

Towards Effective In-context Cross-domain Knowledge Transfer via Domain-invariant-neurons-based Retrieval

Anonymous ACL submission

Abstract

Large language models (LLMs) have made notable progress in logical reasoning, yet still fall short of human-level performance. Current boosting strategies rely on expert-crafted in-domain demonstrations, limiting their applicability in expertise-scarce domains, such as specialized mathematical reasoning, formal logic, or legal analysis. In this work, we demonstrate the feasibility of leveraging cross-domain demonstrating examples to boost the LLMs' reasoning performance. Despite substantial domain differences, many reusable implicit logical structures are shared across domains. In order to effectively retrieve cross-domain examples for unseen domains under investigation, in this work, we further propose an effective retrieval method, called domain-invariant neurons-based retrieval (**DIN-Retrieval**). Concisely, DIN-Retrieval first summarizes a hidden representation that is universal across different domains. Then, during the inference stage, we use the DIN vector to retrieve structurally compatible cross-domain demonstrations for the in-context learning. Experimental results in multiple settings for the transfer of mathematical and logical reasoning demonstrate that our method achieves an average improvement of 1.8% over the state-of-the-art methods ¹.

1 Introduction

In-context learning (ICL) (Brown et al., 2020; Radford et al., 2019) allows large language models (LLMs) to solve new tasks without parameter updates (Dong et al., 2024). With only a few demonstrations, LLMs can adapt rapidly and achieve strong performance across a wide range of tasks and domains (Mueller et al., 2023; Zhou et al., 2023; Wei et al., 2022; Lewkowycz et al., 2022). While recent work has examined out-of-distribution (OOD) robustness in ICL (Tang et al.,

¹Our implementation is available at <https://anonymous.4open.science/r/DIN-Retrieval-6715>

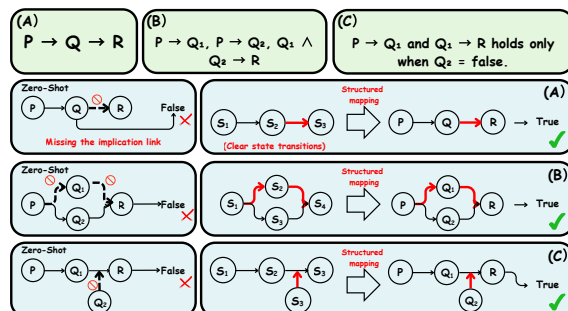


Figure 1: **Three types of failures in zero-shot LLMs.** (A) Missing intermediate links, (B) incomplete branch integration, and (C) ignored blocking conditions

2023; Sun et al., 2024; Siska et al., 2024; Yuan et al., 2024; Honda and Oka, 2025; He et al., 2024; Cheng et al., 2025; He et al., 2025), these studies typically presuppose access to in-domain, expert-annotated demonstrations. Consequently, they haven't considered the practically important setting where human expertise is scarce or unavailable, and effective reasoning must instead be supported by demonstrations drawn from other domains.

Although different domains vary in surface semantics, many reasoning tasks share underlying structural topologies (Besta et al., 2024, 2025; Bu et al., 2025; Zhang et al., 2024; Li et al., 2024). Figure 1 shows three existing types of reasoning structures that the mathematical and logical benchmarks share, yet zero-shot LLMs often fail to realize and therefore reuse them—leading to missing links, incomplete branches, or ignored blocking conditions. Notably, a cross-domain demonstration can restore the correct topology, revealing that LLMs can reuse structural reasoning patterns when appropriately guided (Tan et al., 2025). However, reasoning structures vary widely across tasks, making manual selection of structurally aligned demonstrations unrealistic. Thus, **improving cross-domain performance requires an automated retrieval mechanism capable of identifying examples with**

compatible reasoning structures.

In this work, we conduct the first trial in demonstrating the feasibility of leveraging cross-domain examples to boost ICL performance of LLMs. To achieve effective retrieval of cross-domain samples that are of similar logical structures, we propose a novel retrieval method, called domain-invariant neurons-based retrieval (**DIN-Retrieval**) (Long et al., 2015; Ganin et al., 2016; Zhao et al., 2019; Li et al., 2020; Zhu et al., 2020). Concretely, we identify DINs by selecting neurons whose activation polarities remain consistent across source and target domains based on cross-domain z-score statistics. These neurons define a stable DIN vector used as the retrieval representation, ensuring that similarity is computed within a domain-robust subspace. Conditioning on demonstrations selected through this invariant subspace enables more reliable cross-domain reasoning.

We validate their existence and importance via pruning: removing DINs causes significantly larger perplexity increases than random pruning, indicating their essential role in cross-domain reasoning. Building on this evidence, we then conduct cross-domain ICL experiments on mathematical (GSM8K), logical (PrOntoQA, FOLIO) transfer settings. Across all models and directions, DIN-based retrieval consistently outperforms explanation-based and embedding-based baselines, demonstrating that leveraging these invariant neurons substantially improves ICL robustness under domain shifts.

Our contributions are summarized as follows:

- We introduce DIN-retrieval, a universal neuron-level retrieval method that enables effective cross-domain in-context learning by identifying and exploiting domain-invariant neurons.
- Experimental results on the mutual transfer of multiple mathematical & logical reasoning benchmarks validate that DIN-retrieval consistently enhances ICL performance.
- To the best of our knowledge, this work is the first to demonstrate the feasibility of using cross-domain examples for in-context learning.

We hope this work motivates future research on robust cross-domain ICL.

2 Preliminaries

In this section, we formalize the key background concepts relevant to our study: Domain Adaptation and In-Context Learning.

2.1 Domain Adaptation

Domain adaptation aims to transfer knowledge from a labeled source domain to an unlabeled or under-resourced target domain whose data distribution differs from the source (Pan and Yang, 2010; Sun et al., 2015; Farahani et al., 2020). Let $\mathcal{D}_S = \{(x_i^S, y_i^S)\}_{i=1}^{n_S}$ denote the *source domain* and $\mathcal{D}_T = \{(x_j^T)\}_{j=1}^{n_T}$ the *target domain*, where $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Their input distributions differ:

$$P_S(x) \neq P_T(x),$$

while the underlying prediction function $f : \mathcal{X} \rightarrow \mathcal{Y}$ is assumed to be shared or related across domains.

The goal of domain adaptation is to find a mapping f_θ derived from \mathcal{D}_S that generalizes to target samples $x^T \sim P_T(x)$ without access to labeled target data. A standard approach is to learn representations $h(x) \in \mathbb{R}^d$ that reduce the divergence between source and target feature distributions:

$$\min_{\theta} \text{Dist}\left(\{h(x_i^S)\}_{i=1}^{n_S}, \{h(x_j^T)\}_{j=1}^{n_T}\right)$$

where $\text{Dist}(\cdot, \cdot)$ measures cross-domain divergence

2.2 In-Context Learning

In-context learning (ICL) allows large language models to infer new tasks from contextual examples (Wei et al., 2022; Brown et al., 2020). Unlike in-weights learning, which relies on gradient-based parameter updates, ICL adapts behavior without modifying model weights.

