AdaRetrieval: Towards Effective In-context Cross-domain Knowledge Transfer via Representation-adaptation-based Retrieval

Anonymous ACL submission

Abstract

Although large language models (LLMs) have advanced in logical reasoning, improving their performance traditionally requires extensive domain-specific annotations. This limits their applicability in specialized unseen domains where data collection is impractical. To mitigate this bottleneck, retrieval-augmented in-context learning with cross-domain demonstrations can boost LLM reasoning without updating model parameters. Despite substantial differences, implicit logical structures are shared across domains; however, current retrieval methods compute cosine similarity in unaligned latent spaces, favoring superficial lexical overlap over logical semantics. To retrieve structurally compatible cross-domain examples, we propose the AdaRetrieval knowledge transfer framework. AdaRetrieval projects source demonstrations into a target-aligned space using a lightweight residual mapping function. This projection is optimized via Maximum Mean Discrepancy (MMD) and a semantic regularization loss to preserve intrinsic logical relationships and prevent negative transfer. Experiments across various model scales show our method captures structural isomorphism and consistently outperforms baselines in cross-domain adaptation [1].

1 Introduction

Large Language Models (LLMs) have recently shown strong performance in complex reasoning tasks, such as code generation, mathematical problem-solving, and logical deduction (Wei et al., 2022; Achiam et al., 2023; Huang and Chang, 2023). Despite this progress, their reasoning capabilities often fall short of human proficiency (Valmeekam et al., 2023). Current methods for improving LLM reasoning typically rely on extensive human-annotated datasets. Approaches

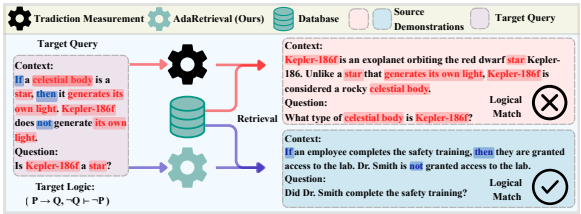


Figure 1: Domain shift in retrieval-based ICL. Traditional methods (left) prioritize literal similarity, retrieving lexically overlapping, logically irrelevant factoids. AdaRetrieval (right) aligns representations to retrieve structurally isomorphic cross-domain demonstrations, providing logical scaffolding ($P \rightarrow Q, \neg Q \vdash \neg P$) triggering LLM reasoning.

such as supervised fine-tuning (SFT) on reasoning trajectories or explicit Chain-of-Thought (CoT) prompting require carefully designed domain-specific exemplars (Ouyang et al., 2022; Zelikman et al., 2022). However, collecting expert-annotated data is often expensive or unfeasible for specialized applications (Yan et al., 2026b). This lack of domain-specific annotations restricts the generalization of existing reasoning techniques to out-of-domain scenarios (Zhou et al., 2023). To address this bottleneck, retrieval-augmented In-Context Learning (ICL) offers a practical alternative (Brown et al., 2020; Dong et al., 2024). By retrieving cross-domain demonstrations, LLMs can transfer structured reasoning knowledge from a data-rich source domain to a target domain directly within the context window, improving performance without parameter updates (Yan et al., 2026a; Lewis et al., 2020; Liu et al., 2026).

However, existing retrieval methodologies encounter a critical obstacle in this cross-domain setting. The fundamental challenge stems from the intrinsic geometry of the LLM's representation space: traditional retrieval methods directly compute cosine similarity within an unaligned latent space. Because source and target domain data aggregate

[1] Our implementation is available at https://anonymous.4open.science/r/AdaRetrieval-6DBE

onto entirely distinct low-dimensional representations, this process is akin to measuring distances across two fundamentally different coordinate systems. Consequently, it disproportionately biases the retrieval towards literal, surface-level structural similarities while completely ignoring the underlying logical semantic relevance, leading to the extraction of logically misaligned pseudo-positive demonstrations (as illustrated in Figure 1).

To address these issues, we propose AdaRetrieval, a cross-domain representation alignment framework that bridges the geometric gap between disparate logical domains before retrieval. Rather than computing similarities in a misaligned space, AdaRetrieval applies a lightweight non-linear residual mapping to project source domain demonstrations into a pseudo-target representation. Our training objective incorporates a Maximum Mean Discrepancy (MMD) loss to align the global feature distributions across domains within a Reproducing Kernel Hilbert Space (RKHS) (Gretton et al., 2012). To prevent this mapping from distorting the semantic relationships among source samples, we introduce a semantic regularization loss that maintains their pairwise cosine similarities. The combined loss balances domain alignment with semantic preservation.

Our main contributions are summarized as follows:

- We identify the representation misalignment as a primary bottleneck in cross-domain ICL retrieval, demonstrating through case studies that traditional similarity metrics fail to capture logical relevance under domain shift.
- We introduce **AdaRetrieval**, a lightweight alignment module that elegantly bridges source and target domains in the representation space using a residual MLP optimized by MMD and semantic regularization losses.
- Extensive empirical studies across various scales of models reveal that AdaRetrieval significantly and consistently outperforms existing baselines in cross-domain logical reasoning tasks.

We hope this work underscores the importance of representation alignment and motivates future research toward achieving robust cross-domain ICL.

2 Preliminaries

2.1 Domain Adaptation

Domain Adaptation (DA) is a specialized case of transfer learning that addresses the distribution mismatch between a labeled source domain $\mathcal{D}_S = \{(x_i^S, y_i^S)\}_{i=1}^{n_S}$ and a target domain $\mathcal{D}_T = \{x_j^T\}_{j=1}^{n_T}$. While both domains typically share the same label space $\mathcal{Y}$, they are characterized by different marginal distributions, i.e., $P_S(x) \neq P_T(x)$.

To mitigate covariate shift, Domain Adaptation (DA) learns a domain-invariant feature extractor $h_\theta : \mathcal{X} \to \mathbb{R}^d$ that maps inputs into a shared latent space, minimizing the discrepancy between domain distributions. The optimization is generally formulated as:

$$\min_{\theta,\phi} \quad \mathcal{L}_S\big(f_\phi(h_\theta(x^S)), y^S\big) + \lambda \cdot \mathrm{dist}\big(h_\theta(P_S), h_\theta(P_T)\big)$$

Here, $\mathcal{L}_S$ is the source supervised loss, f_ϕ the classifier, and $\mathrm{dist}(\cdot,\cdot)$ the domain alignment divergence. Minimizing this objective enables target-domain generalization without extensive annotations.

