETS, BASELINES, AND DETAILS

Datasets. We conduct experiments on datasets drawn from BOSS (Yuan et al., 2023), a widely used benchmark for evaluating OOD generalization in NLP. Within each task, the constituent datasets differ substantially in data source, thereby providing naturally heterogeneous domains for multi-domain OOD evaluation. We consider four tasks: sentiment analysis (SA), toxicity detection (TD), natural language inference (NLI), and named entity recognition (NER). We follow the datasets and task configurations reported in BOSS, with one adaptation: because the available AdvCivil subset contains too few examples to support reliable multi-domain evaluation, we replace it with ImplicitHate, another distributionally distinct dataset included in the original BOSS benchmark.

Baselines. To empirically validate the effectiveness of our proposed method, we compare against multiple recent baselines. ►**kNN** (Liu et al., 2022) We retrieve the k most similar demonstrations from the pooled source domains using pretrained sentence embeddings. ►**DICL** (Kapuriya et al., 2025) We first identify query-relevant candidates by kNN and then apply maximal marginal relevance to select demonstrations that balance relevance and diversity. ►**UCS** (Xin et al., 2026) We combine a determinantal point process objective with estimated cluster coverage to select diverse demonstrations from a query-relevant candidate pool. ►**CCL** (Byun et al., 2026) We train a lightweight supervised VAE to separate task-related and domain-related latent representations, and retrieve demonstrations by similarity in the task-related space. ►**FT** We additionally compare with direct fine-tuning on multiple source domains as a reference. Besides, we evaluate our method using five base LLMs from two model families: `Llama3.2-1B`, `Llama3.1-8B`, `Qwen3-1.7B`, `Qwen3.5-2B`, and `Qwen3.5-9B`. The same LLM is used across retrieval methods within each experimental setting to isolate the effect of demonstration selection. Beyond the locally deployed models, we evaluate CCL and our method on two API-based LLMs, `Qwen3.5-Plus` and `DeepSeek-V4-Flash`.[3]

Details. We initialize the disentangled encoder with all-MiniLM-L6-v2 and train it on the pooled source domains for five epochs using AdamW with a learning rate of 2×10^{-5} and a batch size of 32. The semantic and domain projection heads produce 128-dimensional representations. We set the gradient-reversal coefficient to 1.0, the contrastive temperature to 0.07, and the auxiliary-loss coefficient to $\alpha = 0.1$. At inference time, we retrieve four demonstrations using prototype-guided semantic retrieval with $\beta = 0.1$ and perform greedy decoding. We evaluate up to 5,000 target examples per domain, reporting accuracy for classification tasks and entity-level exact-match accuracy for NER. Additional implementation and preprocessing details are provided in App C.

5.2 QUANTITATIVE RESULTS

As shown in Tab. 1, DRIFT consistently improves OOD ICL performance across diverse tasks and model scales. Among the five open-source LLMs with complete baseline comparisons, DRIFT achieves the highest average accuracy in all cases, outperforming the strongest competing method by 4.99, 2.88, 3.06, 2.96, and 1.89 percentage points, respectively. The improvements are also highly consistent at the task level, with DRIFT achieving the best performance in 18 of the 20 model–task combinations. Similar gains are observed on larger proprietary LLMs, where DRIFT improves the average accuracy of `Qwen3.5-Plus` and `DeepSeek-V4-Flash` by 3.89 and 1.77 points over CCL. Notably, these improvements span substantially different tasks and model scales, suggesting that the benefits are not tied to a particular model or distribution. Overall, the consistent gains support our central hypothesis that suppressing domain-specific information and retrieving demonstrations in the disentangled semantic space can effectively mitigate multi-source interference and

[3]We also evaluated the latest GPT-5.6 model, but observed frequent refusals on TD tasks. Since these refusals prevented a reliable comparison, we exclude its results from the reported experiments.

Table 1: Mean accuracy (%) over the four datasets in each task category. Bold and underlined values indicate the best and second-best ICL results for each model, respectively. Δ denotes discrepancy compared to our method.

LLMs	Method	SA	TD	NLI	NER	Average	Δ
–	FT	60.41	73.78	38.40	64.00	59.15	–
Llama3.2-1B	kNN	49.58	53.20	37.48	28.80	42.27	10.83
	DICL	48.83	58.11	**38.71**	27.84	43.37	9.73
	UCS	50.05	58.05	37.55	28.03	43.42	9.68
	CCL	53.75	68.18	36.03	34.48	48.11	4.99
	DRIFT	**63.09**	**70.85**	38.42	**40.06**	**53.10**	–
Llama3.1-8B	kNN	66.19	70.34	43.58	42.04	55.54	3.49
	DICL	66.44	71.81	43.30	41.02	55.64	3.39
	UCS	65.62	68.94	42.63	41.22	54.60	4.43
	CCL	65.74	**73.07**	42.00	43.81	56.15	2.88
	DRIFT	**69.54**	73.00	**43.78**	**49.81**	**59.03**	–
Qwen3-1.7B	kNN	63.97	47.55	36.15	37.36	46.26	3.54
	DICL	63.93	48.66	36.48	37.89	46.74	3.06
	UCS	62.25	48.30	36.73	38.73	46.50	3.30
	CCL	64.19	48.12	36.46	35.98	46.19	3.61
	DRIFT	**66.09**	**50.84**	**39.17**	**43.09**	**49.80**	–
Qwen3.5-2B	kNN	61.64	59.77	39.18	38.42	49.75	2.96
	DICL	61.39	59.10	39.23	38.30	49.50	3.21
	UCS	61.50	58.47	38.96	38.78	49.43	3.28
	CCL	59.69	63.71	37.83	35.14	49.09	3.62
	DRIFT	**65.11**	**66.45**	**39.87**	**39.42**	**52.71**	–
Qwen3.5-9B	kNN	71.34	72.32	51.04	12.35	51.77	2.03
	DICL	71.61	74.46	50.15	11.42	51.91	1.89
	UCS	69.30	73.45	50.85	11.83	51.36	2.44
	CCL	70.97	70.27	51.01	14.33	51.64	2.16
	DRIFT	**71.92**	**75.18**	**52.71**	**15.39**	**53.80**	–
Qwen3.5-Plus	CCL	75.75	72.50	57.63	63.50	67.34	3.89
	DRIFT	**77.25**	**73.63**	**62.38**	**71.67**	**71.23**	–
DeepSeek-V4-Flash	CCL	77.13	85.75	67.50	78.95	77.33	1.77
	DRIFT	**79.63**	**87.38**	**70.13**	**79.27**	**79.10**	–

improve OOD ICL. For completeness, we also report the detailed results on all individual datasets in App. D.1, providing a more fine-grained comparison across different distribution shifts. Building on these observations, we conduct further analyses to better understand the underlying mechanisms of DRIFT and examine how its key components contribute to robust multi-source OOD ICL.

5.3 Ablation Study

Table 2: Ablation study of DRIFT across four tasks.

LLMs	Pro.	DA	Adv.	Orth.	SA	TD	NLI	NER	Avg.	Δ
Llama3.2-1B	✓	✓	✓	✓	**63.09**	70.85	**38.42**	**40.06**	**53.10**	–
	✗	✓	✓	✓	61.53	70.52	38.04	38.33	52.10	1.00
	✓	✗	✓	✓	63.02	70.79	34.95	39.01	51.94	1.16
	✓	✓	✗	✓	63.06	70.51	36.83	39.22	52.41	0.69
	✓	✓	✓	✗	62.74	**71.03**	35.80	38.39	51.99	1.11
Qwen3-1.7B	✓	✓	✓	✓	**66.09**	50.84	**39.17**	**43.09**	**49.80**	–
	✗	✓	✓	✓	65.52	**51.28**	38.91	42.83	49.63	0.17
	✓	✗	✓	✓	65.55	47.80	35.43	40.53	47.33	2.47
	✓	✓	✗	✓	65.14	50.47	37.13	41.86	48.65	1.15
	✓	✓	✓	✗	65.62	49.59	36.46	41.89	48.39	1.41

We retain the basic task and domain classification objectives and ablate the remaining training objectives and retrieval component of DRIFT. Specifically, **Pro.** denotes the prototype-guided scoring term in demonstration retrieval, which can be removed by setting $\beta = 0$, while **DA**, **Adv.**, and

Orth. correspond to class-conditioned cross-domain alignment, domain-adversarial learning, and representation orthogonality in $\mathcal{L}_{enc}$, respectively.

As shown in Tab. 2, removing any component generally degrades the overall performance, confirming that these components provide complementary benefits. In particular, removing cross-domain alignment (**DA**) causes the largest average degradation on both `Llama3.2-1B` and `Qwen3-1.7B`, with drops of 1.16 and 2.47 points, respectively, highlighting the importance of preserving class-consistent semantic neighborhoods across heterogeneous sources. Removing **Adv.** or **Orth.** also consistently reduces the average performance, supporting the effectiveness of explicitly suppressing domain information and separating semantic and domain-specific representations. Prototype guidance further improves retrieval, although its contribution varies across models, suggesting that class-level semantic anchors provide complementary information beyond instance-level similarity. Overall, the ablation results support the joint design of representation disentanglement, cross-domain semantic alignment, and prototype-guided retrieval in DRIFT. Detailed results of the ablation study are provided in App. D.2.