Formally, each training instance is linearized into an input sequence $\mathbf{x} = (x_1, \dots, x_{|\mathbf{x}|})$ and an output sequence $\mathbf{y} = (y_1, \dots, y_{|\mathbf{y}|})$, where each token belongs to the model vocabulary \mathcal{V} . Given a test input \mathbf{x}_{test} , in-context learning defines its prediction as

$$\mathbf{y}_{\text{test}} \sim \mathcal{P}_{\mathcal{M}} \left(\mathbf{y}_{\text{test}} \mid \underbrace{\mathbf{x}_1, \mathbf{y}_1, \dots, \mathbf{x}_K, \mathbf{y}_K, \mathbf{x}_{\text{test}}}_{\text{In-context prompt}} \right),$$

where the sampling operator denotes the decoding method. Each demonstration $e_i = (\mathbf{x}_i, \mathbf{y}_i)$ is

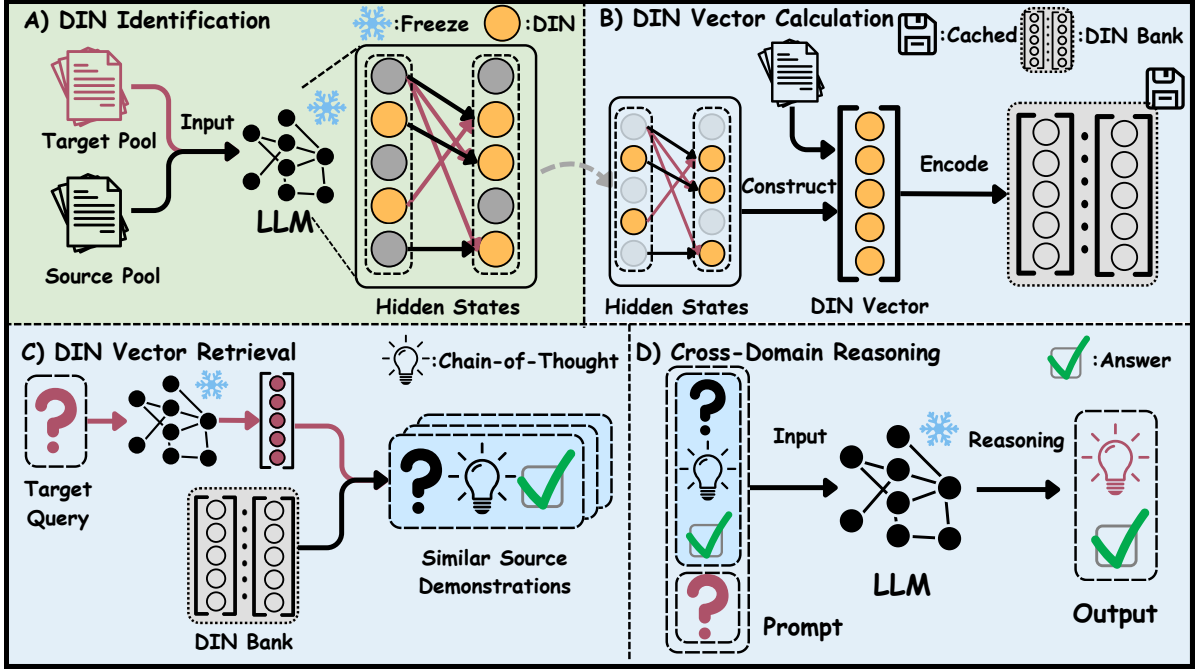


Figure 2: **Overview of the proposed DIN-based ICL framework.** The model identifies domain-invariant neurons (DINs) from source and target activations (A), constructs a stable DIN vector for representation (B), retrieves demonstrations via DIN vector similarity (C), and performs cross-domain chain-of-thought reasoning (D).

drawn from a dataset

$$\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N.$$

This formulation allows the model to condition on the provided examples without updating its parameters, enabling fast adaptation to new tasks without additional training cost.

3 Method

As aforementioned, existing ICL work ignores scenarios where human expert labelling is unavailable. In order to identify domain-invariant neurons through cross-domain alignment and use them to retrieve structurally compatible demonstrations, we present DIN-Retrieval. The following part of this section is organized as follows: we first introduce the DIN identification approach, then we illustrate the cross-domain ICL in detail. As shown in Figure 2, the proposed framework identifies domain-invariant neurons, constructs DIN representations, retrieves structurally aligned demonstrations, and performs cross-domain reasoning based on the retrieved examples.

3.1 DIN Identification

As shown in Part A of Figure 2, to retrieve demonstrations aligned with the target query, we first identify neurons that are stably activated across

the source and target domains. We use a **labeled source dataset** $\mathcal{D}_S = (x_i^S, y_i^S)$ and an **unlabeled target dataset** $\mathcal{D}_T = x_j^T$ to identify such domain-invariant neurons. We define a class of **Domain-Invariant Neurons (DIN)** at each transformer layer. Let $h_t^{(l)}(x) \in \mathbb{R}^d$ denote the hidden state of the t -th token at layer l for input x , where d is the hidden dimension and L_x is the token length. We compute the mean activation vector for a given input as:

$$\bar{h}^{(l)}(x) = \frac{1}{L_x} \sum_{t=1}^{L_x} h_t^{(l)}(x). \quad (1)$$

To measure the relative activation strength of each neuron k across domains, we compute z -scores using the joint statistics of both source and target samples:

$$\mu_k = \mathbb{E}_{x \sim (\mathcal{D}_S \cup \mathcal{D}_T)} [\bar{h}_k^{(l)}(x)], \quad (2)$$

$$\sigma_k = \sqrt{\text{Var}_{x \sim (\mathcal{D}_S \cup \mathcal{D}_T)} [\bar{h}_k^{(l)}(x)]}, \quad (3)$$

$$z_k^S = \frac{\mathbb{E}_{x \sim \mathcal{D}_S} [\bar{h}_k^{(l)}(x)] - \mu_k}{\sigma_k}, \quad (4)$$

$$z_k^T = \frac{\mathbb{E}_{x \sim \mathcal{D}_T} [\bar{h}_k^{(l)}(x)] - \mu_k}{\sigma_k}. \quad (5)$$

Here, z_k^S and z_k^T quantify the standardized activation polarity of neuron k in the source and target domains, respectively.

We define a set of domain-invariant neurons (DIN) at a given layer as the dimensions that exhibit consistent activation polarity across both the source and target domains, exceeding a specified z-score threshold τ . Formally, the DIN candidate set is:

$$\mathcal{I} = \{k \in [1, d] \mid z_k^S > \tau \wedge z_k^T > \tau\} \cup \{k \in [1, d] \mid z_k^S < -\tau \wedge z_k^T < -\tau\} \quad (6)$$

If the size of \mathcal{I} exceeds a pre-defined budget $K = \lfloor k_{\text{ratio}} \cdot d \rfloor$, we select the top- K dimensions with the largest combined z-score magnitude (those maximizing $|z_k^S| + |z_k^T|$).