2.2 Retrieval-Augmented In-Context Learning

Retrieval-augmented ICL enhances the model's capabilities by retrieving relevant demonstration pairs from a corpus $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$ via vector similarity. We utilize an encoding function:

$$\phi : \mathcal{X} \to \mathbb{R}^k,$$

Candidate inputs are indexed into a retrieval database $\mathcal{I} = \{\phi(d_i)\}_{i=1}^{N}$. For a target query q, the top-K demonstrations are retrieved based on the representation $\phi(q)$.

$$\mathcal{R}(\mathbf{x}_{\text{test}}) = \underset{(\mathbf{x}_i,\mathbf{y}_i)\in\mathcal{D}}{\arg\,\mathrm{topK}}\ \mathrm{sim}(\phi(\mathbf{x}_{\text{test}}), \phi(\mathbf{x}_i)),$$

where $\mathrm{sim}(\cdot,\cdot)$ is a similarity metric. The language model $\mathcal{M}$ then generates the ICL prediction $\mathbf{y}_{\text{test}}$ by conditioning on the target query concatenated with the retrieved exemplars $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{K}$:

$$\mathbf{y}_{\text{test}} \sim \mathcal{P}_{\mathcal{M}}\left(\mathbf{y}_{\text{test}} \,\middle|\, \underbrace{\mathbf{x}_1, \mathbf{y}_1, \ldots, \mathbf{x}_K, \mathbf{y}_K, \mathbf{x}_{\text{test}}}_{\textbf{In-context prompt}}\right).$$

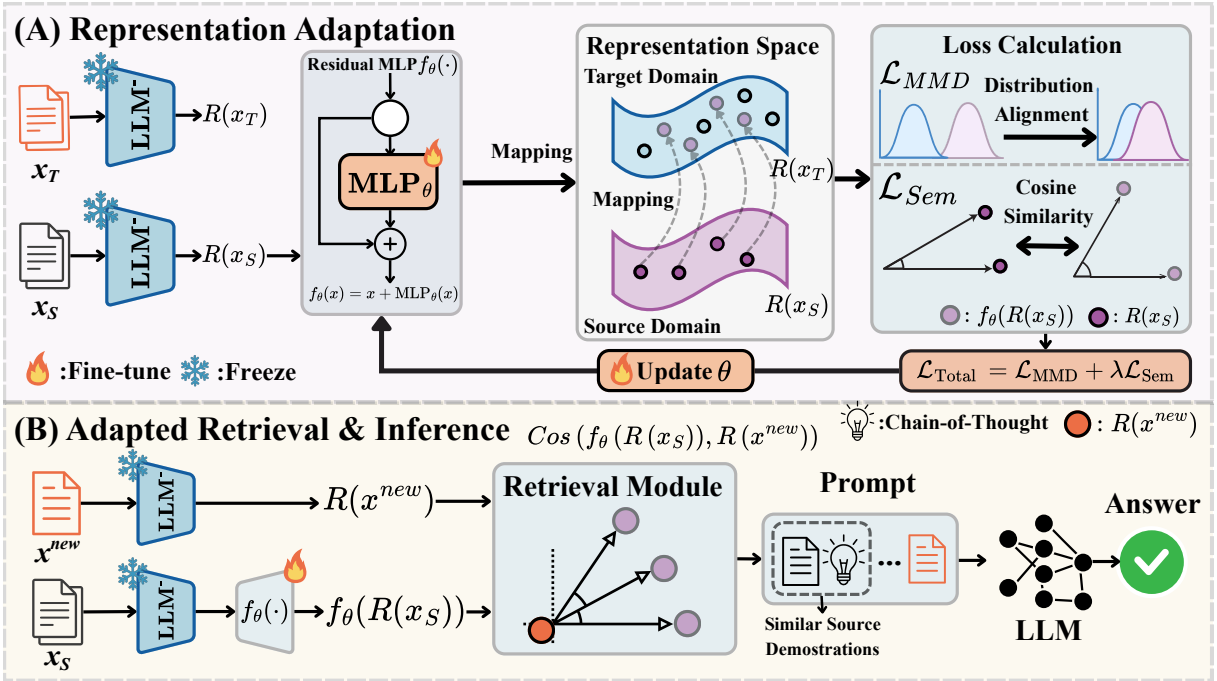


Figure 2: `AdaRetrieval` framework overview. (A) **Representation Adaptation**: Residual MLP adapter f_θ projects source representations into a target-aligned latent space. The objective combines Maximum Mean Discrepancy ($\mathcal{L}_{MMD}$) for distribution alignment and semantic regularization ($\mathcal{L}_{Sem}$) to preserve the intrinsic pairwise structures of source samples. (B) **Adapted Retrieval & Inference**: The trained adapter maps source demonstrations to bridge domain gaps. Logically isomorphic exemplars retrieved via aligned-space similarity facilitate cross-domain in-context reasoning.

By conditioning reasoning on external evidence, retrieval reduces dependence on purely parametric knowledge.

3 Method

As aforementioned, traditional retrieval methods operate in unaligned latent spaces, typically favoring lexical overlap over logical semantics. To mitigate domain shift in in-context learning, AdaRetrieval aligns cross-domain representations prior to retrieval. As Figure 2 illustrates, our framework uses a semantics-preserving residual function to project source demonstrations into pseudo-target representations. To structurally align the domains, the adapter is trained with a dual optimization objective that maintains pairwise cosine similarities. Retrieving these aligned demonstrations enables robust zero-shot reasoning with the frozen language model. The following subsections detail the formulation, adaptation, training, and inference procedures.

3.1 Representation Adaptation

As shwon in Part A in Figure 2, we decompose an LLM sequentially as $M = LLM^- \circ LLM^+$, where $LLM^-(\cdot)$ extracts features from the initial l layers, and $LLM^+(\cdot)$ derives the prediction from the remaining $L - l$ layers. We define an annotated source dataset $D_S = \{(x_S, c_S, y_S)\}$ comprising inputs, CoT rationales, and answers, alongside an unannotated target dataset $D_T = \{x_T\}$ to simulate real-world label scarcity.

In a zero-shot cross-domain setting, latent representations naturally form distinct low-dimensional manifolds, $\mathcal{M}_S$ and $\mathcal{M}_T$. To bridge this representational gap via latent intervention, we introduce an adapter $A_\theta : \mathbb{R}^d \to \mathbb{R}^d$. Implemented as a residual mapping, A_θ explicitly models domain shifts while preserving core semantic features. As depicted in Figure 2(A), we initially extract the original hidden states for both domains at the l-th layer:

$$h_S = LLM^-(x_S), \quad h_T = LLM^-(x_T) \quad (1)$$

To mitigate the severe geometric distortion caused by computing cosine similarity across misaligned manifolds, we propose a latent representation adaptation strategy. We project the source domain onto the target manifold using a residual Multilayer Perceptron (MLP) adapter, $A_\theta(\cdot)$, de-

signed to preserve intrinsic semantics:

$$A_\theta(h_S) = h_S + \text{MLP}_\theta(h_S) \tag{2}$$

This residual connection explicitly retains the foundational features of the source demonstrations, yielding the aligned pseudo-target representation for a source instance x_S as $\tilde{h}_S = A_\theta(h_S)$, where $h_S = LLM^-(x_S)$.

3.1.1 Training Objectives

The mapping function $A_\theta(\cdot)$ is optimized via a dual-objective training strategy that balances distribution alignment with semantic preservation.:

Distribution Alignment Loss We apply Multiple Kernel MMD within an RKHS to minimize the divergence between the mapped source $\{\tilde{h}_S\}$ and target $\{h_T\}$ feature distributions:

$$\mathcal{L}^2_{MMD}(\mathcal{D}_S, \mathcal{D}_T) = \underbrace{\frac{1}{N^2}\sum_{i,j} k(\tilde{h}_{S_i}, \tilde{h}_{S_j})}_{\textbf{Source Domain}} + \underbrace{\frac{1}{M^2}\sum_{i,j} k(h_{T_i}, h_{T_j})}_{\textbf{Target Domain}} - \underbrace{\frac{2}{NM}\sum_{i,j} k(\tilde{h}_{S_i}, h_{T_j})}_{\textbf{Cross Domain}} \tag{3}$$

where $\phi(\cdot)$ is the kernel function mapping.