5.4 Visualization

To examine whether the learned semantic and domain representations are effectively disentangled, we visualize z_s and z_d and quantify their representational similarity using Centered Kernel Alignment (CKA) (Kornblith et al., 2019), where a lower value indicates weaker structural similarity between the two representation spaces. As shown in Fig. 3, z_s and z_d exhibit clearly distinct geometric structures on both DynaSent and ImplicitHate. More importantly, the consistently low CKA scores (0.04 and 0.02, respectively) indicate limited representational overlap between the semantic and domain branches. These results suggest that our framework effectively separates task-relevant semantic information from domain-specific variations while maintaining consistent representations across source and target distributions. Additional visual results are provided in App. D.3.

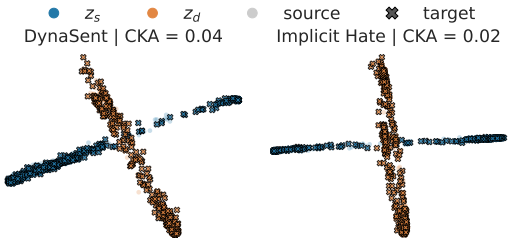


Figure 3: Visualization of representation disentanglement.

5.5 Semantic vs. Domain Retrieval

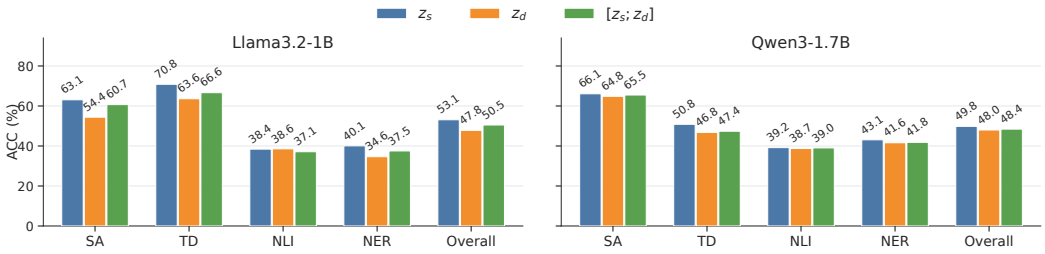


Figure 4: Impact of different retrieval representations.

To directly examine the roles of semantic and domain information in demonstration retrieval, we compare three retrieval spaces: the disentangled semantic representation z_s, the domain representation z_d, and their concatenation $[z_s; z_d]$. For a fair comparison, all settings use the same retrieval strategy and differ only in the representations used to measure demonstration similarity.

As shown in Fig. 4, retrieval with z_s consistently achieves the best performance across all four tasks and both LLMs. In contrast, using z_d leads to substantial degradation, particularly on SA, TD, and NER, indicating that domain similarity alone provides limited guidance for selecting task-relevant demonstrations. Notably, concatenating z_s and z_d also consistently underperforms semantic-only retrieval, with the overall accuracy decreasing from 53.1 to 50.5 on `Llama3.2-1B` and from 49.8

to 48.4 on `Qwen3-1.7B`. This suggests that explicitly retaining domain information in the retrieval representation can interfere with otherwise useful semantic similarity. Overall, these results provide direct empirical support for our theoretical analysis: effective multi-source OOD retrieval should prioritize transferable semantic information while suppressing domain-specific variations. The consistent advantage of z_s further validates the central design of DRIFT, where demonstration retrieval is performed exclusively in the disentangled semantic space.

5.6 Source Impact Analysis

We further investigate how source-domain heterogeneity affects retrieval-based ICL and whether DRIFT remains robust to changes in source composition. For each target domain, we enumerate different combinations of the remaining source domains and measure their discrepancy from the target using MMD^2. Since the absolute scale of MMD^2 and prediction accuracy can vary substantially across target datasets, directly correlating their raw values may conflate source heterogeneity with dataset-specific effects. We therefore standardize MMD^2 within each target domain as $\widetilde{D}_{t,\mathcal{S}} = \frac{D_{t,\mathcal{S}} - \mu_t^D}{\sigma_t^D}$, where $D_{t,\mathcal{S}}$ denotes the MMD^2 between target t and source combination $\mathcal{S}$, and μ_t^D and σ_t^D are computed over all source combinations for the same target. Similarly, we center the ICL performance within each target: $\widetilde{A}_{t,\mathcal{S}} = A_{t,\mathcal{S}} - \mu_t^A$. This normalization isolates the relative effect of source heterogeneity from intrinsic differences across datasets. We then fit LOWESS curves over all source configurations to characterize the potentially nonlinear relationship between source heterogeneity and ICL performance, without imposing a predefined functional form.

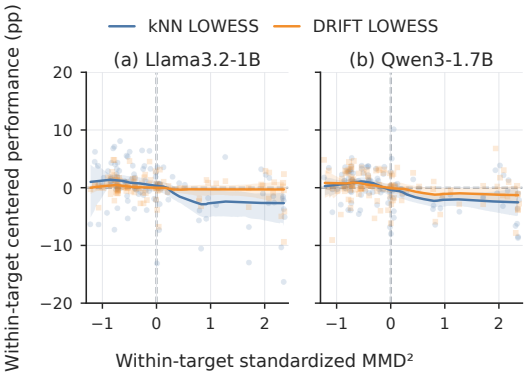


Figure 5: Performance sensitivity to source heterogeneity.

As shown in Fig. 5, kNN becomes increasingly sensitive to source heterogeneity, with a pronounced performance degradation on `Llama3.2-1B` as the source–target discrepancy increases. In contrast, DRIFT exhibits a substantially flatter response, maintaining relatively stable performance across heterogeneous source configurations. Although this effect is less pronounced on `Qwen3-1.7B`, DRIFT still shows reduced performance variation as source discrepancy changes. Moreover, DRIFT consistently exhibits a smaller variance in the performance distribution, providing further evidence of its stability across heterogeneous source configurations. These observations consistently support our hypothesis that disentangling domain-specific information improves the robustness of multi-source retrieval.

6 Conclusion

In this work, we investigate multi-source OOD in-context learning and show that domain-specific information entangled in retrieval representations can introduce source-dependent interference. Based on our theoretical analysis, we propose DRIFT, a domain-disentangled retrieval framework that preserves transferable semantics while suppressing domain-specific information. Extensive experiments across diverse tasks, source domains, and LLMs demonstrate consistent improvements over strong retrieval baselines, while further analyses confirm improved semantic retrieval and robustness to source heterogeneity. These findings highlight the importance of disentangling transferable semantics from domain-specific variations for robust multi-source OOD ICL.

Limitations. Despite our broad evaluation, the rapidly expanding landscape of LLMs and downstream tasks cannot be exhaustively covered. DRIFT also assumes available source-domain labels, a common yet potentially restrictive assumption in domain-adversarial learning. Extending evaluation and relaxing this assumption remain important directions for future work.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. *Advances in neural information processing systems*, 29, 2016.

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.

Hoyoon Byun, Gyeongdeok Seo, Joonseong Kang, Taero Kim, Jihee Kim, and Kyungwoo Song. Ccl: Causal-aware in-context learning for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 38:83676–83695, 2026.

Guanyu Cai, Lianghua He, MengChu Zhou, Hesham Alhumade, and Die Hu. Learning smooth representation for unsupervised domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8):4181–4195, 2021.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

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.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pp. 5815–5826. PMLR, 2021.

Soo Min Kwon, Alec S Xu, Can Yaras, Laura Balzano, and Qing Qu. Out-of-distribution generalization of in-context learning: A low-dimensional subspace perspective. *arXiv preprint arXiv:2505.14808*, 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, 2025.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd workshop on knowledge extraction and integration for deep learning architectures*, pp. 100–114, 2022.

Manas Madine. Bridging distribution gap via semantic rewriting with llms to enhance ood robustness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 334–344, 2024.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. *Advances in neural information processing systems*, 21, 2008.

Aaron Mueller, Albert Webson, Jackson Petty, and Tal Linzen. In-context learning generalizes, but not always robustly: The case of syntax. 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. 4761–4779, 2024.

Madhur Panwar, Kabir Ahuja, and Navin Goyal. In-context learning through the bayesian prism. In *International Conference on Learning Representations*, volume 2024, pp. 49789–49843, 2024.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1406–1415. IEEE, 2019a.

Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations. In *International conference on machine learning*, pp. 5102–5112. PMLR, 2019b.

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.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Rui Song, Fausto Giunchiglia, Yingji Li, Mingjie Tian, and Hao Xu. Tacit: A target-agnostic feature disentanglement framework for cross-domain text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 18999–19007, 2024.