3.2 Cross-domain In-Context Learning

As shown in Part B of Figure 2, for each input x , we compute its DIN representation $\mathbf{v}_{\text{DIN}}(x) = \bigoplus_{l \in \mathcal{L}} h^{(l)}(x)_{\mathcal{I}^{(l)}}$ by aggregating activations over the identified invariant neurons. Then, in Part C of Figure 2, given a target-domain query x_{test}^T , we retrieve source-domain demonstrations based on their similarity in the Domain-Invariant Neuron (DIN) space. The similarity between the target query and a source example is defined as:

$$\text{Sim}(\mathbf{v}_q, \mathbf{v}_i) = \frac{\mathbf{v}_q \cdot \mathbf{v}_i}{\|\mathbf{v}_q\| \|\mathbf{v}_i\| + \epsilon},$$

where $\mathbf{v}_q = \mathbf{v}_{\text{DIN}}(x_{\text{test}}^T)$ and $\mathbf{v}_i = \mathbf{v}_{\text{DIN}}(x_i^S)$.

To encourage diversity among selected demonstrations, we applied Maximal Marginal Relevance (MMR):

$$\text{Score}(i) = \lambda \cdot \cos(\mathbf{v}_q, \mathbf{v}_i) - (1 - \lambda) \cdot \max_{j \in \mathcal{S}} \cos(\mathbf{v}_i, \mathbf{v}_j),$$

where \mathcal{S} denotes the set of already selected examples.

Finally, in Part D of Figure 2, the top- k (with $k = 2$ in this work) retrieved source examples are concatenated with the target query to form the in-context prompt:

$$\hat{y} \sim \mathcal{P}_{\text{LM}} \left(\underbrace{[(x_1^S, y_1^S), \dots, (x_k^S, y_k^S)]}_{\text{Source Domain}}, \underbrace{x_{\text{test}}^T}_{\text{Target Domain}} \right) \quad (7)$$

By retrieving demonstrations in the domain-invariant space, the resulting prompt emphasizes structurally aligned reasoning patterns, enabling more robust cross-domain generalization.

Algorithm 1: DIN-Retrieval

Input: Source-domain pool

$\mathcal{D}_S = (x_i^S, y_i^S)$, Target-domain pool $\mathcal{D}_T = x_j^T$, Target query instance x_{test} , LLM \mathcal{M} , Activation threshold τ , Neuron selection ratio k_{ratio} , Number of retrieved demonstrations k

Output: Model Prediction \hat{y}

DIN Identification. Compute neuron-wise statistics (z_k^S, z_k^T) and select domain-invariant neurons

$$\mathcal{I} = \{k \in [1, d] \mid z_k^S > \tau \wedge z_k^T > \tau\} \cup \{k \in [1, d] \mid z_k^S < -\tau \wedge z_k^T < -\tau\}$$

Construct DIN representation

$$\mathbf{v}_{\text{DIN}}(x) = \bigoplus_l h^{(l)}(x)_{\mathcal{I}^{(l)}}.$$

Cross-Domain ICL. Retrieve top- k source examples by cosine similarity

$$\text{Sim}(x_q, x_i) = \frac{\mathbf{v}_q \cdot \mathbf{v}_i}{\|\mathbf{v}_q\| \|\mathbf{v}_i\|},$$

optionally refined by MMR, and predict

$$\hat{y} \sim \mathcal{P}_{\mathcal{M}}([(x_1^S, y_1^S), \dots, (x_k^S, y_k^S)], x_{\text{test}}).$$

4 Experiments

In this section, we evaluate the existence and usefulness of Domain-Invariant Neurons through pruning analysis and DIN-based ICL retrieval. We begin by outlining our experimental setup and then address three research questions: **RQ1** — Do DINs exist, and are they functionally important for cross-domain reasoning? **RQ2** — Is DIN-retrieval effective in improving ICL’s reasoning performance? **RQ3** — How do DIN-retrieval retrieved cross-domain demonstrations boost the reasoning performance in essence?

4.1 Experimental Setup

Backbone Models

We evaluated our approach using a diverse suite of open-source large language models, covering multiple architectures and scales. Specifically, we used LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024), Gemma-3-12B, and Gemma-3-27B (Team, 2025a), as well as Qwen2.5 and Qwen3 model families—each tested at 7B/8B, 14B, and 32B parameter sizes (Yang et al., 2024; Team, 2025b). To

Model	DIN	Random	Δ	Significant
LLaMA3.1-8B	62.7 \pm 2.3	60.3 \pm 1.8	+2.4	▼
Qwen2.5-7B	62.8 \pm 2.0	59.5 \pm 1.6	+3.3	▼
Qwen2.5-14B	70.7 \pm 1.2	68.8 \pm 1.2	+1.8	▼
Qwen3-8B	85.5 \pm 0.8	84.0 \pm 1.9	+1.5	▼

Table 1: **Comparison of cross-domain reasoning (GSM8K \rightarrow FOLIO) accuracy between DIN-ICL and random neuron selection across different models.** Δ denotes accuracy gain, and ▼ indicates statistically significant improvement ($p < 0.05$).

ensure the generality of our findings across architectures and capacities, we included both moderate-sized (7–14B) and larger (27–32B) variants. Appendix A reports implementation and decoding configurations.

Datasets & Tasks

We study **cross-domain reasoning** between **mathematical** and **logical** tasks, evaluating model generalization in both directions. We use **GSM8K** (Cobbe et al., 2021) for mathematical reasoning, and **PrOntoQA** (Saparov and He, 2022) and **FOLIO** (Han et al., 2022) for logical reasoning.

Baselines

To benchmark the effectiveness of our proposed framework, we compare against recent and representative methods:

- **Zero-Shot** (Wei et al., 2022) performs the target task without any in-context examples, relying only on its pretrained knowledge.
- **X-ICL** (He et al., 2024) enhances in-context learning by using LLM-generated natural language explanations to improve model performance.
- **Set-BSR** (Gupta et al., 2023) greedily selects examples to maximize token-level semantic coverage of the query based on bidirectional similarity.

4.2 Existence and Importance of DIN (RQ1)

To assess the existence and functional importance of domain-invariant neurons (DIN), we perform pruning analysis on LLaMA-3.1-8B-Instruct as described in Section 3.1. For each of the last six layers ($\ell = -6$ to -1), we compare the perplexity (PPL) increase from pruning DINs versus random dimensions of equal size, averaging over 300 trials. Statistical significance is evaluated using both empirical p -values and normal approximation.