Semantic Regularization To prevent negative transfer and explicitly preserve the relative topological structure among source samples, we introduce a pairwise similarity preservation loss:

$$\mathcal{L}_{Sem} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\Big(\cos(\tilde{h}_{S_i}, \tilde{h}_{S_j}) - \cos(h_{S_i}, h_{S_j})\Big)^2\Big)^2 \tag{4}$$

The overall objective function is a weighted sum of the aforementioned losses:

$$\mathcal{L}_{Total} = \mathcal{L}_{MMD} + \lambda\mathcal{L}_{Sem} \tag{5}$$

3.2 Adapted Retrieval & Inference

As shwon in Part B in Figure 2, during inference for a new target query $x_T \in \mathcal{D}_T$, the base LLM remains frozen. To execute zero-shot reasoning, we bypass the misaligned space by evaluating cosine similarity between the adapted source demonstrations and the original target query:

Algorithm 1 `AdaRetrieval`

Require: Source data $D_S = \{(x_S, c_S, y_S)\}$, Target data $D_T = \{x_T\}$, LLM component $LLM^-(\cdot)$, full LLM $\mathcal{P}_{LM}$
Require: Hyperparameters λ (regularization weight), η (learning rate), K (retrieval size)
Ensure: Trained adapter A^*_θ, Predicted answer $\hat{y}$

Phase 1: Representation Adaptation
while not converged **do**
  Sample batches $\{x_S\} \sim \mathcal{D}_S$, $\{x_T\} \sim \mathcal{D}_T$
  Encode states: $h_S \leftarrow LLM^-(x_S)$, $h_T \leftarrow LLM^-(x_T)$
  Project source representations: $\tilde{h}_S \leftarrow A_\theta(h_S) = h_S + \text{MLP}_\theta(h_S)$
  Compute alignment & semantic losses: $\mathcal{L}_{MMD}(\tilde{h}_S, h_T)$, $\mathcal{L}_{Sem}(\tilde{h}_S, h_S)$
  Update adapter parameters: $\theta \leftarrow \theta - \eta\nabla_\theta\big(\mathcal{L}_{MMD} + \lambda\mathcal{L}_{Sem}\big)$
end while
Phase 2: Adapted Retrieval and Inference
Given target query $x_{new} \in \mathcal{D}_T$, encode $h_{new} \leftarrow LLM^-(x_{new})$
Retrieve top-K demonstrations $\mathcal{D}^*_K \subset \mathcal{D}_S$ by maximizing $Score(x_S, x_{new}) = \cos\left(A^*_\theta(LLM^-(x_S)), h_{new}\right)$
return Prediction $\hat{y} \sim \mathcal{P}_{LM}\big([\mathcal{D}^*_K \oplus x_{new}]\big)$

$$\text{Score}(x_S, x_T) = \cos(A_\theta(LLM^-(x_S)), LLM^-(x_T)) = \cos(\tilde{h}_S, h_T) \tag{6}$$

We retrieve the top-K most logically relevant source demonstrations, denoted as an ordered set $\mathcal{D}^*_K = \{(x_{S_1}, y_{S_1}), \ldots, (x_{S_K}, y_{S_K})\}$. Finally, to generate the predicted answer $\hat{y}$ for the target query x_T, we formulate the cross-domain In-Context Learning (ICL) inference process:

$$\hat{y} \sim \mathcal{P}_{\text{LM}}\left(\underbrace{[(x_{S_1}, y_{S_1}), \ldots, (x_{S_k}, y_{S_k})]}_{\textbf{Source Domain}}, \underbrace{x_T}_{\textbf{Target Domain}}\right) \tag{7}$$

As shown in Algorithm 1, this lightweight representation alignment framework empowers the frozen LLM to perform robust zero-shot reasoning for cross-domain target queries efficiently and effectively.

Table 1: **Comprehensive cross-domain reasoning results.** Our method consistently outperforms zero-shot inference and existing retrieval baselines across various source-to-target domain pairs. The evaluation includes both main backbones (14B/32B/12B/27B) and smaller variants (3B/4B/7B/8B). Best results among same-sized models are highlighted. Avg. denotes the average score across all tasks.

	Source Domain → Target Domain									Source Domain → Target Domain								
	Qwen-2.5-3B-Instruct									Gemma-3-4B-it								
Method	*P-L*	*L-P*	*F-L*	*L-F*	*P-F*	*F-P*	*G-L*	*C-L*	Avg.	*P-L*	*L-P*	*F-L*	*L-F*	*P-F*	*F-P*	*G-L*	*C-L*	Avg.
Zero-shot	41.7	52.8	41.7	63.2	63.2	52.8	40.5	40.0	49.5	43.5	54.0	43.5	64.5	64.5	54.0	42.0	41.5	50.9
Embed	45.3	54.3	45.3	60.8	64.2	51.8	44.0	43.5	51.2	46.5	55.5	46.5	62.5	65.5	53.0	45.0	44.5	52.4
BM25	46.0	54.3	44.0	59.8	63.2	51.8	44.5	43.5	50.9	47.5	55.5	45.5	61.5	64.5	53.0	45.5	44.5	52.2
ConE	51.3	52.8	46.7	63.7	60.3	53.5	46.0	44.5	52.4	52.5	54.0	48.0	65.0	61.5	55.0	47.5	46.0	53.7
DIN	49.2	57.8	49.4	63.7	66.4	55.5	53.5	52.5	56.0	50.5	59.0	50.5	65.0	67.5	57.0	55.0	54.0	57.3
`AdaRetrieval`	52.0	54.1	50.6	66.1	68.6	56.0	54.0	53.0	**56.8**	53.5	55.5	52.0	67.5	69.5	57.5	55.5	54.5	**58.2**
	Qwen-2.5-7B-Instruct									Llama-3.1-8B-Instruct								
Zero-shot	62.0	62.5	62.0	69.6	69.6	62.5	60.0	59.5	63.5	63.5	64.0	63.5	71.0	71.0	64.0	61.5	61.0	64.9
Embed	64.7	62.0	62.3	68.6	66.2	62.7	62.5	61.0	63.8	66.0	63.5	64.0	70.0	67.5	64.5	64.0	62.5	65.3
BM25	60.7	61.2	61.0	69.1	71.1	63.7	59.5	58.0	63.0	62.0	62.5	62.5	70.5	72.5	65.0	61.0	59.5	64.4
ConE	63.0	61.3	61.7	63.2	73.5	62.5	64.0	62.0	63.9	64.5	63.0	63.0	64.5	75.0	64.0	65.5	63.5	65.4
DIN	67.5	65.6	69.4	71.2	72.0	67.5	69.0	67.5	68.7	69.0	67.0	71.0	72.5	73.5	69.0	70.5	69.0	70.2
`AdaRetrieval`	71.0	67.5	70.6	71.5	74.0	67.6	69.5	68.0	**69.9**	72.5	69.0	72.0	73.0	75.5	69.5	71.0	69.5	**71.5**
	Qwen-2.5-14B-Instruct									Gemma-3-12B-it								
Zero-shot	79.7	67.7	79.7	68.1	68.1	67.7	77.0	76.5	73.1	78.0	66.0	78.0	66.5	66.5	66.0	75.5	75.0	71.4
Embed	77.3	71.5	76.7	74.0	72.1	70.8	78.0	77.5	74.7	75.5	70.0	75.0	72.5	70.5	69.0	76.5	76.0	73.1
BM25	73.7	71.3	78.3	74.0	75.0	72.7	79.0	78.5	75.3	72.0	69.5	76.5	72.5	73.5	71.0	77.5	77.0	73.7
ConE	78.7	70.0	79.3	70.6	75.5	73.8	80.0	79.5	75.9	77.0	68.5	77.5	69.0	74.0	72.0	78.5	78.0	74.3
DIN	80.8	71.4	79.2	75.9	75.9	72.9	80.5	80.0	77.1	79.0	70.0	77.5	74.5	74.5	71.5	79.0	78.5	75.6
`AdaRetrieval`	81.3	71.8	82.0	76.9	77.4	71.8	81.0	80.5	**77.8**	80.0	70.5	80.5	75.5	76.0	70.5	79.5	79.0	**76.4**
	Qwen-2.5-32B-Instruct									Gemma-3-27B-it								
Zero-shot	83.0	70.7	83.0	71.1	71.1	70.7	81.0	80.0	76.3	82.0	69.5	82.0	70.0	70.0	69.5	80.0	79.0	75.3
Embed	83.3	71.8	83.7	72.5	70.6	71.3	82.0	81.5	77.1	82.5	70.5	82.5	71.5	69.5	70.0	81.0	80.5	76.0
BM25	86.7	74.0	85.7	71.6	74.5	70.7	83.0	82.5	78.6	85.5	73.0	84.5	70.5	73.5	69.5	82.0	81.5	77.5
ConE	86.0	72.8	83.7	72.1	72.5	69.3	83.5	83.0	77.9	85.0	71.5	82.5	71.0	71.5	68.0	82.5	82.0	76.8
DIN	86.8	70.7	85.2	77.9	77.9	72.2	85.5	84.5	80.1	85.5	69.5	84.0	76.5	76.5	71.0	84.5	83.5	78.9
`AdaRetrieval`	89.0	72.5	89.6	75.0	75.9	74.0	86.0	85.0	**80.9**	88.0	71.5	88.5	74.0	75.0	73.0	85.0	84.0	**79.9**