Rui Song, Yingji Li, Fausto Giunchiglia, and Hao Xu. Retrieval augmentation for out-of-distribution robustness in non-knowledge intensive in-context learning. *Science China Information Sciences*, 69(2):122108, 2026a.

Rui Song, Yingji Li, Jian Li, Fausto Giunchiglia, and Hao Xu. From fake to real: Mitigating out-of-distribution bias in in-context learning via feedback supervision from large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 15197–15213, 2026b.

Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072, 2022.

Qixun Wang, Yifei Wang, Xianghua Ying, and Yisen Wang. Can in-context learning really generalize to out-of-distribution tasks? In *International Conference on Learning Representations*, volume 2025, pp. 83553–83574, 2025.

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.

Hao Xu, Rite Bo, Fausto Giunchiglia, Yingji Li, and Rui Song. Toward robust in-context learning: Leveraging out-of-distribution proxies for target inaccessible demonstration retrieval. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20540–20556, 2026.

Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations. *Advances in Neural Information Processing Systems*, 36:58478–58507, 2023.

Guanglin Zhou, Zhongyi Han, Shaoan Xie, Shiming Chen, Biwei Huang, Liming Zhu, Xin Gao, Lina Yao, and Salman Khan. From small to large: In-context learning as a new paradigm for domain generalization. *International Journal of Computer Vision*, 134(1):9, 2026.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021.

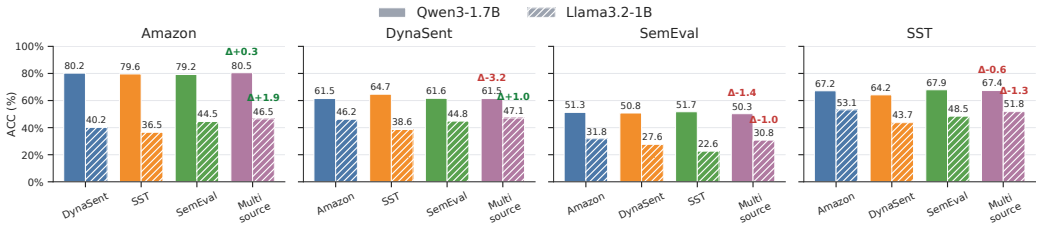


Figure 6: Preliminary study of single-source and multi-source demonstration retrieval for OOD ICL across four SA target domains and two LLMs.

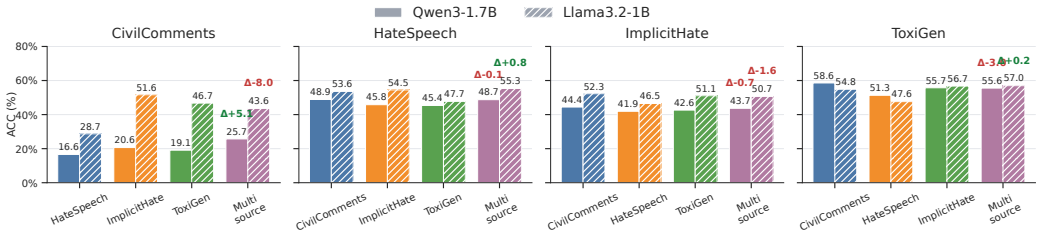


Figure 7: Preliminary study of single-source and multi-source demonstration retrieval for OOD ICL across four TD target domains and two LLMs.

A More Comparisons Between Single-source and Multi-source Retrieval Results

The preliminary experiments follow a configuration similar to that of our main experiments, except that we set $k = 2$ to reduce inference cost. As shown in Figs. 6 and 7, performance degradation caused by directly pooling multiple source domains remains evident across different target domains and LLMs, with a drop of up to 8.0 percentage points on CivilComments. Meanwhile, multi-source retrieval can occasionally be beneficial. For example, consistent improvements are observed when Amazon serves as the target domain, suggesting that heterogeneous sources may sometimes provide complementary task-relevant semantics. Nevertheless, the overall results reinforce our central observation: simply pooling heterogeneous sources does not guarantee better OOD ICL performance, and can introduce substantial interference when the source distributions are poorly aligned with the target.

B Further Explanation of the Theorem and Theoretical Proof

In this section, we provide more detailed derivations and additional explanations of our theoretical results to facilitate a clearer understanding of the underlying assumptions and proofs.

B.1 Explanation of the Bayesian ICL Formulation

Following the Bayesian interpretation of ICL (Panwar et al., 2024; Zhou et al., 2026), we introduce a latent task hypothesis Θ and interpret the demonstrations as evidence for updating the model's belief over Θ. Given a query x_q, let $p_0(\theta \mid x_q)$ denote the prior distribution over the latent hypothesis before observing the retrieved demonstrations $\mathcal{C}_R = \{(x_i, y_i)\}_{i=1}^{K}$. By Bayes' rule, the posterior distribution is

$$p_R(\theta \mid x_q, \mathcal{C}_R) = \frac{p(\mathcal{C}_R \mid \theta, x_q) p_0(\theta \mid x_q)}{p(\mathcal{C}_R \mid x_q)}. \tag{14}$$

Table 3: Summary of the main mathematical notation used in this paper.

Symbol	Description	Symbol	Description
$\mathcal{D}_m$	m-th source domain	M	Number of source domains
$\mathcal{B}$	Pooled source bank, $\bigcup_{m=1}^{M} \mathcal{D}_m$	x_q	Query from the unseen target domain
x_i	i-th source candidate	y_i, d_i	Task and domain labels of x_i
y_q, Y_q	Query label and corresponding random variable	R	Demonstration retrieval strategy
$R^\star$	Oracle semantic retriever	$\mathcal{C}_R$	Demonstrations selected by R
K	Number of retrieved demonstrations	z^s, z^d	Semantic and domain-specific representations
z_i^s, z_i^d	Representations of source example x_i	z_q^s, z_q^d	Representations of query x_q
d_s	Desired semantic distance	$d^\star$	Ideal semantic metric
$\hat{d}_s$	Learned semantic distance	$\hat{d}_e$	Domain-entangled retrieval distance
b_d	Domain-induced perturbation	λ	Weight of domain perturbation
b_m	Perturbation for source domain $\mathcal{D}_m$	b_0	Common domain-related perturbation
h_m	Source-heterogeneity perturbation	δ_s	Bound on semantic approximation error
B_D	Bound on domain-induced perturbation	H_M	Bound on heterogeneity perturbation
S_R	Semantic retrieval cost of R	$S_{R^\star}$	Semantic retrieval cost of $R^\star$
S_s	Cost of semantic retrieval	S_{ent}	Cost of entangled retrieval
S_{multient}	Cost of multi-domain entangled retrieval	Θ, θ	Latent task hypothesis and its realization
$p_0(\theta \mid x_q)$	Prior over task hypotheses	$p_\theta(y \mid x)$	Label likelihood under θ
$p_R(\theta \mid x_q, \mathcal{C}_R)$	Posterior induced by retrieved context	P_R	Predictive distribution induced by R
P_s	Predictive distribution of semantic retrieval	$P^\star$	Oracle predictive distribution
P_{multient}	Predictive distribution of multi-domain entangled retrieval	D_{KL}	Kullback–Leibler divergence
L	Semantic smoothness (Lipschitz) constant	UB_s	Bayesian upper bound for semantic retrieval
UB_{multient}	Upper bound for multi-domain entangled retrieval	B	Training mini-batch size
$\mathcal{L}_{\text{cls}}$	Task-label classification loss	$\mathcal{L}_{\text{adv}}$	Domain-adversarial loss
$\mathcal{L}_{\text{dom}}$	Domain classification loss	$\mathcal{L}_{\text{orth}}$	Representation orthogonality loss
$\mathcal{L}_{\text{DA}}$	Class-conditioned cross-domain alignment loss	$\mathcal{L}_{\text{enc}}$	Overall encoder training objective
α	Weight of auxiliary training objectives	$\mathcal{I}$	Set of contrastive anchor indices
$\mathcal{P}(i)$	Cross-domain same-class positives for anchor i	τ	Contrastive temperature
$\mathcal{S}_c$	Source examples belonging to class c	p_c	Semantic prototype of class c
p_{y_i}	Prototype associated with label y_i	$s(x_q, x_i)$	Prototype-guided retrieval score
β	Weight of prototype similarity	k	Number of demonstrations selected at inference
$D_{t,\mathcal{S}}$	MMD^2 between target t and source set $\mathcal{S}$	μ_t^D, σ_t^D	Mean/std. of $D_{t,\mathcal{S}}$ for target t
$\widetilde{D}_{t,\mathcal{S}}$	Standardized source–target discrepancy	$A_{t,\mathcal{S}}$	ICL performance under source set $\mathcal{S}$
μ_t^A	Mean performance for target t	$\widetilde{A}_{t,\mathcal{S}}$	Within-target centered performance
N	Number of NER evaluation samples	G_i, P_i	Ground-truth and predicted entity sets