Results show that pruning DINs consistently leads to greater degradation than random pruning. In the source domain, DIN pruning causes 5.2–8.1 % relative increase in PPL across layers ($\ell = -6$

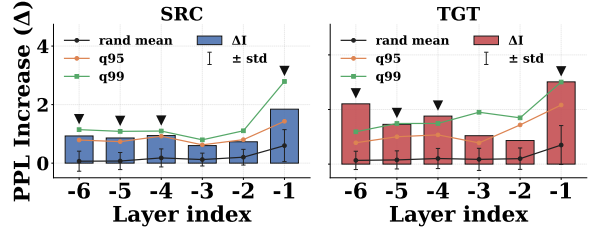


Figure 3: **Perplexity increase from pruning DINs vs. random neurons across the last six layers.** Results are averaged over 300 trials. The solid line denotes mean PPL increase after pruning DINs, while the dashed line and shaded areas indicate the random pruning baseline (mean with 95th and 99th percentiles) ▼ indicates statistically significant improvement ($p < 0.05$).

to -2), significantly exceeding the random pruning baseline (at $\ell = -6$, PPL rises by +7.99% when pruning DINs, compared with only +0.07% under random pruning ($p_{\text{emp}} = 0.0332$)). Consistently, Table 1 shows that using DIN-selected neurons for cross-domain in-context learning yields statistically significant accuracy gains over random neuron selection across multiple models.

Overall, pruning DINs from layers -6 to -1 leads to significantly greater degradation than random pruning in both domains, confirming that a compact set of domain-invariant neurons are both identifiable and functionally important for cross-domain generalization.

4.3 Cross-domain ICL improvement (RQ2)

To evaluate whether DIN can support improved generalization in cross-domain reasoning, we conduct a large-scale comparative study across diverse model families, sizes, and transfer directions. Specifically, we compare three ICL strategies: ZERO-SHOT CoT prompting without in-context examples, X-ICL strengthens ICL by using LLM-generated explanations to build more robust demonstration prompts, and our proposed DIN-guided retrieval, which operates within subspaces defined by consistent cross-domain activation patterns.

Table 2 reports accuracy results on four transfer directions: FOLIO \rightarrow GSM8K, GSM8K \rightarrow FOLIO, GSM8K \rightarrow PrOntoQA, and PrOntoQA \rightarrow GSM8K, across nine open-source models ranging from 7B to 32B parameters. In all cases, DIN-based retrieval either matches or outperforms the baselines, with especially pronounced gains in more challenging settings such as GSM \rightarrow FOL (e.g., +3.0 on Qwen2.5-14B and +4.1 on Qwen3-8B) and Pronto \rightarrow GSM (e.g., +0.6 on Qwen3-8B and +0.4 on Qwen2.5-7B). Compared to ZERO-SHOT, DIN

Method	Model Series	Parameters	Source Domain → Target Domain				
			FOL→GSM	GSM→FOL	GSM→PRO	PRO→GSM	Average
ZERO-SHOT	Qwen-2.5	7B	91.1	<i>61.8</i>	95.6	89.9	84.6
		14B	93.4	<i>67.4</i>	91.8	94.3	84.2
		32B	91.7	<i>70.3</i>	99.8	<i>91.7</i>	87.3
	Qwen-3	8B	93.3	<i>81.7</i>	100.0	92.4	<i>91.8</i>
		14B	89.9	84.5	96.2	90.5	90.2
		32B	<i>94.6</i>	82.2	100.0	<i>94.6</i>	92.3
	Gemma-3	12B	93.7	61.0	98.8	93.2	84.5
		27B	<i>94.3</i>	67.9	98.2	94.6	88.75
	LLaMA-3.1	8B	81.6	56.3	88.8	81.7	77.1
	X-ICL	Qwen-2.5	7B	89.6 _{-1.5}	59.7 _{-2.1}	95.4 _{-0.2}	89.2 _{-0.7}
14B			92.9 _{-0.5}	67.4 _{+0.0}	94.2 _{+2.4}	93.2 _{-1.1}	84.8 _{+0.6}
32B			91.6 _{-0.1}	66.0 _{-4.3}	99.6 _{-0.2}	91.6 _{-0.1}	85.7 _{-1.6}
Qwen-3		8B	93.1 _{-0.2}	81.2 _{-0.5}	99.6 _{-0.4}	92.2 _{-0.2}	91.5 _{-0.4}
		14B	89.9 _{+0.0}	83.1 _{-1.4}	94.8 _{-1.4}	91.1 _{+0.6}	89.3 _{-0.9}
		32B	94.6 _{+0.0}	83.6 _{+1.4}	100.0 _{+0.0}	94.6 _{+0.0}	92.7 _{+0.4}
Gemma-3		12B	92.6 _{-1.1}	62.5 _{+1.5}	97.8 _{-1.0}	92.1 _{-1.1}	84.3 _{-0.2}
		27B	93.1 _{-1.2}	68.4 _{+0.5}	97.6 _{-0.6}	93.1 _{-1.5}	88.1 _{-0.7}
LLaMA-3.1		8B	81.4 _{-0.2}	55.5 _{-0.8}	81.0 _{-7.8}	80.4 _{-1.3}	74.6 _{-2.5}
DIN-ICL (OURS)		Qwen-2.5	7B	89.7 _{-1.4}	63.5 _{+1.7}	96.8 _{+1.2}	90.3 _{+0.4}
	14B		93.4 _{+0.0}	70.4 _{+3.0}	94.8 _{+3.0}	94.0 _{-0.3}	86.2 _{+2.0}
	32B		92.1 _{+0.4}	71.4 _{+1.1}	99.6 _{-0.2}	92.1 _{+0.4}	87.7 _{+0.4}
	Qwen-3	8B	94.6 _{+1.3}	85.8 _{+4.1}	100.0 _{+0.0}	93.0 _{+0.6}	93.3 _{+1.5}
		14B	91.0 _{+1.1}	85.0 _{+0.5}	97.0 _{+0.8}	91.2 _{+0.7}	91.0 _{+0.8}
		32B	95.0 _{+0.4}	84.0 _{+1.8}	100.0 _{+0.0}	95.0 _{+0.4}	93.0 _{+0.7}
	Gemma-3	12B	93.7 _{+0.0}	65.5 _{+4.5}	99.0 _{+0.2}	92.7 _{-0.5}	86.1 _{+1.6}
		27B	95.1 _{+0.8}	68.9 _{+1.0}	99.2 _{+1.0}	93.9 _{-0.7}	89.3 _{+0.6}
	LLaMA-3.1	8B	81.5 _{-0.1}	63.3 _{+7.0}	88.6 _{-0.2}	81.6 _{-0.1}	78.7 _{+1.6}

Table 2: **Performance (%) across cross-domain tasks using different ICL strategies.** Each cell under DIN-ICL (OURS) includes a delta compared to the corresponding ZERO-SHOT result. The final column reports average accuracy across tasks, where **underlined bold** denotes the best and *italic* denotes the second best. Significance testing was assessed via an unequal variances t-test in comparison with ZERO-SHOT: ▼ represents a p-value lower than 0.05.

341 achieves an average gain of +0.5–2.0 points on
342 most Qwen and Gemma models, and notably out-
343 performs X-ICL across all sizes of LLaMA3.1,
344 where the latter shows degradation (e.g., -7.8 on
345 GSM→PRO) while DIN remains stable.