4 Experiments

In this section, we evaluates `AdaRetrieval`'s cross-domain reasoning adaptation by addressing three core research questions:

4.1 Backbone Model

We evaluate `AdaRetrieval` across a diverse set of open-source large language models to ensure our findings generalize across different architectures and parameter scales. Our experiments utilize **Gemma-3-it** (12B and 27B) (Team, 2025), and the **Qwen-2.5-Instruct** (3B, 7B,14B and 32B) (Yang et al., 2024). By encompassing both moderate and larger capacities, we thoroughly demonstrate the scalability of our approach.

4.2 Datasets & Tasks

We study cross-domain reasoning between mathematical and logical tasks. We use **GSM8K** (Cobbe et al., 2021) for mathematical reasoning, and **LogicalDeduction** (Nguyen et al., 2025), **FOLIO** (Han et al., 2022), **ProofWriter** (Tafjord et al., 2021) for logical reasoning, and **CommonSenseQA** for commonsense reasoning (Talmor et al., 2019).

4.3 Baselines

To benchmark the effectiveness of `AdaRetrieval`, we compare our framework against a standard **Zero-Shot** baseline, alongside recent and representative methods categorized into four main paradigms:

- **Zero-shot** (Wei et al., 2022): Evaluates the base model's direct reasoning capabilities without providing any external context, demonstrations, or parameter updates.
- **Embedding** (Lewis et al., 2020): A dense vector retrieval baseline that fetches semantically relevant context using neural text embeddings.
- **BM25** (Robertson and Zaragoza, 2009): A non-trainable sparse text retrieval method that fetches relevant context based on exact keyword matching.
- **ConE** (Peng et al., 2024): A context intervention method that utilizes perplexity-based reranking to select the most helpful retrieved examples.
- **DIN-Retrieval** (Yan et al., 2026a): Universal-neurons-based Retrieval

4.4 RQ1: Empirical Evaluation on Cross-Domain Benchmarks

Figure 3: Cross-domain retrieval case study. Traditional retrieval (**Middle**) gets trapped by superficial lexical overlap in unaligned spaces, causing non-deterministic branching and negative transfer. **AdaRetrieval** (**Bottom**) captures structural isomorphism via representation alignment. It successfully retrieves a zero-lexical-overlap demonstration sharing the identical linear transitive chaining structure ($X \to Y \to Z$), providing precise logical scaffolding for the target query.

Target Query (LogicalDeduction): Fact: $(\text{Yellow} < \text{Red}) \wedge (\text{Red} < \text{Green}) \implies$ New Fact: $(\text{Yellow} < \text{Red} < \text{Green})$

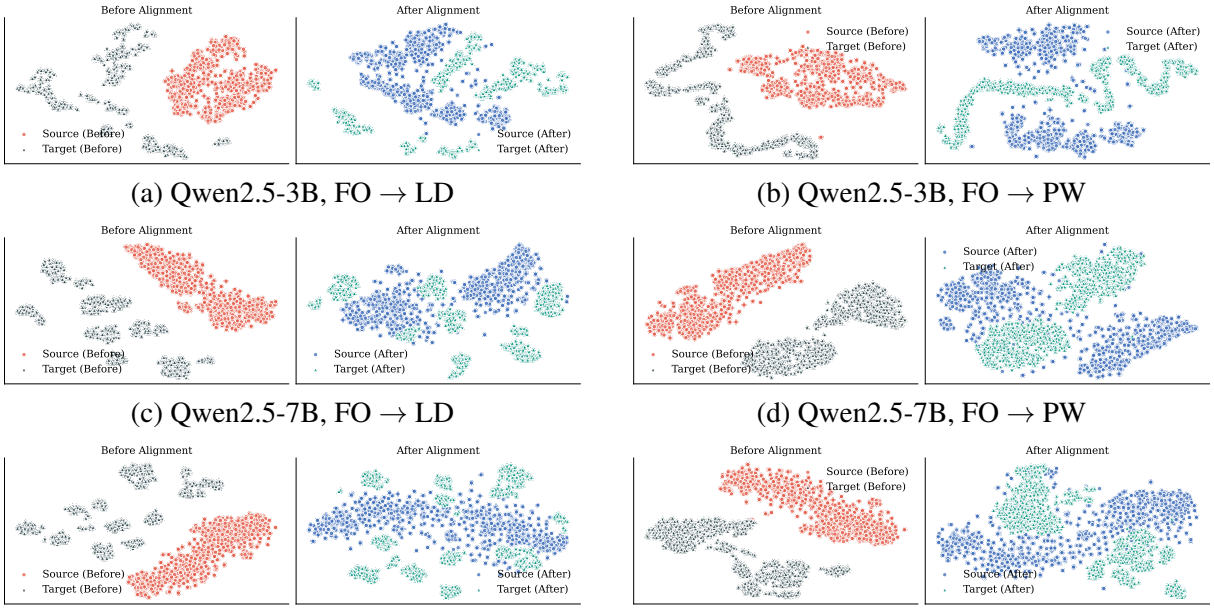
Context:
The following paragraphs each describe a set of three objects arranged in a fixed order. The statements are logically consistent within each paragraph. On a shelf, there are three books: a yellow book, a green book, and a red book. The red book is to the left of the green book. The yellow book is to the left of the red book.
Question:
Which of the following is true?
A) The yellow book is the second from the left. B) The green book is the second from the left. C) The red book is the second from the left. **Book:Text matching**

BM25 (FOLIO): Fact: $(P \vee Q) \implies$ Branching: $[Q \to \text{Dead End}] \cup [P \to \text{Uncertain matching with } \exists]$