Under the standard conditional-independence assumption that the demonstrations are independent given the latent hypothesis, and that their labels are generated according to $p_\theta(y_i \mid x_i)$, we have

$$p(\mathcal{C}_R \mid \theta, x_q) = \prod_{i=1}^{K} p_\theta(y_i \mid x_i). \tag{15}$$

Substituting Eq. 15 into Eq. 14 gives

$$p_R(\theta \mid x_q, \mathcal{C}_R) = \frac{p_0(\theta \mid x_q) \prod_{i=1}^{K} p_\theta(y_i \mid x_i)}{Z_R(x_q)}, \tag{16}$$

where

$$Z_R(x_q) = \int p_0(\theta \mid x_q) \prod_{i=1}^{K} p_\theta(y_i \mid x_i) d\theta \tag{17}$$

is the normalization constant. Equivalently,

$$p_R(\theta \mid x_q, \mathcal{C}_R) \propto p_0(\theta \mid x_q) \prod_{i=1}^{K} p_\theta(y_i \mid x_i). \tag{18}$$

The prediction for the query is then obtained by marginalizing the latent hypothesis Θ. Specifically,

$$\begin{aligned} P_R(y_q \mid x_q, \mathcal{C}_R) &= \int p(y_q, \theta \mid x_q, \mathcal{C}_R) d\theta \\ &= \int p(y_q \mid x_q, \theta, \mathcal{C}_R) p_R(\theta \mid x_q, \mathcal{C}_R) d\theta. \end{aligned} \tag{19}$$

Assuming that, conditioned on the latent hypothesis Θ and the query x_q, the query label is independent of the demonstrations,

$$p(y_q \mid x_q, \theta, \mathcal{C}_R) = p_\theta(y_q \mid x_q), \tag{20}$$

which yields the posterior predictive distribution

$$\boxed{P_R(y_q \mid x_q, \mathcal{C}_R) = \int p_\theta(y_q \mid x_q) p_R(\theta \mid x_q, \mathcal{C}_R) d\theta.} \tag{21}$$

Therefore, the retrieved demonstrations affect prediction through the following Bayesian updating process:

$$\mathcal{C}_R \to p_R(\Theta \mid x_q, \mathcal{C}_R) \to P_R(Y_q \mid x_q, \mathcal{C}_R). \tag{22}$$

B.2 Explanation of Disentangled Retrieval

For each input x, we decompose its representation into a semantic component z^s and a domain-specific component z^d, where z^s is expected to preserve task-relevant information and z^d captures domain-dependent characteristics. The desired retrieval metric therefore depends only on the semantic component:

$$d_s(x_q, x_i) = d(z_q^s, z_i^s).$$

In a conventional entangled representation, however, semantic and domain-specific information jointly affect the similarity between the query and a candidate demonstration. We model this effect as

$$d_e(x_q, x_i) = d_s(x_q, x_i) + \lambda b_d(x_q, x_i),$$

where $b_d(x_q, x_i)$ denotes the change in retrieval distance caused by domain-specific information, and λ controls its relative influence. This formulation should be understood as an analytical decomposition of the retrieval metric rather than a specific architectural assumption. In particular, b_d can be either positive or negative, indicating that domain information may make a candidate appear farther from or closer to the query than suggested by semantic similarity alone. Our method explicitly separates the two factors and performs retrieval using only d_s, thereby preventing domain-specific information in OOD tasks from directly participating in demonstration ranking.

B.3 Proof of Theorem 1

Proof. Let $\mathcal{C}_R = \{(x_i, y_i)\}_{i=1}^K$ and $\mathcal{C}_\star = \{(x_i^\star, y_i)\}_{i=1}^K$ denote the contexts retrieved by R and $R^\star$, respectively, with the same label composition. Define

$$\mathcal{L}_R(\theta) = \prod_{i=1}^{K} p_\theta(y_i \mid x_i), \qquad \mathcal{L}_{R^\star}(\theta) = \prod_{i=1}^{K} p_\theta(y_i \mid x_i^\star). \tag{23}$$

By Assumption 1 and the triangle inequality of $d^\star$,

$$\left|\log \frac{\mathcal{L}_R(\theta)}{\mathcal{L}_{R^\star}(\theta)}\right| \le L \sum_{i=1}^{K} d^\star(x_i, x_i^\star) \le L\left[S_R(x_q) + S_{R^\star}(x_q)\right]. \tag{24}$$

Let $A = L\left[S_R(x_q) + S_{R^\star}(x_q)\right]$, then

$$e^{-A} \le \frac{\mathcal{L}_R(\theta)}{\mathcal{L}_{R^\star}(\theta)} \le e^A.$$

Equivalently,

$$e^{-A}\mathcal{L}_{R^\star}(\theta) \le \mathcal{L}_R(\theta) \le e^A \mathcal{L}_{R^\star}(\theta).$$

Multiplying all terms by the common non-negative prior $p_0(\theta \mid x_q)$ gives

$$e^{-A} p_0(\theta \mid x_q)\mathcal{L}_{R^\star}(\theta) \le p_0(\theta \mid x_q)\mathcal{L}_R(\theta) \le e^A p_0(\theta \mid x_q)\mathcal{L}_{R^\star}(\theta).$$

Integrating over θ, we obtain

$$e^{-A} Z_{R^\star} \le Z_R \le e^A Z_{R^\star},$$

where

$$Z_R = \int p_0(\theta \mid x_q)\mathcal{L}_R(\theta)\, d\theta, \qquad Z_{R^\star} = \int p_0(\theta \mid x_q)\mathcal{L}_{R^\star}(\theta)\, d\theta \tag{25}$$

are the corresponding Bayesian normalization constants in Definition 1. Hence,

$$e^{-A} \le \frac{Z_R}{Z_{R^\star}} \le e^A, \tag{26}$$

which implies

$$\left|\log \frac{Z_{R^\star}}{Z_R}\right| \le A. \tag{27}$$

Using the Bayesian posteriors

$$p_R(\theta) = \frac{p_0(\theta \mid x_q)\mathcal{L}_R(\theta)}{Z_R}, \qquad p_{R^\star}(\theta) = \frac{p_0(\theta \mid x_q)\mathcal{L}_{R^\star}(\theta)}{Z_{R^\star}}, \tag{28}$$

their ratio can be written as

$$\frac{p_R(\theta)}{p_{R^\star}(\theta)} = \frac{\mathcal{L}_R(\theta)}{\mathcal{L}_{R^\star}(\theta)} \frac{Z_{R^\star}}{Z_R}. \tag{29}$$

Therefore,

$$\left|\log \frac{p_R(\theta)}{p_{R^\star}(\theta)}\right| \le \left|\log \frac{\mathcal{L}_R(\theta)}{\mathcal{L}_{R^\star}(\theta)}\right| + \left|\log \frac{Z_{R^\star}}{Z_R}\right| \le 2A. \tag{30}$$

Consequently,

$$D_{\mathrm{KL}}\left(p_R(\Theta) \| p_{R^\star}(\Theta)\right) \le 2A. \tag{31}$$

Since the two predictive distributions are obtained by passing the corresponding posteriors through the same conditional prediction channel $p_\theta(y_q \mid x_q)$, the data-processing inequality gives

$$D_{\mathrm{KL}}\left(P_R(Y_q \mid x_q) \| P_{R^\star}(Y_q \mid x_q)\right) \le D_{\mathrm{KL}}\left(p_R(\Theta) \| p_{R^\star}(\Theta)\right) \le 2L\left[S_R(x_q) + S_{R^\star}(x_q)\right]. \tag{32}$$

□

B.4 Proof of Theorem 2

Proof. Under Assumption 2, the learned semantic distance satisfies

$$\left|\hat{d}_s(x_q, x_i) - d_s(x_q, x_i)\right| \leq \delta_s. \tag{33}$$

Let R_s denote the K demonstrations selected according to $\hat{d}_s$, and let $R^\star$ denote the oracle retriever minimizing the total semantic distance under d_s. Since R_s minimizes the total distance measured by $\hat{d}_s$, we have

$$\sum_{x_i \in R_s} \hat{d}_s(x_q, x_i) \leq \sum_{x_i \in R^\star} \hat{d}_s(x_q, x_i). \tag{34}$$

Combining this optimality condition with Assumption 2 gives

$$S_s(x_q) \leq \sum_{x_i \in R_s} \hat{d}_s(x_q, x_i) + K\delta_s \leq \sum_{x_i \in R^\star} \hat{d}_s(x_q, x_i) + K\delta_s \leq S_{R^\star}(x_q) + 2K\delta_s. \tag{35}$$

Therefore,

$$S_s(x_q) \leq S_{R^\star}(x_q) + 2K\delta_s. \tag{36}$$

For the entangled retrieval metric, recall that

$$\hat{d}_e(x_q, x_i) = \hat{d}_s(x_q, x_i) + \lambda b_d(x_q, x_i). \tag{37}$$

Similiar, by Assumption 2,

$$\left|\hat{d}_e(x_q, x_i) - d_s(x_q, x_i)\right| \leq \left|\hat{d}_s(x_q, x_i) - d_s(x_q, x_i)\right| + \lambda |b_d(x_q, x_i)| \leq \delta_s + \lambda B_D. \tag{38}$$