346 These improvements suggest that DIN identifies
347 a more transferable latent subspace for selecting
348 effective demonstrations in ICL, especially when
349 domain shift is significant. Notably, models with
350 larger capacity (e.g., Qwen3-32B) show strong
351 baseline performance where DIN yields smaller
352 gains (e.g., +0.4 or no change), whereas smaller or
353 less robust models benefit more from DIN-guided
354 selection. This pattern further supports the hypoth-

355 esis that DIN acts as a stabilizing inductive bias
356 under domain mismatch.

357 Table 3 compares our proposed DIN-ICL with
358 the non-parametric Set-BSR baseline (Gupta et al.,
359 2023) across four transfer directions (FOL→GSM,
360 GSM→FOL, GSM→Pro, and Pro→GSM) and
361 multiple model families. Overall, DIN-ICL consis-
362 tently outperforms Set-BSR across all model scales
363 and domains, confirming the effectiveness of iden-
364 tifying domain-invariant neurons for cross-domain
365 reasoning. While Set-BSR already improves over
366 standard token-level retrieval by maximizing cov-
367 erage of query semantics, it remains sensitive to
368 domain-specific embedding shifts. By contrast,

Method	Model Series	Parameters	Source Domain \rightarrow Target Domain				
			FOL \rightarrow GSM	GSM \rightarrow FOL	GSM \rightarrow PRO	PRO \rightarrow GSM	Average
SET-BSR	Qwen-2.5	7B	89.6	59.7	95.4	89.2	83.5
		14B	92.9	67.4	94.2	93.2	84.8
		32B	91.6	66.0	99.6	91.6	85.7
	Qwen-3	8B	93.1	81.2	99.6	92.2	91.5
		14B	89.9	83.1	94.8	91.1	89.3
		32B	94.6	83.6	100.0	94.6	92.7
	Gemma-3	12B	92.6	62.5	97.8	92.1	84.3
		27B	93.1	68.4	97.6	93.1	88.1
	LLaMA-3.1	8B	81.4	55.5	81.0	80.4	74.6
	DIN-ICL (OURS)	Qwen-2.5	7B	89.7 ^{+0.1}	63.5 ^{▼+3.8}	96.8 ^{▼+1.4}	90.3 ^{▼+1.1}
14B			93.4 ^{+0.5}	70.4 ^{▼+3.0}	93.9 ^{-0.3}	94.0 ^{+0.8}	86.2 ^{▼+1.1}
32B			92.1 ^{+0.5}	71.4 ^{▼+5.4}	99.6 ^{+0.0}	92.1 ^{+0.5}	87.7 ^{▼+2.0}
Qwen-3		8B	94.6 ^{▼+1.5}	85.8 ^{▼+4.6}	100.0 ^{+0.4}	93.0 ^{+0.8}	93.3 ^{▼+1.8}
		14B	91.0 ^{▼+1.1}	85.0 ^{▼+1.9}	97.0 ^{▼+2.2}	90.9 ^{-0.2}	90.8 ^{▼+1.5}
		32B	95.0 ^{+0.4}	84.0 ^{+0.4}	100.0 ^{+0.0}	95.0 ^{+0.4}	93.0 ^{+0.3}
Gemma-3		12B	93.7 ^{▼+1.1}	65.5 ^{▼+3.0}	99.0 ^{▼+1.2}	92.7 ^{+0.6}	86.1 ^{▼+1.8}
		27B	95.1 ^{▼+2.0}	68.9 ^{+0.5}	99.2 ^{▼+1.6}	92.8 ^{-0.3}	89.0 ^{▼+0.9}
LLaMA-3.1		8B	81.5 ^{+0.1}	63.3 ^{▼+7.8}	88.6 ^{▼+7.6}	81.6 ^{▼+1.2}	78.7 ^{▼+4.1}

Table 3: Comparison between retrieval-based ICL (Set-BSR) and DIN-ICL.

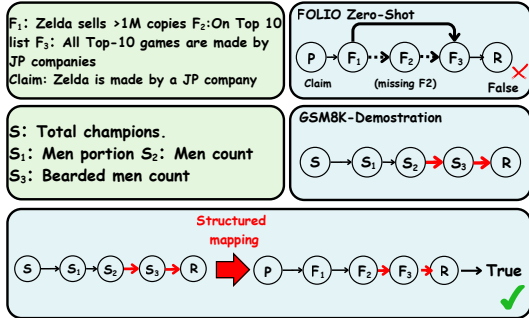


Figure 4: Case study illustrating a linear reasoning topology.

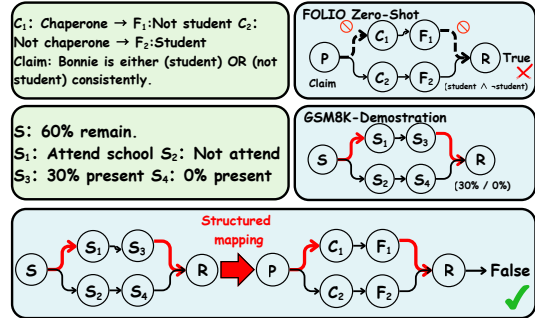


Figure 5: Case study illustrating a binary branching reasoning topology.

DIN-ICL aligns retrieval within a neuron-stable subspace, effectively reducing variance in cross-domain similarity estimation. On average, DIN-ICL achieves +0.8–1.5 pp higher accuracy than Set-BSR and up to +3 pp gains in high-shift settings such as GSM \rightarrow FOL. Notably, the performance gap diminishes as model size grows, suggesting that large LMs already encode partially domain-invariant representations, yet DIN-ICL further stabilizes retrieval, yielding the most consistent improvements across all configurations.

In summary, DIN-based retrieval provides consistent, model-agnostic improvements in cross-domain in-context reasoning and outperforms traditional full-space similarity approaches, particularly

when zero-shot performance is weak or unstable.

4.4 Structural Case Analysis of Cross-Domain Reasoning Failures (RQ3)

Figures 4, 5, and 6 present representative examples illustrating how DIN-ICL repairs broken reasoning topologies under cross-domain transfer.

In the first case, the zero-shot model fails to perform correct case analysis, as it collapses two conflicting branches into a single path, thereby missing the required multi-branch structure. The demonstration retrieved via DIN-retrieval provides an isomorphic two-branch topology, enabling structured mapping and allowing the model to recover the correct branch separation.