Context:
All books written by Tomas have sold more than 1 million copies. Some books that have won Hugo Awards are written by Tomas. All books about the future are inspired by Shakespeare. 2001 is a book either inspired from Shakespeare or it has sold more than 1 million copies. 2001 is a book that has either sold more than 1 million copies or it was written by Tomas.
Question:
Based on the above information, is the following statement true, false, or uncertain? 2001 is a book that has won a Hugo Award.
A) True B) False C) Uncertain

AdaRetrieval (FOLIO): Fact: $(X < Y) \wedge (Y < Z) \implies$ New Fact: $(X < Z)$

Context:
If entertainment is interesting, it will be popular. All popular things are well-known.
Question:
Based on the above information, is the following statement true, false, or uncertain? If entertainment is interesting, it will be well-known.
A) True B) False C) Uncertain



(a) Qwen2.5-3B, FO → LD (b) Qwen2.5-3B, FO → PW

(c) Qwen2.5-7B, FO → LD (d) Qwen2.5-7B, FO → PW

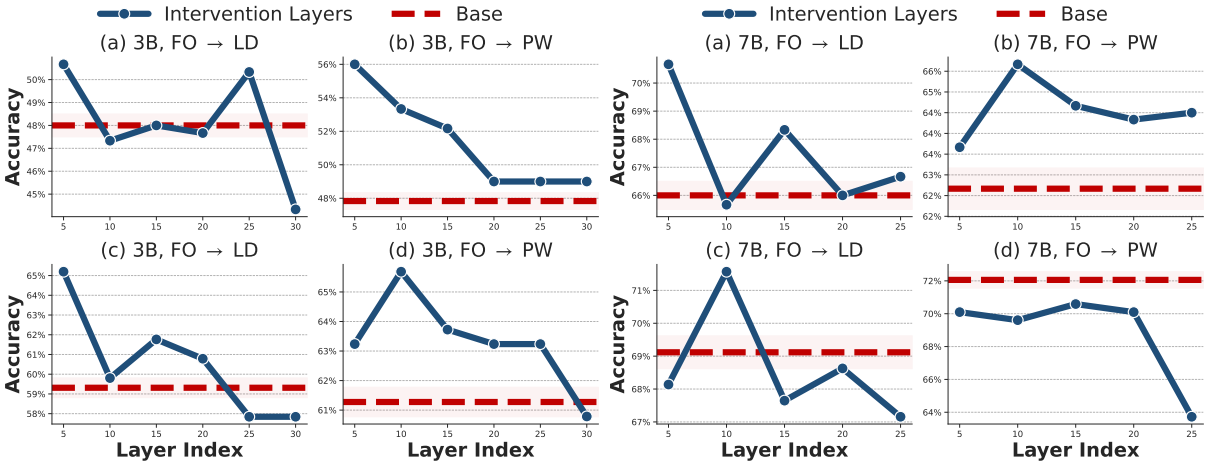
(e) Qwen2.5-14B, FO → LD (f) Qwen2.5-14B, FO → PW

Figure 4: T-SNE visualization of feature distributions before and after alignment across different Qwen2.5 model sizes on the CS → FO task. The aligned features demonstrate better domain overlap.

AdaRetrieval consistently establishes a new state-of-the-art across all evaluated model families and sizes, demonstrating robust generalization from small-scale to large-scale architectures. On the Qwen-2.5-3B-Instruct backbone, **AdaRetrieval** achieves an average accuracy of 56.7%, representing a 4.2% absolute improvement over the Zero-shot baseline (52.5%) and outperforming the strongest retrieval baseline, ConE, which stands at 54.7%. This performance gap persists and often widens as model capacity increases. For the largest evaluated configuration, Gemma-3-27B-it, **AdaRetrieval** reaches an average of 85.0%, which is a 7.8% gain over the Zero-shot performance of 77.2% and a 2.6% lead over the specialized ConE method at 82.4%. The proposed method shows exceptional performance in specific domain transitions that require deep logical alignment rather

Table 2: Ablation study of **AdaRetrieval** across different backbones (`Qwen2.5-32B-Instruct` and `Gemma3-27B-it`). We report the impact of removing the MMD penalty ($\mathcal{L}_{MMD}$) and the reasoning loss ($\mathcal{L}_{Reason}$) on cross-domain transfer performance.

Backbone	Ablation	*PW→LD*	*LD→PW*	*FO→LD*	*LD→FO*	*PW→FO*	*FO→PW*	*GS→LD*	*CS→LD*	Avg.
Qwen2.5-32B -Instruct	**AdaRetrieval**	89.0	72.5	89.6	75.0	86.0	74.0	75.9	85.0	**80.9**
	- $\mathcal{L}_{MMD}$	85.5	70.2	84.3	72.1	82.4	71.0	73.1	82.2	77.6
	- $\mathcal{L}_{Reason}$	84.1	68.8	81.5	70.5	79.8	69.2	71.4	80.7	75.8
Gemma3-27B -it	**AdaRetrieval**	88.0	71.5	88.5	74.0	85.0	73.0	75.0	84.0	**79.9**
	- $\mathcal{L}_{MMD}$	84.6	69.1	83.2	70.5	81.5	69.8	72.0	81.3	76.5
	- $\mathcal{L}_{Reason}$	83.2	67.5	80.5	68.8	78.4	68.0	70.2	79.8	74.6



Qwen2.5-3B-Instruct Qwen2.5-7B-Instruct

Figure 5: Impact of latent intervention at different layer indices on cross-domain reasoning accuracy across various model scales and task transitions.

than surface-level pattern matching. In the P-L transition for Gemma-3-4B-it, **AdaRetrieval** yields 73.0%, significantly surpassing the 60.7% achieved by the Zero-shot approach and the 55.3% by Embedding-based retrieval. A similar trend is observed in the F-L task on Qwen-2.5-14B-Instruct, where **AdaRetrieval** reaches 81.6%, whereas traditional BM25 and Embedding-based retrieval lag behind at 78.3% and 76.7%, respectively. These results underscore the effectiveness of the representation alignment strategy in bridging the gap between disparate logical domains such as Propositional Logic (P) and First-order Logic (F), where lexical overlap is minimal. We further conduct a case study to validate the efficacy of our method. As illustrated in Figure 3, traditional retrieval baselines operate in unaligned latent spaces and get trapped by superficial lexical overlap. Consequently, this retrieval process yields misaligned exemplars with branched, non-deterministic chaining topologies, ultimately inducing negative transfer that impairs downstream reasoning performance. Conversely, **AdaRetrieval** overcomes superficial discrepancies via latent manifold alignment. Notably, despite a complete absence of lexical overlap, it successfully retrieves a perfectly isomorphic demonstration sharing the identical deterministic linear transitive chaining structure. Consequently, transferring the reasoning pattern facilitates robust knowledge transfer, thereby ensuring accurate structural mapping.

4.5 RQ2: Efficacy of Representation Alignment

In this subsection, we focus on validating the actual effect of our representation alignment. We qualitatively assess the efficacy of our latent-space adaptation by visualizing the representation distributions of both source and target domains before and after applying the **AdaRetrieval** framework. Employing t-SNE dimensionality reduction to project the high-dimensional hidden states into a two-dimensional space, we first observe the inherent domain shift present in the base model. As illustrated in the left panel of Figure 4, the source representations and the target queries aggregate into distinctly isolated low-dimensional data representations. This severe geometric disjointness visually corroborates why traditional retrieval methods, which directly compute cosine similarity within

this unaligned latent space, fail to capture cross-domain relevance and instead retrieve logically misaligned pseudo-positives.