Let R_{ent} denote the K demonstrations selected according to $\hat{d}_e$. Since R_{ent} minimizes the total distance measured by $\hat{d}_e$,

$$\sum_{x_i \in R_{\text{ent}}} \hat{d}_e(x_q, x_i) \leq \sum_{x_i \in R^\star} \hat{d}_e(x_q, x_i). \tag{39}$$

Applying the same argument yields

$$\begin{aligned} S_{\text{ent}}(x_q) &\leq \sum_{x_i \in R_{\text{ent}}} \hat{d}_e(x_q, x_i) + K(\delta_s + \lambda B_D) \\ &\leq \sum_{x_i \in R^\star} \hat{d}_e(x_q, x_i) + K(\delta_s + \lambda B_D) \\ &\leq S_{R^\star}(x_q) + 2K(\delta_s + \lambda B_D). \end{aligned} \tag{40}$$

Hence,

$$S_{\text{ent}}(x_q) \leq S_{R^\star}(x_q) + 2K(\delta_s + \lambda B_D). \tag{41}$$

For a candidate x_i from source domain $\mathcal{D}_m$, the **multi-domain** entangled metric is

$$\hat{d}_e^{(m)}(x_q, x_i) = \hat{d}_s(x_q, x_i) + \lambda b_m(x_q, x_i), \tag{42}$$

where $b_m = b_0 + h_m$. By the bounded perturbation assumption,

$$|b_m| \leq |b_0| + |h_m| \leq B_D + H_M. \tag{43}$$

Therefore,

$$\left|\hat{d}_e^{(m)}(x_q, x_i) - d_s(x_q, x_i)\right| \leq \delta_s + \lambda(B_D + H_M). \tag{44}$$

Applying the same Top-K retrieval argument as above gives

$$S_{\text{multi}}(x_q) \leq S_{R^\star}(x_q) + 2K\left[\delta_s + \lambda(B_D + H_M)\right]. \tag{45}$$

This completes the proof. □

B.5 Proof of Corollary 1

Proof. By Theorem 1, for any retrieval strategy R,

$$D_{\mathrm{KL}}\left(P_R \| P_{R^\star}\right) \leq 2L\left[S_R(x_q) + S_{R^\star}(x_q)\right]. \tag{46}$$

For semantic retrieval, Theorem 2 gives

$$S_s(x_q) \leq S_{R^\star}(x_q) + 2K\delta_s. \tag{47}$$

Substituting it into the Bayesian stability bound yields

$$D_{\mathrm{KL}}\left(P_s \| P_{R^\star}\right) \leq 4L\left[S_{R^\star}(x_q) + K\delta_s\right]. \tag{48}$$

For domain-entangled retrieval, we similarly have

$$S_{\mathrm{ent}}(x_q) \leq S_{R^\star}(x_q) + 2K(\delta_s + \lambda B_D), \tag{49}$$

which gives the intermediate bound

$$D_{\mathrm{KL}}\left(P_{\mathrm{ent}} \| P_{R^\star}\right) \leq 4L\left[S_{R^\star}(x_q) + K(\delta_s + \lambda B_D)\right]. \tag{50}$$

For multi-domain entangled retrieval, Theorem 2 further gives

$$S_{\mathrm{multient}}(x_q) \leq S_{R^\star}(x_q) + 2K\left[\delta_s + \lambda(B_D + H_M)\right]. \tag{51}$$

Applying Theorem 1 again, we obtain

$$D_{\mathrm{KL}}\left(P_{\mathrm{multient}} \| P_{R^\star}\right) \leq 4L\left[S_{R^\star}(x_q) + K\left(\delta_s + \lambda(B_D + H_M)\right)\right]. \tag{52}$$

Therefore, the gap between the derived multi-domain entangled and semantic upper bounds is

$$\boxed{\mathrm{UB}_{\mathrm{multient}} - \mathrm{UB}_s = 4LK\lambda(B_D + H_M).} \tag{53}$$

This completes the proof. □

C More Details

Here, we provide more detailed information on the datasets, prompt templates, model configurations, and other experimental settings.

Tasks and preprocessing. We evaluate four NLP tasks from BOSS under a unified generative ICL formulation. For sentiment analysis (SA), we use Amazon, DynaSent, SemEval, and SST, and map their annotations to a shared three-class label space consisting of *negative*, *positive*, and *neutral*. For toxicity detection (TD), we use CivilComments, ToxiGen, ImplicitHate, and HateSpeech, and normalize their annotations into the binary labels *benign* and *toxic*. For natural language inference (NLI), we use MNLI, ANLI, ContractNLI, and WANLI, concatenate each premise–hypothesis pair into a single encoder input, and unify the labels as *entailment*, *neutral*, and *contradiction*. For named entity recognition (NER), we use FewNERD, CoNLL, ENER, and WNUT. We convert all NER datasets to a sentence-level BIO format and retain the three entity types shared most consistently across datasets: *person*, *location*, and *organization*; all remaining entity types are mapped to O, and sentences containing no retained entity are removed. The metrics for computing NER performance are as follows:

$$\mathrm{EntityAcc} = \frac{\sum_{i=1}^{N} |G_i \cap P_i|}{\sum_{i=1}^{N} |G_i|}, \tag{54}$$

where G_i denotes the set of ground-truth entities in the i-th sample, and P_i denotes the set of entities predicted by the model. For every dataset, we strip leading and trailing whitespace, collapse consecutive whitespace characters, and remove invisible characters. Classification inputs are further truncated to 100 whitespace-separated tokens when constructing ICL prompts. We adopt a leave-one-domain-out protocol in which each dataset is treated as the target domain in turn, while the remaining datasets from the same task are pooled as labeled source domains. Additionally, all prompt templates are adopted from the BOSS benchmark.

For the NER task, where each sentence may contain multiple entities and does not have a naturally defined sentence-level class label, we construct a proxy label using the first non-O BIO tag in the sentence. We retain only person, location, and organization entities, while mapping all other entity types to O. Source instances sharing the same proxy label are then used to construct class prototypes by averaging their semantic representations. During retrieval, each candidate demonstration is scored by combining its semantic similarity to the query with the query's similarity to the prototype associated with the candidate's proxy label.

Language-model inference. For the main experiments, we locally deploy five LLMs of different families and scales: Llama-3.2-1B, Llama-3.1-8B, `Qwen3-1.7B`, Qwen3.5-2B, and Qwen3.5-9B from HuggingFace. All local models are loaded in FP16 and used without parameter updates. We use `all-MiniLM-L6-v2` as the embedding backbone for both representation learning and demonstration retrieval. For each target query, we prepend four retrieved input-label demonstrations to a task-specific instruction and perform deterministic greedy decoding. We set the maximum number of newly generated tokens to 3 for SA, TD, and NLI and to 20 for NER. The same prompt template, demonstration budget, and decoding configuration are used across all retrieval methods for a given LLM. For all target domains, we evaluate on the corresponding test sets. If a test set contains more than 5,000 samples, we randomly sample a subset of 5,000 examples for evaluation. To examine whether the results extend to larger models that cannot be deployed locally, we additionally compare CCL and our method using Qwen3.5-Plus and DeepSeek-V4-Flash through OpenAI-compatible APIs. Owing to API cost, these experiments use the first 200 examples of each target domain while retaining the same four-shot prompts and evaluation procedure.

D More Results

D.1 Detailed Results on Different Datasets

Table 4: Accuracy (%) on sentiment analysis datasets. Avg. is the arithmetic mean over the four datasets. Gap denotes the decrease in average accuracy relative to DRIFT.