The second case shows a chained reasoning fail-

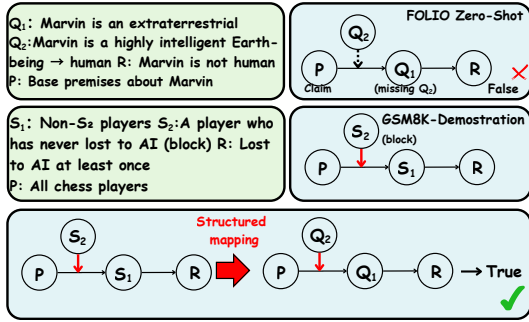


Figure 6: Case study illustrating a blocking-condition reasoning topology.

ure, where the zero-shot model omits a necessary intermediate implication. The linear step-by-step structure in the GSM8K demo supplies an appropriate scaffold, helping the model reconstruct the missing link. Finally, Figure 6 illustrates a blocking-condition topology. The zero-shot model ignores the blocking branch, leading to an invalid conclusion. DIN-ICL retrieves a GSM8K demonstration with the same blocking structure, enabling the model to reinstate the blocked path and reach the correct inference. All complete examples can be found in the table in the Appendix A.5.

Together, these cases demonstrate that DIN-ICL boosts cross-domain reasoning by supplying demonstrations whose internal structures match the required reasoning topology of the target query.

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). 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 data or prompts, while overlooking the model’s internal transferability. Recent work has begun examining cross-domain representation alignment (Aghajanyan et al., 2020), including neuron-level alignment in multilingual settings (Huang et al., 2025). Although neuron-level analyses are well explored (Chen et al., 2024; Sajjad et al., 2022), the existence and role of domain-invariant neurons under domain shift remain unknown. We address this gap by leveraging domain-invariant neurons (DINs) to improve cross-domain generalization.

5.2 Neuron-Level Analysis and Functional Attribution

Understanding the functional roles of individual neurons has been central to interpretability research (Chen et al., 2025; Sajjad et al., 2022; Antverg and Belinkov, 2021). Prior work has identified knowledge-related neurons (Dai et al., 2021) and memory behaviour in feed-forward layers (Geva et al., 2020), while lesion-based methods quantify the contribution of specific components (Voita et al., 2019; Meng et al., 2022; Li and Janson, 2024) or confidence-regulation neurons (Stolfo et al., 2024). In contrast, we identify domain-invariant neurons using cross-domain z-score polarity consistency and show via lesion tests that removing them substantially degrades performance, highlighting their importance for cross-domain reasoning.

5.3 Example Selection for In-Context Learning

In-context learning (ICL) is highly sensitive to demonstration selection (Luo et al., 2024). Existing retrieval methods rely on semantic similarity (Rubin et al., 2021), 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., 2023), but largely operate on surface-level or input-level cues. Recent work has begun using internal token representations (Liu et al., 2023). In contrast, our method retrieves demonstrations in the DIN vector space, leveraging domain-robust internal dimensions for more stable and transferable prompting.

6 Conclusion

We presented DIN-ICL, a framework that leverages Domain-Invariant Neurons (DINs) to improve cross-domain in-context learning. By identifying neurons with consistent activation polarity across domains and using them to form a DIN-based retrieval subspace, our method selects demonstrations that capture transferable reasoning structure. Experiments across multiple models and math–logic transfer settings show that DIN-ICL consistently improves cross-domain accuracy over zero-shot and strong retrieval baselines while maintaining in-domain performance. These results highlight neuron-level invariance as a useful inductive bias for robust cross-domain reasoning.

483
484
485
486
487
488
489
490
491
492
493
494

495
496
497
498
499

500
501
502

503
504
505
506
507
508
509

510
511
512
513
514
515
516
517
518

519
520
521
522
523
524
525
526
527
528
529
530
531
532

533
534
535

Limitations

First, our DIN identification method uses a simple polarity-consistency rule and fixed thresholds, which may not capture more complex invariance. Second, experiments are limited to reasoning domains (GSM8K, PrOntoQA, FOLIO); broader domains should be explored. The causal role of identified neurons remains preliminary, and observed gains, though consistent, are modest. Future work may integrate DIN-guided retrieval with adaptive or fine-tuning-based methods for stronger cross-domain generalization.

References

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

Omer Antverg and Yonatan Belinkov. 2021. On the pitfalls of analyzing individual neurons in language models. *arXiv preprint arXiv:2110.07483*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. [Graph of thoughts: Solving elaborate problems with large language models](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690.

Maciej Besta, Florim Memedi, Zhenyu Zhang, Robert Gerstenberger, Guangyuan Piao, Nils Blach, Piotr Nyczyk, Marcin Copik, Grzegorz Kwasniewski, Jürgen Müller, Lukas Gianinazzi, Ales Kubicek, Hubert Niewiadomski, Aidan O’Mahony, Onur Mutlu, and Torsten Hoefler. 2025. [Demystifying chains, trees, and graphs of thoughts](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12):10967–10989.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tianpeng Bu, Minying Zhang, Hongtao Duan, Shurui Li, Lulu Hu, and Yu Li. 2025. [Enhanced data synthesis for LLM through reasoning structures generated](#)

[by hierarchical GFlowNet](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15931–15958, Vienna, Austria. Association for Computational Linguistics. 536
537
538
539

Lihu Chen, Adam Dejl, and Francesca Toni. 2024. Analyzing key neurons in large language models. *arXiv e-prints*, pages arXiv–2406. 540
541
542

Lihu Chen, Adam Dejl, and Francesca Toni. 2025. Identifying query-relevant neurons in large language models for long-form texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23595–23604. 543
544
545
546
547

Zhoujun Cheng, Shibo Hao, Tianyang Liu, Fan Zhou, Yutao Xie, Feng Yao, Yuexin Bian, Yonghao Zhuang, Nilabjo Dey, Yuheng Zha, Yi Gu, Kun Zhou, Yuqi Wang, Yuan Li, Richard Fan, Jianshu She, Chengqian Gao, Abulhair Saparov, Haonan Li, Taylor W. Killian, Mikhail Yurochkin, Zhengzhong Liu, Eric P. Xing, and Zhiting Hu. 2025. [Revisiting reinforcement learning for llm reasoning from a cross-domain perspective](#). *Preprint*, arXiv:2506.14965. 548
549
550
551
552
553
554
555
556

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*. 557
558
559
560
561
562

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2021. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*. 563
564
565
566

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. [A survey on in-context learning](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics. 567
568
569
570
571
572
573
574

Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R. Arabnia. 2020. [A brief review of domain adaptation](#). *Preprint*, arXiv:2010.03978. 575
576
577

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35. 578
579
580
581
582

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2020. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*. 583
584
585

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and Abhinav Pandey. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783. 586
587
588