Conversely, the After Alignment panel of Figure 4 illustrates a profound structural shift following our proposed intervention. By mapping the source domain demonstrations through the trained residual adapter, the resulting pseudo-target representations effectively bridge the spatial gap. Optimized by the Maximum Mean Discrepancy objective, the mapped source features now highly overlap with the target domain's reasoning manifold. This confirms that **AdaRetrieval** successfully projects cross-domain data into a shared geometric space, ensuring that subsequent retrieval operations are accurately grounded in isomorphic logical semantics rather than superficial literal similarities.

4.6 RQ3: Further Analysis and Ablation Studies

To systematically evaluate the contributions of individual optimization objectives, we executed ablation experiments on Qwen-2.5 and Gemma-3 architectures. Removing the Maximum Mean Discrepancy penalty markedly undermines cross-domain transfer performance because the adapter is incapable of reconciling the source and target latent representations within the RKHS. Consequently, retrieval reverts to a dependence on surface-level textual overlaps rather than underlying semantic logic; similarly, ablating the semantic reasoning objective drastically compromises precision. Since this regularization term retains the mutual geometric configurations of source samples, its removal warps the inherent reasoning connections and triggers detrimental transfer effects. These results demonstrate that holistic distribution matching and fine-grained semantic retention are indispensable for stable cross-domain in-context learning.

Furthermore, our analysis indicates that latent-space adaptation achieves peak efficacy within the middle-to-higher layers. Shallow-layer interventions offer negligible improvements because they primarily operate on rudimentary surface patterns. In contrast, final-layer adaptations tend to hamper model efficacy by interfering with the model's unembedding and token-generation processes. This substantiates the hypothesis that domain-robust deduction mechanisms are largely concentrated inside the intermediate and deep feature subspaces.

5 Related Work

5.1 Generalization in LLM

Large language models (LLMs) often degrade under domain shifts (Öncel et al., 2024; Oh et al., 2025; Li et al., 2024). Existing approaches, such as data-centric adaptation (Wang et al., 2024), prompt calibration (Zhao et al., 2021; Honda and Oka, 2025; He et al., 2024), and parameter-efficient tuning (Hu et al., 2022), primarily modify external inputs. Consequently, they overlook the model's intrinsic representational transferability. In particular, recent work has begun to examine the alignment of cross-domain representation (Aghajanyan et al., 2020), including neuron-level alignment in multilingual settings (Huang et al., 2025). Although foundational neuronal level analyzes are well explored (Chen et al., 2024; Sajjad et al., 2022; Li et al., 2025c,b,a), their potential to resolve logical cross-domain shifts remains largely unexplored.

5.2 Example Selection for ICL

Despite ICL's inherent sensitivity to demonstration selection (Luo et al., 2024), conventional retrieval methods (relying on semantic similarity (Rubin et al., 2022), dense retrievers (Wang et al., 2023), uncertainty signals (Ling et al., 2024; Huang et al., 2024; Margatina et al., 2023), coverage-based selection (Gupta et al., 2023), or MMR-based diversification (Liu et al., 2024)) largely operate on surface-level cues, failing to capture deeper reasoning structures. To mitigate this, recent studies leverage internal token representations (Liu et al., 2024).

6 Conlusion

We present **AdaRetrieval**, a novel framework for effective in-context cross-domain knowledge transfer. To overcome inherent domain shifts degrading traditional methods, we optimize an adapter via MMD and semantic regularization, yielding domain-invariant representations that align disparate data. This alignment enables retrieving structurally isomorphic demonstrations grounded in logical semantics over superficial literal overlap. Multi-scale evaluations confirm our approach significantly outperforms traditional baselines. To mitigate data scarcity in specialized applications, our framework unlocks complex target-domain reasoning. Future work should validate the universality of cross-domain in-context learning across borders and more challenging tasks.

Limitations

While our framework effectively mitigates domain shift in cross-domain in-context learning, several limitations remain. First, caching dense latent states for large-scale models imposes substantial VRAM overhead, constraining training on standard hardware. Second, although our alignment method synergizes with Domain-Invariant Neurons (DIN), jointly optimizing neuron-level invariants and representation-level manifold alignment within a fully unified architecture is left for future work. Finally, since current evaluations focus heavily on formal logic and mathematics, the framework's adaptability to highly unstructured generative tasks or cross-lingual transfers requires further empirical validation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. 2020. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Lihu Chen, Adam Dejl, and Francesca Toni. 2024. Analyzing key neurons in large language models. *arXiv e-prints*, pages arXiv–2406.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

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

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773.

Shivanshu Gupta, Matt Gardner, and Sameer Singh. 2023. Coverage-based example selection for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13924–13950.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, David Peng, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Shafiq Joty, Alexander R. Fabbri, Wojciech Kryscinski, Xi Victoria Lin, Caiming Xiong, and Dragomir Radev. 2022. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*.

Xuanli He, Yuxiang Wu, Oana-Maria Camburu, Pasquale Minervini, and Pontus Stenetorp. 2024. Using natural language explanations to improve robustness of in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13477–13499.

Ukyo Honda and Tatsushi Oka. 2025. Exploring explanations improves the robustness of in-context learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23693–23714.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Chongxuan Huang, Yongshi Ye, Biao Fu, Qifeng Su, and Xiaodong Shi. 2025. From neurons to semantics: Evaluating cross-linguistic alignment capabilities of large language models via neurons alignment. *arXiv preprint arXiv:2507.14900*.

Hsiu-Yuan Huang, Zichen Wu, Yutong Yang, Junzhao Zhang, and Yunfang Wu. 2024. Unlocking the power of llm uncertainty for active in-context example selection. *arXiv preprint arXiv:2408.09172*.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the association for computational linguistics: ACL 2023*, pages 1049–1065.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Yunxin Li, Xinyu Chen, Shenyuan Jiang, Haoyuan Shi, Zhenyu Liu, Xuanyu Zhang, Nanhao Deng, Zhenran

Xu, Yicheng Ma, Meishan Zhang, et al. 2025a. Unimoe-2.0-omni: Scaling language-centric omnimodal large model with advanced moe, training and data. *arXiv preprint arXiv:2511.12609*.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. 2025b. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3424–3439.

Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shenyuan Jiang, Xintong Wang, Jifang Wang, Shouzheng Huang, Xinping Zhao, Borui Jiang, Lanqing Hong, Longyue Wang, Zhuotao Tian, Baoxing Huai, Wenhan Luo, Weihua Luo, Zheng Zhang, Baotian Hu, and Min Zhang. 2025c. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *arXiv preprint arXiv:2505.04921*.

Zhiming Li, Yushi Cao, Yan Zheng, Xu Liu, Bozhi Wu, Tianlin Li, Xiufeng Xu, Junzhe Jiang, Yon Shin Teo, Shang-Wei Lin, et al. 2024. Improving neural logic machines via failure reflection. In *Forty-first International Conference on Machine Learning*.

Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, et al. 2024. Uncertainty quantification for in-context learning of large language models. *arXiv preprint arXiv:2402.10189*.

Le Liu, Zhiming Li, Jianzhi Yan, Zike Yuan, Shiwei Chen, Youcheng Pan, Buzhou Tang, Qingcai Chen, Yang Xiang, and Danny Dongning Sun. 2026. Reason analogically via cross-domain prior knowledge: An empirical study of cross-domain knowledge transfer for in-context learning. *arXiv preprint arXiv:2604.05396*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.