Model	Method	DynaSent	SemEval	SST	Amazon	Avg.	Gap
	FT	42.91	59.48	64.71	74.56	60.41	N/A
Llama3.2-1B	kNN	45.89	44.62	56.51	51.30	49.58	13.51
	DICL	45.98	44.80	55.45	49.08	48.83	14.26
	UCS	46.33	45.06	57.33	51.48	50.05	13.04
	CCL	46.44	48.38	58.79	61.40	53.75	9.34
	DRIFT	50.61	62.30	63.42	76.02	**63.09**	–
Llama3.1-8B	kNN	60.04	51.20	69.40	84.12	66.19	3.35
	DICL	59.66	51.28	70.11	84.70	66.44	3.10
	UCS	59.38	50.80	69.28	83.00	65.62	3.92
	CCL	59.14	52.14	70.34	81.34	65.74	3.80
	DRIFT	60.19	62.42	71.63	83.92	**69.54**	–
Qwen3-1.7B	kNN	59.78	49.44	65.53	81.14	63.97	2.12
	DICL	59.98	49.12	64.36	82.28	63.93	2.16
	UCS	60.73	49.40	57.17	81.70	62.25	3.84
	CCL	59.95	51.04	65.89	79.88	64.19	1.90
	DRIFT	59.55	58.18	66.12	80.50	**66.09**	–
Qwen3.5-2B	kNN	55.67	45.50	62.37	83.02	61.64	3.47
	DICL	55.30	45.06	61.08	84.12	61.39	3.72
	UCS	54.86	44.82	62.37	83.94	61.50	3.61
	CCL	55.44	44.66	58.73	79.94	59.69	5.42
	DRIFT	57.32	56.28	65.18	81.64	**65.11**	–
Qwen3.5-9B	kNN	69.07	61.56	70.22	84.52	71.34	0.58
	DICL	68.69	61.90	70.81	85.02	71.61	0.31
	UCS	59.36	61.74	72.10	83.98	69.30	2.62
	CCL	67.99	63.32	71.28	81.28	70.97	0.95
	DRIFT	67.78	64.70	70.40	84.80	**71.92**	–
Qwen3.5-Plus	CCL	73.50	68.50	79.00	82.00	75.75	1.50
	DRIFT	74.00	70.00	84.50	80.50	**77.25**	–
DeepSeek-V4-Flash	CCL	81.00	68.50	74.00	85.00	77.13	2.50
	DRIFT	84.50	73.00	75.50	85.50	**79.63**	–

SA. Tab. 4 reports the results on four sentiment analysis datasets. DRIFT achieves the highest average accuracy across all seven LLMs, demonstrating consistent effectiveness across different model families and scales. The improvements are particularly pronounced on smaller models: compared with the strongest retrieval baseline, DRIFT improves the average accuracy by 9.34, 3.10, 1.90, and 3.47 percentage points on Llama3.2-1B, Llama3.1-8B, `Qwen3-1.7B`, and Qwen3.5-2B, respectively. Notably, DRIFT raises `Llama3.2-1B` from 53.75% to 63.09%, substantially narrowing the gap to the fully supervised FT baseline (60.41%) and even surpassing it on average. The gains remain consistent on stronger LLMs, reaching 71.92%, 77.25%, and 79.63% on Qwen3.5-9B, Qwen3.5-Plus, and DeepSeek-V4-Flash, respectively. Across individual target domains, DRIFT shows particularly strong improvements on SemEval, where conventional retrieval methods often suffer substantial degradation.

Table 5: Accuracy (%) on toxicity detection datasets. Avg. is the arithmetic mean over the four datasets. Gap denotes the decrease in average accuracy relative to DRIFT.

Model	Method	HateSpeech	ImplicitHate	ToxiGen	CivilComments	Avg.	Gap
	FT	79.26	70.02	71.14	74.72	73.78	N/A
Llama3.2-1B	kNN	67.84	58.04	53.32	33.62	53.20	17.65
	DICL	68.89	59.94	59.97	43.66	58.11	12.74
	UCS	67.15	60.30	62.23	42.50	58.05	12.80
	CCL	76.04	61.64	65.03	70.02	68.18	2.67
	DRIFT	78.09	64.78	69.28	71.24	**70.85**	–
Llama3.1-8B	kNN	77.20	63.14	66.89	74.12	70.34	2.66
	DICL	77.44	64.38	70.86	74.58	71.81	1.19
	UCS	76.96	53.92	71.34	73.52	68.94	4.06
	CCL	79.90	66.30	72.75	73.32	**73.07**	−0.07
	DRIFT	79.54	63.76	71.94	76.76	73.00	–
Qwen3-1.7B	kNN	61.26	48.94	56.52	23.48	47.55	3.29
	DICL	62.95	51.66	58.24	21.80	48.66	2.18
	UCS	60.53	51.71	58.78	22.18	48.30	2.54
	CCL	61.19	50.42	58.91	21.96	48.12	2.72
	DRIFT	64.21	52.60	63.03	23.50	**50.84**	–
Qwen3.5-2B	kNN	69.73	58.70	62.90	47.76	59.77	6.68
	DICL	70.54	57.24	64.49	44.12	59.10	7.35
	UCS	69.37	57.48	63.16	43.86	58.47	7.98
	CCL	75.26	61.20	67.59	50.80	63.71	2.74
	DRIFT	78.05	62.20	70.21	55.34	**66.45**	–
Qwen3.5-9B	kNN	79.30	57.94	78.59	73.46	72.32	2.86
	DICL	78.09	64.42	78.72	76.62	74.46	0.72
	UCS	77.76	63.70	80.19	72.14	73.45	1.73
	CCL	73.65	59.58	80.85	66.98	70.27	4.91
	DRIFT	83.08	62.78	79.65	75.22	**75.18**	–
Qwen3.5-Plus	CCL	83.00	50.00	86.00	71.00	72.50	1.13
	DRIFT	79.00	59.00	82.50	74.00	**73.63**	–
DeepSeek-V4-Flash	CCL	86.00	87.50	88.00	81.50	85.75	1.63
	DRIFT	86.50	89.00	90.00	84.00	**87.38**	–

TD. A similar overall trend is observed on toxicity detection in Tab. 5, where DRIFT achieves the best average performance on most LLMs and remains competitive on Llama3.1-8B. More importantly, the results reveal substantial dataset-specific variation. In particular, CivilComments presents a pronounced challenge for smaller LLMs: conventional retrieval methods achieve only 33.62-43.66% on `Llama3.2-1B` and 21.80–23.48% on `Qwen3-1.7B`. DRIFT substantially improves the former to 71.24%, approaching the supervised FT result of 74.72%, while the latter remains challenging at 23.50%. As model capacity increases, this discrepancy is considerably reduced, with DRIFT reaching 55.34%, 75.22%, and 84.00% on Qwen3.5-2B, Qwen3.5-9B, and DeepSeek-V4-Flash, respectively. These results highlight that multi-source OOD ICL can exhibit strong target-specific difficulty, while the effectiveness of disentangled retrieval depends jointly on the target-domain shift and the underlying LLM capability.

NLI. On NLI, DRIFT generally achieves the best average performance in Tab. 6, although the gains are more moderate than those on SA and TD. A notable dataset-specific pattern emerges on WANLI, where DRIFT consistently delivers strong performance across different LLMs. In contrast, the advantage is less pronounced on ANLI. As ANLI is constructed adversarially to expose model weaknesses, performance on this dataset may depend more heavily on the intrinsic adversar-

Table 6: Accuracy (%) on NLI datasets. Avg. is the arithmetic mean over the four datasets. Gap denotes the decrease in average accuracy relative to DRIFT.

Model	Method	ANLI	ContractNLI	WANLI	MNLI	Avg.	Gap
	FT	30.00	41.21	45.48	36.90	38.40	N/A
Llama3.2-1B	kNN	32.54	44.56	38.92	33.92	37.48	0.94
	DICL	33.12	46.59	40.52	34.62	**38.71**	−0.29
	UCS	32.73	44.57	39.40	33.50	37.55	0.87
	CCL	33.32	38.70	38.90	33.20	36.03	2.39
	DRIFT	32.47	38.87	43.80	38.54	38.42	–
Llama3.1-8B	kNN	34.84	45.75	41.08	52.64	43.58	0.20
	DICL	34.82	45.69	39.65	53.06	43.30	0.48
	UCS	34.73	42.46	39.96	53.36	42.63	1.15
	CCL	34.02	40.07	41.05	52.84	42.00	1.78
	DRIFT	32.70	44.47	45.28	52.66	**43.78**	–
Qwen3-1.7B	kNN	35.82	21.23	39.23	48.32	36.15	3.02
	DICL	34.30	25.84	37.45	48.34	36.48	2.69
	UCS	34.88	25.78	38.08	48.18	36.73	2.44
	CCL	32.77	24.55	40.30	48.22	36.46	2.71
	DRIFT	31.56	33.19	44.83	47.12	**39.17**	–
Qwen3.5-2B	kNN	34.69	37.32	39.90	44.80	39.18	0.69
	DICL	36.02	37.38	38.98	44.56	39.23	0.64
	UCS	34.96	37.20	38.85	44.82	38.96	0.91
	CCL	32.97	32.12	39.90	46.32	37.83	2.04
	DRIFT	32.76	36.81	44.98	44.94	**39.87**	–
Qwen3.5-9B	kNN	45.72	39.23	52.33	66.90	51.04	1.67
	DICL	46.72	34.03	51.28	68.56	50.15	2.56
	UCS	47.54	36.30	51.52	68.04	50.85	1.86
	CCL	47.85	33.61	53.70	68.88	51.01	1.70
	DRIFT	46.61	40.62	54.38	69.22	**52.71**	–
Qwen3.5-Plus	CCL	63.00	35.50	56.00	76.00	57.63	4.75
	DRIFT	64.50	42.00	62.50	80.50	**62.38**	–
DeepSeek-V4-Flash	CCL	74.00	43.00	73.00	80.00	67.50	2.63
	DRIFT	73.00	46.50	78.00	83.00	**70.13**	–

ial robustness and reasoning capability of the underlying LLM, rather than solely on robustness to distribution shift. This may limit the gains achievable through improved OOD demonstration retrieval alone. Nevertheless, DRIFT maintains strong overall performance across heterogeneous NLI domains, with particularly clear gains on stronger LLMs, improving over CCL by 4.75 and 2.63 percentage points on Qwen3.5-Plus and DeepSeek-V4-Flash, respectively.