589	Shivanshu Gupta, Matt Gardner, and Sameer Singh. 2023. Coverage-based example selection for in-context learning . In <i>Findings of the Association for Computational Linguistics: EMNLP 2023</i> , pages 13924–13950, Singapore. Association for Computational Linguistics.	Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. <i>Advances in neural information processing systems</i> , 35:3843–3857.	646 647 648 649 650
595	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. Fabri, Wojciech Kryscinski, Xi Victoria Lin, Caiming Xiong, and Dragomir Radev. 2022. Folio: Natural language reasoning with first-order logic . <i>arXiv preprint arXiv:2209.00840</i> .	Maximilian Li and Lucas Janson. 2024. Optimal ablation for interpretability. <i>Advances in Neural Information Processing Systems</i> , 37:109233–109282.	651 652 653
605	Feng He, Zijun Chen, Xinnian Liang, Tingting Ma, Yunqi Qiu, Shuangzhi Wu, and Junchi Yan. 2025. Protoreasoning: Prototypes as the foundation for generalizable reasoning in llms . <i>Preprint</i> , arXiv:2506.15211.	Shuang Li, Chi Liu, Qiuxia Lin, Binhui Xie, Zhengming Ding, Gao Huang, and Jian Tang. 2020. Domain conditioned adaptation network. In <i>Proceedings of the AAAI conference on artificial intelligence</i> , volume 34, pages 11386–11393.	654 655 656 657 658
610	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 <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 13477–13499, Bangkok, Thailand. Association for Computational Linguistics.	Yuhan Li, Zhixun Li, Peisong Wang, Jia Li, Xiangguo Sun, Hong Cheng, and Jeffrey Xu Yu. 2024. A survey of graph meets large language model: Progress and future directions . <i>Preprint</i> , arXiv:2311.12399.	659 660 661 662
618	Ukyo Honda and Tatsushi Oka. 2025. Exploring explanations improves the robustness of in-context learning . In <i>Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 23693–23714, Vienna, Austria. Association for Computational Linguistics.	Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyong Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, et al. 2024. Uncertainty quantification for in-context learning of large language models. <i>arXiv preprint arXiv:2402.10189</i> .	663 664 665 666 667
624	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. <i>ICLR</i> , 1(2):3.	Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. <i>arXiv preprint arXiv:2307.03172</i> .	668 669 670 671 672
628	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. <i>arXiv preprint arXiv:2507.14900</i> .	Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In <i>International conference on machine learning</i> , pages 97–105. PMLR.	673 674 675 676
633	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. <i>arXiv preprint arXiv:2408.09172</i> .	Man Luo, Xin Xu, Yue Liu, Panupong Pasupat, and Mehran Kazemi. 2024. In-context learning with retrieved demonstrations for language models: A survey. <i>arXiv preprint arXiv:2401.11624</i> .	677 678 679 680
637	Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In <i>Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles</i> .	Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. 2023. Active learning principles for in-context learning with large language models. <i>arXiv preprint arXiv:2305.14264</i> .	681 682 683 684
644	Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh,	Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. <i>Advances in neural information processing systems</i> , 35:17359–17372.	685 686 687 688
645		Aaron Mueller, Albert Webson, Jackson Petty, and Tal Linzen. 2023. In-context learning generalizes, but not always robustly: The case of syntax. <i>arXiv preprint arXiv:2311.07811</i> .	689 690 691 692
		Changdae Oh, Zhen Fang, Shawn Im, Xuefeng Du, and Yixuan Li. 2025. Understanding multimodal llms under distribution shifts: An information-theoretic approach. <i>arXiv preprint arXiv:2502.00577</i> .	693 694 695 696
		Fırat Öncel, Matthias Bethge, Beyza Ermis, Mirco Ravanelli, Cem Subakan, and Çağatay Yıldız. 2024. Adaptation odyssey in llms: Why does additional pre-training sometimes fail to improve? <i>arXiv preprint arXiv:2410.05581</i> .	697 698 699 700 701

702	Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning . <i>IEEE Transactions on Knowledge and Data Engineering</i> , 22(10):1345–1359.	Elena Voita, Rico Sennrich, and Ivan Titov. 2019. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. <i>arXiv preprint arXiv:1909.01380</i> .	753 754 755 756 757
705	Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. <i>OpenAI blog</i> , 1(8):9.	Jinchao Wang, Qian Zhang, Jun Ma, and Xinyu Sun. 2024. Analysing neurons across languages and tasks in large language models. In <i>Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing</i> .	758 759 760 761 762
709	Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2021. Learning to retrieve prompts for in-context learning. <i>arXiv preprint arXiv:2112.08633</i> .	Liang Wang, Nan Yang, and Furu Wei. 2023. Learning to retrieve in-context examples for large language models. <i>arXiv preprint arXiv:2307.07164</i> .	763 764 765
712	Hassan Sajjad, Nadir Durrani, and Fahim Dalvi. 2022. Neuron-level interpretation of deep nlp models: A survey. <i>Transactions of the Association for Computational Linguistics</i> , 10:1285–1303.	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. <i>Advances in neural information processing systems</i> , 35:24824–24837.	766 767 768 769 770
716	Abulhair Saparov and He He. 2022. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. <i>arXiv preprint arXiv:2210.01240</i> .	Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing . In <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations</i> , pages 38–45, Online. Association for Computational Linguistics.	771 772 773 774 775 776 777 778 779 780 781 782
719	Charlotte Siska, Katerina Marazopoulou, Melissa Ailem, and James Bono. 2024. Examining the robustness of LLM evaluation to the distributional assumptions of benchmarks . In <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 10406–10421, Bangkok, Thailand. Association for Computational Linguistics.	An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. <i>arXiv preprint arXiv:2407.10671</i> .	783 784 785 786 787 788 789 790 791 792 793 794 795 796 797 798
727	Alessandro Stolfo, Ben Wu, Wes Gurnee, Yonatan Belinkov, Xingyi Song, Mrinmaya Sachan, and Neel Nanda. 2024. Confidence regulation neurons in language models. <i>Advances in Neural Information Processing Systems</i> , 37:125019–125049.	Yu Yuan, Lili Zhao, Kai Zhang, Guangting Zheng, and Qi Liu. 2024. Do LLMs overcome shortcut learning? an evaluation of shortcut challenges in large language models . In <i>Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing</i> , pages 12188–12200, Miami, Florida, USA. Association for Computational Linguistics.	799 800 801 802 803 804 805
732	Shiliang Sun, Honglei Shi, and Yuanbin Wu. 2015. A survey of multi-source domain adaptation . <i>Information Fusion</i> , 24:84–92.	Ge Zhang, Mohammad Ali Alomrani, Hongjian Gu, Jiaming Zhou, Yaochen Hu, Bin Wang, Qun Liu, Mark Coates, Yingxue Zhang, and Jianye Hao. 2024. Path-of-thoughts: Extracting and following paths for robust relational reasoning with large language models . <i>Preprint</i> , arXiv:2412.17963.	806 807 808 809 810 811
735	Zechen Sun, Yisheng Xiao, Juntao Li, Yixin Ji, Wenliang Chen, and Min Zhang. 2024. Exploring and mitigating shortcut learning for generative large language models. In <i>Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)</i> , pages 6883–6893.		
742	Xue Wen Tan, Nathaniel Tan, Galen Lee, and Stanley Kok. 2025. The shape of reasoning: Topological analysis of reasoning traces in large language models . <i>Preprint</i> , arXiv:2510.20665.		
746	Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. 2023. Large language models can be lazy learners: Analyze shortcuts in in-context learning. <i>arXiv preprint arXiv:2305.17256</i> .		
750	Gemma Team. 2025a. Gemma 3 .		
751	Qwen Team. 2025b. Qwen3 technical report . <i>Preprint</i> , arXiv:2505.09388.		