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

Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. 2023. Active learning principles for in-context learning with large language models. *arXiv preprint arXiv:2305.14264*.

Phuong Minh Nguyen, Tien Dang, and Naoya Inoue. 2025. Non-interactive symbolic-aided chain-of-thought for logical reasoning. In *Proceedings of the 39th Pacific Asia Conference on Language, Information and Computation*, pages 329–340.

Changdae Oh, Zhen Fang, Shawn Im, Xuefeng Du, and Yixuan Li. 2025. Understanding multimodal llms under distribution shifts: An information-theoretic approach. *arXiv preprint arXiv:2502.00577*.

Fırat Öncel, Matthias Bethge, Beyza Ermis, Mirco Ravanelli, Cem Subakan, and Çağatay Yıldız. 2024. Adaptation odyssey in llms: Why does additional pretraining sometimes fail to improve? *arXiv preprint arXiv:2410.05581*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2024. 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)*, pages 9090–9101.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*, volume 4. Now Publishers Inc.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. 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*, pages 2655–2671.

Hassan Sajjad, Nadir Durrani, and Fahim Dalvi. 2022. Neuron-level interpretation of deep nlp models: A survey. *Transactions of the Association for Computational Linguistics*, 10:1285–1303.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

Gemma Team. 2025. Gemma 3.

Karthik Valmeekam, Sarath Sreedharan, Matthew Marquez, Alberto Olmo, and Subbarao Kambhampati. 2023. On the planning abilities of large language models (a critical investigation with a proposed benchmark). *arXiv preprint arXiv:2302.06706*.

Liang Wang, Nan Yang, and Furu Wei. 2023. Learning to retrieve in-context examples for large language models. *arXiv preprint arXiv:2307.07164*.

Weixuan Wang, Barry Haddow, Minghao Wu, Wei Peng, and Alexandra Birch. 2024. Sharing matters: Analysing neurons across languages and tasks in llms. *arXiv preprint arXiv:2406.09265*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Jianzhi Yan, Zhiming Li, Le Liu, Zike Yuan, Shiwei Chen, Youcheng Pan, Buzhou Tang, Yang Xiang, and Danny Dongning Sun. 2026a. Towards effective in-context cross-domain knowledge transfer via domain-invariant-neurons-based retrieval. *arXiv preprint arXiv:2604.05383*.

Jianzhi Yan, Le Liu, Buzhou Tang, Yang Xiang, Dongning Sun, and Zhiming Li. 2026b. Coda: Towards effective cross-domain knowledge transfer via cot-guided domain adaptation. *arXiv preprint arXiv:2604.19488*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *eprint arXiv: 2407.10671*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

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

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

7 Appendix

7.1 Implementation details

Hyperparameters We summarize the comprehensive hyperparameter configurations for **AdaRetrieval** and all evaluated baselines in Table 6. For our proposed framework, the nonlinear adapter is optimized using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 1×10^{-4} and a training batch size of 32 for a maximum of 50 epochs. Based on our empirical sensitivity analysis, we set the default steering strength to $\alpha = 1.5$ and the MMD alignment penalty to $\lambda = 1.0$.

Training dataset details As shown in Tabel 4, to evaluate the cross-domain reasoning adaptation of **AdaRetrieval**, we utilize four diverse datasets spanning mathematical and logical reasoning domains. For the source domain, each instance consists of an input question x_S, a Chain-of-Thought (CoT) rationale c_S, and a final answer y_S. In contrast, the target domain datasets are treated as unlabeled, containing only input questions x_T without any CoT annotations or labels during training.

- **GSM8K**: This dataset consists of high-quality grade school math word problems that require multi-step reasoning to solve. In our experiments, it serves as the primary representative for the mathematical reasoning domain.
- **LogicalDeduction**: This dataset is used to test deductive reasoning capabilities, where the model must infer conclusions from a set of provided logical premises.
- **FOLIO**: A natural language reasoning dataset based on first-order logic. It requires the model to determine whether a conclusion is True, False, or Uncertain based on complex logical constraints.
- **ProofWriter**: This dataset focuses on generating implications and proofs over natural language. It challenges the model to perform multi-step reasoning to derive implications from a given theory.

7.2 Prompt Templates

We adopt task-specific *system prompts* for each dataset to ensure (i) a consistent reasoning style across domains and (ii) a unified, strictly formatted final answer for reliable automatic evaluation. All prompts follow a two-stage structure: the model is first instructed to provide a short step-by-step reasoning, and then output the final answer on a separate line using a standardized format (e.g., `Final answer: A`). This design minimizes ambiguity in option extraction and avoids dataset-specific parsing heuristics.

FOLIO. FOLIO requires three-way classification (*entailment* / *contradiction* / *unknown*). We extend the same reasoning template to support three options:

```
You are a careful reasoner. Think step by step
concisely.
Then on a new line, output exactly:
'Final answer: A' or 'Final answer: B' or 'Final
answer: C'.
```

ProofWriter. ProofWriter is a logical reasoning task typically formulated as entailment-style verification. We adopt the same binary template as ProntoQA:

```
You are a careful reasoner. Think step by step
concisely.
Then on a new line, output exactly: 'Final
answer: A' or 'Final answer: B'.
```

LogicalDeduction. LogicalDeduction is a multiple-choice deductive reasoning task. Let $\mathcal{O} = \{A, B, C, D\}$ denote the option set (dataset-dependent). We instruct the model to output the selected option letter in a strict format:

```
You are a careful reasoner. Think step by step
concisely.
Then on a new line, output exactly: 'Final
answer: <option letter>'.
```

Table 3: Implementation details and hyperparameter settings.

Method	Hyperparameter	Value
AdaRetrieval	Steering Layer (l)	$15 \sim 34$
	Steering Strength (α)	1.5
	MMD Penalty (λ)	1.0
	Learning Rate	$1e^{-4}$
	Train Batch Size	32
	Eval Batch Size	8
	Max Epochs	50
	Optimizer	Adam (Kingma and Ba, 2014)
	Initialization	Zero
Embedding	Embedding Model	bge-large-en-v1.5 (Xiao et al., 2024)
	Top-k	4
BM25	Term Frequency (k_1)	1.5
	Length Normalization (b)	0.75
ConE	Top-k	30
	Number of Examples (N)	4
DIN	Top-k	30
	Number of Examples (N)	4

Table 4: Task categorization and datasets used for cross-domain reasoning evaluation.