NER exhibits substantially stronger model-specific variation than the other tasks, reflecting the different capabilities of LLMs in structured entity prediction in Tab. 7. Despite this variation, DRIFT achieves the highest average accuracy across all evaluated LLMs. The gains are particularly substantial on Llama models, improving over CCL by 5.58 and 6.00 percentage points on `Llama3.2-1B` and Llama3.1-8B, respectively, with consistent improvements across most target datasets. A particularly large gain is observed on Qwen3.5-Plus, where DRIFT improves the average accuracy from 63.50% to 71.67%, including a 14.60-point improvement on ENER. In contrast, Qwen3.5-9B performs poorly across all methods, suggesting that NER performance is strongly influenced by the underlying model's task-specific capability. Nevertheless, DRIFT retains the highest average accuracy even in this challenging setting. Overall, these results indicate that domain-disentangled retrieval remains effective under substantial variation in both target domains and underlying model capabilities.

D.2 More Ablition Results

Tab. 8–11 present the full ablation results across the four tasks. Overall, the ablations reveal complementary yet task-dependent contributions rather than uniform gains from every component. On SA, the complete model achieves the best average performance on both LLMs, while the relatively moderate degradation of individual variants suggests that different components provide complementary refinements. On TD, the effects are more dataset- and model-dependent, with certain ablations occasionally outperforming the full model, reflecting the substantial heterogeneity among toxicity datasets. In contrast, DA plays a particularly important role on NLI, where its re-

Table 7: Accuracy (%) on named entity recognition datasets. Avg. is the arithmetic mean over the four datasets. Gap denotes the decrease in average accuracy relative to DRIFT.

Model	Method	CoNLL	ENER	FewNERD	WNUT	Avg.	Gap
	FT	72.46	54.77	62.75	66.01	64.00	N/A
Llama3.2-1B	kNN	38.64	29.15	24.08	23.32	28.80	11.26
	DICL	36.93	27.26	24.59	22.56	27.84	12.22
	UCS	37.64	28.20	23.70	22.56	28.03	12.03
	CCL	45.27	32.05	27.22	33.38	34.48	5.58
	DRIFT	48.69	40.27	35.92	35.37	**40.06**	–
Llama3.1-8B	kNN	53.02	34.63	37.96	42.53	42.04	7.77
	DICL	52.04	31.84	39.34	40.85	41.02	8.79
	UCS	51.41	33.88	38.77	40.83	41.22	8.59
	CCL	54.73	33.99	39.43	47.10	43.81	6.00
	DRIFT	56.85	44.25	48.14	50.00	**49.81**	–
Qwen3-1.7B	kNN	43.40	40.03	27.76	38.26	37.36	5.73
	DICL	43.38	39.71	31.57	36.89	37.89	5.20
	UCS	43.61	41.19	30.50	39.63	38.73	4.36
	CCL	39.35	38.38	28.38	37.80	35.98	7.11
	DRIFT	45.58	43.44	38.51	44.82	**43.09**	–
Qwen3.5-2B	kNN	43.71	35.71	36.29	37.95	38.42	1.00
	DICL	43.08	35.28	37.62	37.20	38.30	1.12
	UCS	43.40	36.51	37.56	37.65	38.78	0.64
	CCL	39.83	32.24	33.91	34.60	35.14	4.28
	DRIFT	43.09	37.77	38.92	37.91	**39.42**	–
Qwen3.5-9B	kNN	12.76	11.32	11.30	14.02	12.35	3.04
	DICL	11.28	11.10	10.67	12.65	11.42	3.97
	UCS	10.93	11.24	10.97	14.18	11.83	3.56
	CCL	14.17	11.27	14.21	17.68	14.33	1.06
	DRIFT	13.74	17.28	12.10	18.45	**15.39**	–
Qwen3.5-Plus	CCL	74.85	55.48	59.63	64.04	63.50	8.17
	DRIFT	80.04	70.08	67.71	68.83	**71.67**	–
DeepSeek-V4-Flash	CCL	89.82	71.62	77.66	76.71	78.95	0.32
	DRIFT	88.04	72.41	78.90	77.72	**79.27**	–

Table 8: Full ablation results on sentiment analysis datasets. Avg. is the arithmetic mean over the four datasets. Bold and underlined values denote the best and second-best results for each model, respectively.

Model	Pro.	DA	Adv.	Orth.	DynaSent	SemEval	SST	Amazon	Avg.
Llama3.2-1B	✓	✓	✓	✓	50.61	62.30	63.42	**76.02**	**63.09**
	✗	✓	✓	✓	50.52	58.68	64.01	72.92	61.53
	✓	✗	✓	✓	**51.62**	59.98	65.18	75.30	63.02
	✓	✓	✗	✓	50.69	**62.84**	63.31	75.42	63.06
	✓	✓	✓	✗	49.88	59.86	**66.00**	75.22	62.74
Qwen3-1.7B	✓	✓	✓	✓	**59.55**	**58.18**	66.12	80.50	**66.09**
	✗	✓	✓	✓	59.09	55.64	66.59	80.76	65.52
	✓	✗	✓	✓	58.13	58.02	64.95	**81.11**	65.55
	✓	✓	✗	✓	58.25	55.22	66.58	80.50	65.14
	✓	✓	✓	✗	57.20	57.10	**67.29**	80.90	65.62

moval decreases the average accuracy from 38.42 to 34.95 on `Llama3.2-1B` and from 39.17 to 35.43 on `Qwen3-1.7B`. A similar tendency is observed on NER, especially for `Qwen3-1.7B`, where removing DA results in a 2.56-point drop. This suggests that class-conditioned cross-domain alignment is particularly beneficial for tasks involving fine-grained semantic relations or structured predictions. Meanwhile, Pro., Adv., and Orth. exhibit more task-specific effects, and the best-performing ablation occasionally varies across individual target datasets. Taken together, these results indicate that no single component universally dominates; rather, their integration provides more balanced and robust representations across heterogeneous domains.

D.3 More Visual Results

Fig. 8 provides the complete visualization of the learned semantic and domain representations across all evaluated datasets, covering sentiment analysis (SA), toxicity detection (TD), natural language

Table 9: Full ablation results on toxicity detection datasets. Avg. is the arithmetic mean over the four datasets. Bold and underlined values denote the best and second-best results for each model, respectively.

Model	Pro.	DA	Adv.	Orth.	HateSpeech	Implicit Hate	ToxiGen	Civil Comments	Avg.
Llama3.2-1B	✓	✓	✓	✓	78.09	64.78	69.28	71.24	70.85
	✗	✓	✓	✓	76.88	**67.20**	**70.35**	67.64	70.52
	✓	✗	✓	✓	76.99	64.34	68.35	**73.48**	70.79
	✓	✓	✗	✓	**79.90**	66.02	68.75	67.38	70.51
	✓	✓	✓	✗	78.41	65.26	68.62	71.82	**71.03**
Qwen3-1.7B	✓	✓	✓	✓	64.21	52.60	63.03	**23.50**	50.84
	✗	✓	✓	✓	63.52	**55.43**	**63.56**	22.60	**51.28**
	✓	✗	✓	✓	62.91	50.78	54.78	22.72	47.80
	✓	✓	✗	✓	**64.69**	54.70	62.10	20.38	50.47
	✓	✓	✓	✗	63.54	52.52	59.84	22.44	49.59

Table 10: Full ablation results on natural language inference datasets. Avg. is the arithmetic mean over the four datasets. Bold and underlined values denote the best and second-best results for each model, respectively.

Model	Pro.	DA	Adv.	Orth.	ANLI	ContractNLI	WANLI	MNLI	Avg.
Llama3.2-1B	✓	✓	✓	✓	**32.47**	38.87	43.80	38.54	**38.42**
	✗	✓	✓	✓	31.80	**41.39**	43.33	35.64	38.04
	✓	✗	✓	✓	28.63	27.93	43.30	**39.92**	34.95
	✓	✓	✗	✓	30.43	32.54	**45.43**	38.92	36.83
	✓	✓	✓	✗	31.33	28.29	44.38	39.20	35.80
Qwen3-1.7B	✓	✓	✓	✓	31.56	**33.19**	44.83	47.12	**39.17**
	✗	✓	✓	✓	**33.28**	30.80	43.05	**48.50**	38.91
	✓	✗	✓	✓	29.17	23.45	46.08	43.02	35.43
	✓	✓	✗	✓	31.13	26.31	**46.35**	44.72	37.13
	✓	✓	✓	✗	32.15	23.21	46.13	44.34	36.46

inference (NLI), and named entity recognition (NER). Consistent with the observations in the main text, z_s and z_d exhibit substantially different geometric structures across all datasets, with uniformly low CKA scores ranging from 0.01 to 0.12. Notably, this separation remains consistent across diverse tasks and source–target distributions, suggesting that the learned disentanglement is not specific to a particular dataset or task. Together, these results provide further evidence that our framework consistently separates semantic and domain-specific information under heterogeneous distribution shifts.