- 812 Han Zhao, Remi Tachet Des Combes, Kun Zhang, and
813 Geoffrey Gordon. 2019. On learning invariant rep-
814 resentations for domain adaptation. In *International*
815 *conference on machine learning*, pages 7523–7532.
816 PMLR.
- 817 Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and
818 Sameer Singh. 2021. Calibrate before use: Improv-
819 ing few-shot performance of language models. In
820 *International conference on machine learning*, pages
821 12697–12706. PMLR.
- 822 Yuhang Zhou, Paiheng Xu, Xiaoyu Liu, Bang An, Wei
823 Ai, and Furong Huang. 2023. Explore spurious cor-
824 relations at the concept level in language models for
825 text classification. *arXiv preprint arXiv:2311.08648*.
- 826 Yongchun Zhu, Fuzhen Zhuang, Jindong Wang, Guolin
827 Ke, Jingwu Chen, Jiang Bian, Hui Xiong, and Qing
828 He. 2020. Deep subdomain adaptation network for
829 image classification. *IEEE transactions on neural*
830 *networks and learning systems*, 32:1713–1722.

A Appendix

A.1 Implementation Details

A.1.1 Model Inference

All experiments are conducted using vLLM (Kwon et al., 2023) as the inference backend to ensure efficient serving of large models and fast hidden-state extraction. Unless otherwise specified, model precision is set to FP16, following the default mixed-precision configuration of vLLM. We use HuggingFace Transformers (Wolf et al., 2020) for model loading, tokenization, and hidden-state access.

A.1.2 Generation Hyperparameters

Across all experiments—including cross-domain ICL evaluation, DIN retrieval, and case studies—we use the following decoding configuration:

Category	Setting
Temperature	0.6
Sampling (Top-p/k)	Top-p = 1.0, Top-k = 50
Max Gen Length	8192 tokens
Random Seed	1-30

Table 4: Decoding setup used throughout all experiments.

A.2 Hyperparameter Analysis

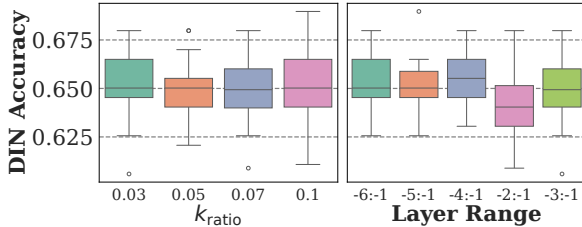


Figure 7: **Effect of key hyperparameters on DIN-based ICL performance.** Left: Increasing k_{ratio} generally leads to slightly higher DIN accuracy. Right: DIN subspaces extracted from deeper layers tend to outperform shallower ones.

We investigate how key hyperparameters affect the effectiveness of DIN selection and its downstream impact on in-context reasoning. Specifically, we analyze the influence of the selection ratio k_{ratio} and the choice of layer range used to extract domain-invariant neurons.

Figure 7 (left) shows that increasing k_{ratio} from 0.03 to 0.1 leads to slightly higher DIN accuracy on average, although the variance remains large. This suggests that using more neurons provides richer signal for cross-domain generalization,

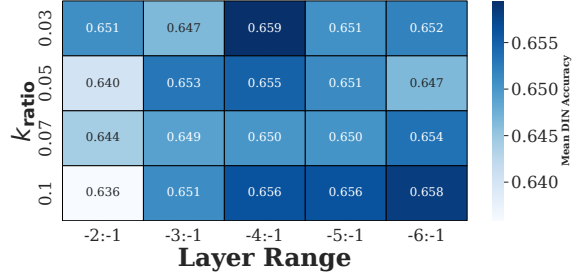


Figure 8: **Mean DIN accuracy across different combinations of k_{ratio} and layer range.**

but over-selection may introduce noise. Figure 7 (right) compares different layer ranges, showing that deeper layers consistently yield higher accuracy than shallower ones. This is consistent with prior findings that later layers in LLMs encode more task-specific and transferable representations.

To jointly analyze the interaction between the two hyperparameters, we plot a heatmap in Figure 8. The results confirm that deeper layer ranges and moderate k_{ratio} values yield the most reliable DIN subspaces across tasks. Notably, the highest DIN accuracy (0.659) is achieved with $k_{ratio} = 0.03$ and layer range L-4:-1, indicating that quality can sometimes outweigh quantity when selecting stable neurons.

These results highlight the importance of careful hyperparameter tuning when applying DIN-based retrieval in practice. We adopt the best-performing settings in subsequent experiments unless otherwise noted.

A.3 Prompts

We adopt task-specific system prompts for each dataset to ensure consistent reasoning style and unified answer formatting across domains. All prompts follow a two-stage structure: (1) the model is instructed to provide a short step-by-step reasoning; and (2) the final answer must be output on a separate line using a strict standardized format. This design avoids ambiguity in option extraction and enables reliable automatic evaluation.

PrOntoQA. For binary logical reasoning tasks in PrOntoQA, the system prompt is:

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

FOLIO. FOLIO requires three-way classification (entailment / contradiction / unknown). We

894 extend the same reasoning template to support three
895 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'.
```

896
897 **GSM8K.** For math word problems, we follow the
898 conventional Chain-of-Thought (CoT) prompting
899 format:

```
You are a careful math reasoner. Solve step by  
step concisely.  
Then on a new line, output exactly: 'Final  
answer: <number>'.
```

900 901 **A.4 Dataset Statistics**

902 To evaluate cross-domain reasoning, we use three
903 representative datasets covering mathematical and
904 logical tasks: GSM8K, FOLIO, and PrOntoQA.
905 These datasets differ in reasoning style, language
906 form, and label space, offering a diverse benchmark
907 for assessing the transferability of domain-invariant
908 neurons. Table 5 summarizes their key statistics.

Dataset	#Train	#Test	Labels
GSM8K	7,473	1,319	Numeric
FOLIO	5,216	500	A/B/C
PrOntoQA	22,000	2,000	A/B (Yes/No)

Table 5: Statistics of datasets used in our cross-domain experiments.

909 **A.5 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	<i>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</i> 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 $100 - 60 = 40$ 40 champions have been men because $25 \times .4 = 10$ 10 champions have been men with beards because $10 \times .4 = 4$
DIN-ICL	<i>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</i> Final answer: B (False)

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

Setting	Model Reasoning and Output
Context	<p>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.</p>
Zero-shot	<p><i>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</i> Final answer: A (True)</p>
Demonstration	<p>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: $100 - 40 = 60$ 60 percent of students are remaining. $60 / 2 = 30$ 30 percent are present in school.</p>
DIN-ICL	<p><i>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 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</i> Final answer: B (False)</p>

Table 7: Comparison between zero-shot reasoning and DIN-ICL reasoning on a cross-domain example. DIN-guided retrieval corrects the logical inconsistency made by zero-shot prompting.