Task Type	Dataset Name	Description	Train Data Size	Test Data Size
Math	GSM8K (Cobbe et al., 2021)	Math Problems	7037	1319
Logical	LogicalDeduction (Nguyen et al., 2025)	Deductive logical reasoning	1200	300
	FOLIO (Han et al., 2022)	First-order logic reasoning	1004	204
	ProofWriter (Tafjord et al., 2021)	Generating implications	3000	600

7.3 Data Example

Setting	Model Reasoning and Output
Context	A Japanese game company created the game The Legend of Zelda.All games on the Top 10 list are made by Japanese game companies.If a game sells more than one million copies, then it will be included in the Top 10 list.The Legend of Zelda sold more than one million copies.
Zero-shot	*Since the Legend of Zelda sold more than one million copies, it should be included in the Top 10 list. However, it is stated that the Legend of Zelda is not on the Top 10 list. This creates a contradiction with the information that all games on the Top 10 list are made by Japanese game companies and the Legend of Zelda is made by a Japanese game company. Therefore, the statement "The Legend of Zelda is not on the Top 10 list" must be true. Final answer: A* **Final answer: A (True)**
Demostration	Question: In the last 25 years, 60% of Scrabble champions have been women and the rest have been men. There is one champion per year. Of the men, 40% had a beard. How many men with beards have been Scrabble champion? Reasoning: 40% of champions are men because 100 - 60 = «100-60=40»40 10 champions have been men because 25 x .4 = «25*.4=10»10 4 champions have been men with beards because 10 x .4 = «10*.4=4»4
Cross-domain	*Since the Legend of Zelda sold more than one million copies and all games on the Top 10 list are made by Japanese game companies, and the Legend of Zelda is made by a Japanese game company, it must be on the Top 10 list. Therefore, the statement "The Legend of Zelda is not on the Top 10 list" is false. Final answer: B* **Final answer: B (False)**

Table 5: Comparison between zero-shot reasoning and cross-domain reasoning reasoning on a cross-domain example. DIN-guided retrieval corrects the logical inconsistency made by zero-shot prompting.

Setting	Model Reasoning and Output
Context	People in this club who perform in school talent shows often attend and are very engaged with school events.People in this club either perform in school talent shows often or are inactive and disinterested community members.People in this club who chaperone high school dances are not students who attend the school.All people in this club who are inactive and disinterested members of their community chaperone high school dances.All young children and teenagers in this club who wish to further their academic careers and educational opportunities are students who attend the school. Bonnie is in this club and she either both attends and is very engaged with school events and is a student who attends the school or is not someone who both attends and is very engaged with school events and is not a student who attends the school.
Zero-shot	*Let's break down the information step by step:1. If Bonnie chaperones high school dances, then she is not a student who attends the school (from the third statement).2. If Bonnie does not chaperone high school dances, then she performs in school talent shows often (from the second statement).3. If Bonnie performs in school talent shows often, then she is very engaged with school events (from the first statement).4. If Bonnie is very engaged with school events, then she is a student who attends the school (from the first statement).5. From steps 1 and 4, we can conclude that Bonnie cannot be both a student who attends the school and chaperone high school dances.6. From steps 2 and 3, we can conclude that if Bonnie does not chaperone high school dances, then she is a student who attends the school.7. From steps 5 and 6, we can conclude that Bonnie is not a student who attends the school if she chaperones high school dances, and she is a student who attends the school if she does not chaperone high school dances.8. From the fourth statement, we can conclude that if Bonnie is a student who attends the school, then she is a young child or teenager who wishes to further her academic career and educational opportunities.9. From steps 7 and 8, we can conclude that Bonnie is a young child or teenager who wishes to further her academic career and educational opportunities if she does not chaperone high school dances.10. From the fifth statement, we can conclude that if Bonnie is an inactive and disinterested member of the community, then she chaperones high school dances.11. From steps 7 and 10, we can conclude that Bonnie is an inactive and disinterested member of the community if she chaperones high school dances.12. From steps 9 and 11, we can conclude that Bonnie is both a young child or teenager who wishes to further her academic career and educational opportunities and an inactive and disinterested member of the community if she does not chaperone high school dances.Therefore, the statement is true.Final answer: A* **Final answer: A (True)**
Demostration	Question: Forty percent of the students have elected to learn from home during the pandemic. The remaining students are divided into two equal groups, only one of which is physically in school on any day. What percent of students are present in school? Reasoning: 100-40=«100-40=60»60 percent of students are remaining. 60/2=«60/2=30»30 percent are present in school.
Cross-domain	*Letś break down the information step by step:1. If Bonnie chaperones high school dances, then she is not a student who attends the school (from the context).2. If Bonnie does not chaperone high school dances, then she performs in school talent shows often (from the given statement).3. If Bonnie performs in school talent shows often, then she is very engaged with school events (from the context).4. If Bonnie is very engaged with school events, then she is a student who attends the school (from the context).5. From steps 1 and 4, we have a contradiction: Bonnie cannot both be a student who attends the school and not be a student who attends the school.6. Therefore, the given statement "If Bonnie either chaperones high school dances or, if she does not, she performs in school talent shows often" must be false, because it leads to a contradiction.Final answer: B* **Final answer: B (False)**

Table 6: Comparison between zero-shot reasoning and cross-domain reasoning on a cross-domain example. DIN-guided retrieval corrects the logical inconsistency made by zero-shot prompting.

Setting	Model Reasoning and Output
Context	The bear does not chase the lion. The bear visits the lion. The bear visits the tiger. The lion chases the bear. The lion chases the tiger. The lion visits the bear. The rabbit chases the tiger. The rabbit is big. The rabbit is not green. The rabbit is young. The rabbit visits the tiger. The tiger chases the rabbit. The tiger is rough. The tiger visits the bear. If something chases the bear and it sees the rabbit then the bear does not visit the rabbit. If something chases the lion then the lion is rough. If the rabbit chases the tiger and the tiger visits the rabbit then the rabbit is not big. If something chases the lion then the lion does not visit the tiger. If something chases the rabbit and it chases the tiger then the rabbit chases the bear. If something is young and it does not see the rabbit then it sees the bear. If something visits the lion then it is young. If the tiger visits the bear and the bear is big then the tiger visits the lion. Young things are big.
Question	Based on the above information, is the following statement true, false, or unknown? The bear visits the lion.
Zero-shot	*Let's break down the information step by step: 1. The bear needs the cow. 2. The cow needs the rabbit. 3. The cow sees the bear. 4. The cow visits the lion. 5. The lion is nice. 6. The lion is red. 7. The lion does not see the cow. 8. The lion visits the cow. 9. The rabbit sees the cow. 10. The rabbit visits the cow. 11. If something needs the cow and the cow visits the rabbit then it visits the cow. 12. If the lion needs the bear then the bear sees the cow. 13. If the lion sees the bear and the bear is rough then the bear does not see the cow. 14. If something visits the lion then the lion visits the bear. 15. If something visits the bear then it needs the bear. 16. If something visits the bear then the bear needs the cow. 17. If something sees the cow then it visits the bear. From the information, we know: - The rabbit sees the cow and visits the cow. - According to point 11, if the lion needs the bear, then the bear sees the cow. - According to point 13, if the lion sees the bear and the bear is rough, then the bear does not see the cow. - Since the lion visits the cow (point 8), and the rabbit visits the cow (point 9), the rabbit needs the cow (point 15). - According to point 16, if the rabbit visits the bear, then the bear needs the cow. - Since the bear needs the cow (point 1), and the rabbit needs the cow (point 15), the bear needs the cow. - According to point 13, if the lion sees the bear and the bear is rough, then the bear does not see the cow. But we know the bear sees the cow (point 3). - Therefore, the bear is not rough (point 13). So, the rabbit is not rough.* **Final answer: A (True)**
Demostration	Context: No reptile has fur. All snakes are reptiles. Question: Based on the above information, is the following statement true, false, or uncertain? Some snake has fur. Options: A) True B) False C) Uncertain Reasoning: 1. "No reptile has fur" means: If something is a reptile, it does not have fur. 2. "All snakes are reptiles" means: If something is a snake, it is a reptile. 3. If all snakes are reptiles, and no reptile has fur, then no snake has fur. 4. The statement "Some snake has fur" would be false, because there is no snake that has fur. The correct option is: B Answer: B

Table 7: FOLIO-ProofWriter