Fig. 9 further presents the representations along the first three principal components. The semantic and domain representations remain geometrically distinct across all 16 datasets, consistent with our 2D observations. This additional perspective shows that the observed separation is not merely an artifact of two-dimensional projection, further supporting the effectiveness of representation disentanglement in DRIFT.

E USE OF AI

AI-assisted tools were used to improve the clarity and linguistic quality of the manuscript and to assist in searching for relevant literature. All scientific content, methodological design, experimental analysis, and conclusions were developed and verified by the authors.

Table 11: Full ablation results on named entity recognition datasets. Avg. is the arithmetic mean over the four datasets. Bold and underlined values denote the best and second-best results for each model, respectively.

Model	Pro.	DA	Adv.	Orth.	CoNLL	ENER	FewNERD	WNUT	Avg.
Llama3.2-1B	✓	✓	✓	✓	**48.69**	**40.27**	**35.92**	35.37	**40.06**
	✗	✓	✓	✓	47.00	38.21	35.32	32.77	38.33
	✓	✗	✓	✓	48.44	38.57	33.49	35.54	39.01
	✓	✓	✗	✓	47.86	39.84	34.59	34.60	39.22
	✓	✓	✓	✗	48.51	35.80	33.38	**35.89**	38.39
Qwen3-1.7B	✓	✓	✓	✓	45.58	43.44	**38.51**	**44.82**	**43.09**
	✗	✓	✓	✓	**47.31**	42.48	38.38	43.14	42.83
	✓	✗	✓	✓	42.52	43.82	35.55	40.24	40.53
	✓	✓	✗	✓	45.14	44.48	38.17	39.63	41.86
	✓	✓	✓	✗	44.91	**45.43**	37.14	40.09	41.89

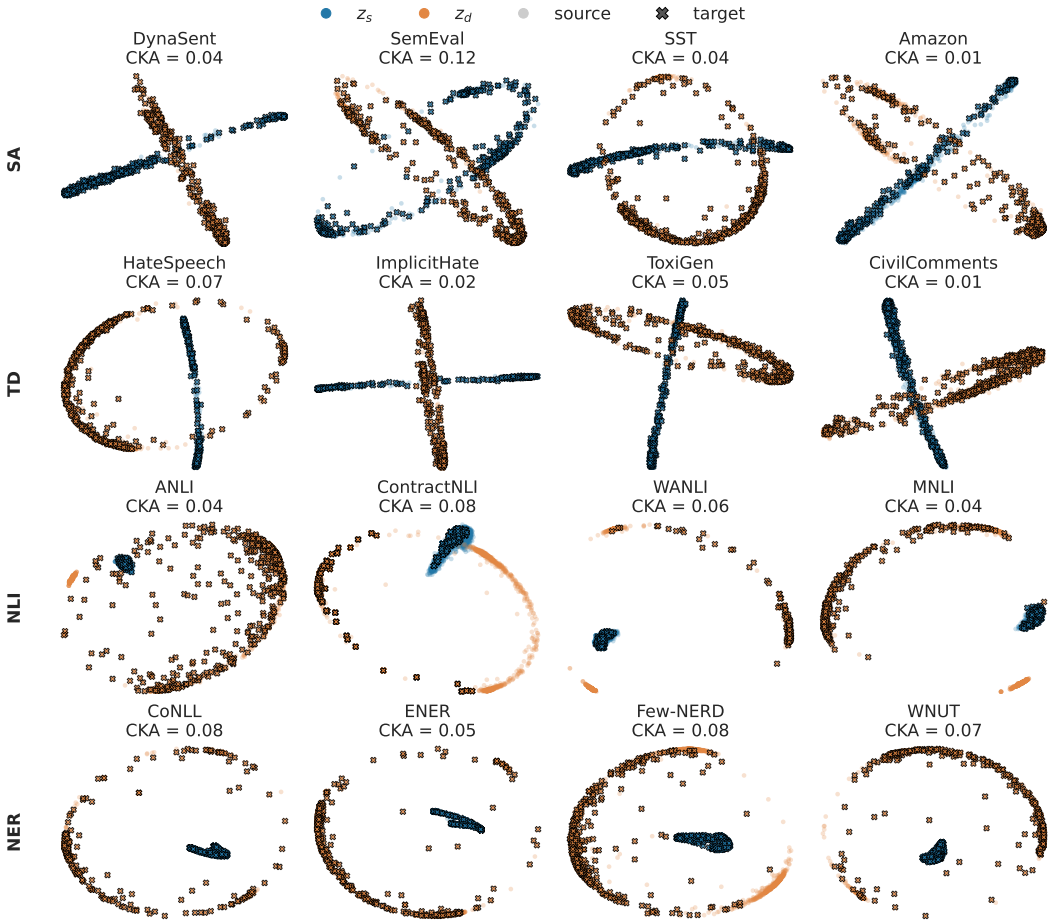


Figure 8: Visualization of semantic representations z_s and domain representations z_d across all evaluated datasets. Colors distinguish the two representation spaces, while marker types distinguish source and target samples. The consistently low CKA values indicate limited representational similarity between z_s and z_d across different tasks and distribution shifts.

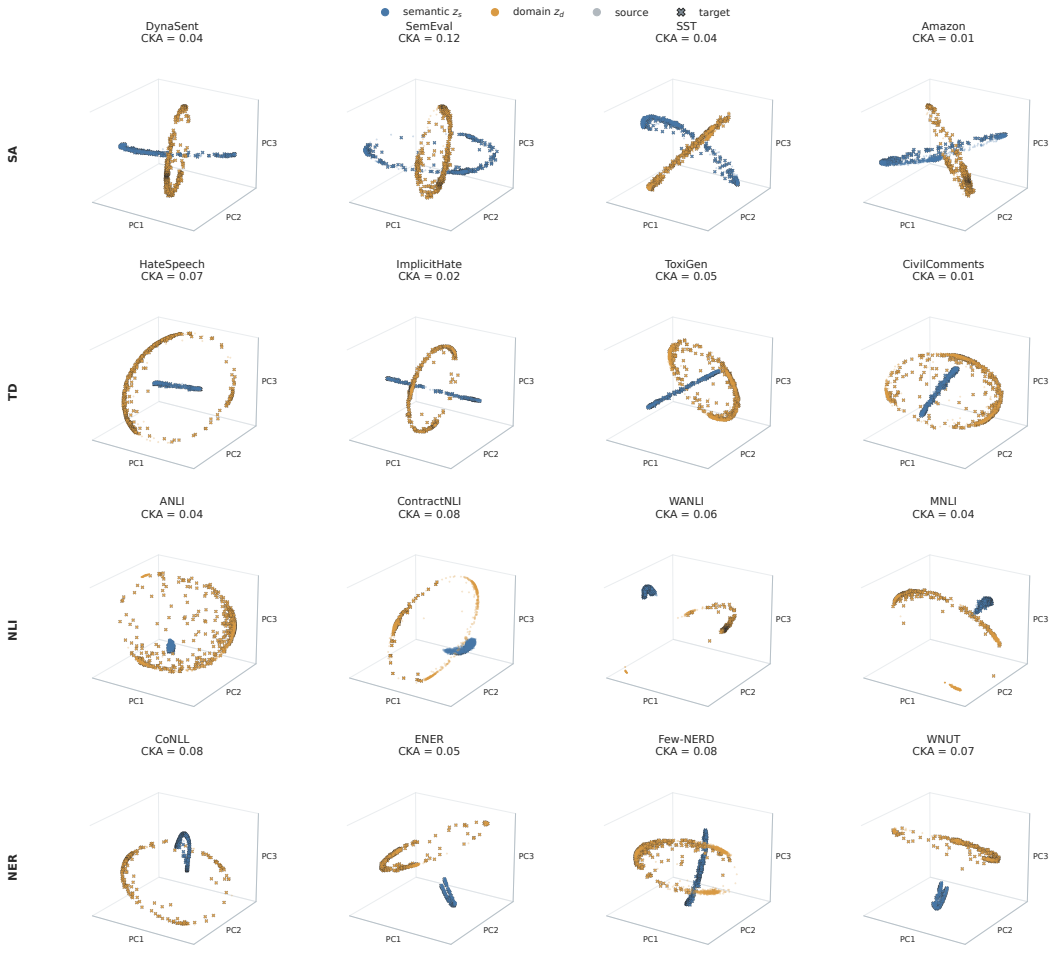


Figure 9: 3D-visualization of semantic representations z_s and domain representations z_d across all evaluated datasets.