Task-Related In-Context Learning

Wenqiang Wang[1], **Peng Chen**[1], **Yan Xiao**[1], **Yangshijie Zhang**[3],
Xiaoyue Lu[1], Jianjie huang[1], Xiaochun Cao[1,2*]
[1]Shenzhen Campus of Sun Yat-sen University, [2]Peng Cheng Laboratory
[3]Lanzhou University
wangwq69@mail2.sysu.edu.cn, 23pchen@stu.edu.cn
huangjj67@mail2.sysu.edu.cn, caoxiaochun@mail.sysu.edu.cn

Abstract

Standard in-context learning (ICL) assumes identical output spaces between test and retrieval datasets (fully aligned). However, in practice, these datasets can be fully aligned, but also can be fully disjoint in label space (Output space), forming an information continuum from rich to scarce. Naive ICL often becomes ineffective under such mismatches. In this work, we challenge this assumption by demonstrating that the retrieval dataset need not perfectly align with the test dataset, as long as it remains related to the target task. We propose Task-Related In-Context Learning (TRICL), a unified framework for ICL under output-space mismatch, designed to cover the full continuum of scenarios. TRICL first identifies demonstrations in the mismatched retrieval dataset that are relevant to the test label space via a lightweight Bayesian probabilistic criterion, and uses them to form a related dataset. TRICL then perform ICL on the related dataset to obtain preliminary predictions; finally, TRICL leverage these intermediate predictions to reduce and transform the output space of the original test task, thereby improving the performance of LLMs. In the *fully disjoint* scenario, as long as the retrieval dataset is task-related to the test task, TRICL achieves state-of-the-art (SOTA) results across three LLMs, three task types, and four datasets. Moreover, TRICL remains effective in the *fully aligned* scenaris, consistently yielding strong gains over competitive baselines. Moreover, TRICL also extends to generative task.

1 Introduction

In-context learning (ICL) enables large language models (LLMs) to solve NLP tasks using a small set of retrieved demonstrations (Liu et al., 2023; Ho et al., 2024). Standard ICL implicitly assumes that the test dataset, task, and the retrieval dataset share an identical output space, so retrieved demonstrations can be applied directly to inference. While this assumption is frequently violated in the real world, where test and retrieval datasets are not fully aligned. Moreover, the label space of the task is identical to that of the test dataset.

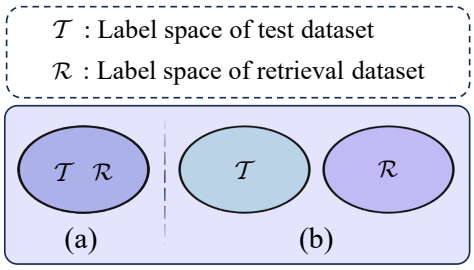


Figure 1: The different scenarios of the label space of test dataset and retrieval dataset

As illustrated in Figure 1, the output spaces of the test dataset $\mathcal{T}$ and the retrieval dataset $\mathcal{R}$ can relate in several ways. Figure 1 (a) corresponds to the standard ICL setting, where $\mathcal{T}$ and $\mathcal{R}$ are fully aligned. Figure 1 (b) represents the fully disjoint case, where $\mathcal{R}$ shares no label-level overlap with $\mathcal{T}$; for example, in weakly supervised settings, supervision is provided only through weak signals (e.g., metadata or heuristic annotations), which do not define matching labels but may still be informative for inference.

At the fully disjoint settings, the retrieval dataset provides no direct label-level evidence, making a naive application of standard ICL particularly fragile, as confirmed by our empirical study (Section 2.2). This setting can be viewed as a natural relaxation of the label-space alignment assumption underlying standard ICL. In the absence of label overlap, retrieval labels themselves are no longer directly informative for the target task. Instead, any useful signal must come from shared underlying attributes or latent factors that correlate with the target decision. As illustrated in Figure 2, the

*Corresponding author: Xiaochun Cao

Findings of the Association for Computational Linguistics: ACL 2026, pages 43378–43400

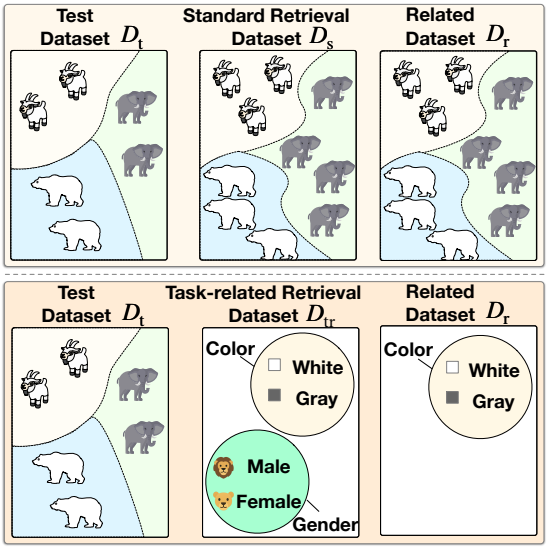


Figure 2: Illustrative example of the task-related retrieval dataset and related dataset.

test task predicts animal categories, whereas the retrieval datasets are labeled by related attributes such as color or gender. Although these attribute labels are not aligned with the target labels, attributes (e.g., color) can still constrain the prediction space. We refer to such residual but informative cues as *task-related signals*.

To formalize the above observation, we introduce the concept of *Task-Related Retrieval Datasets (TRRD)*. A TRRD refers to a retrieval dataset that remains semantically or functionally related to the target task. Prior work has considered partially disjoint output-space scenarios. For example, incomplete in-context learning (Wang and Zhang, 2025) assumes that the retrieval label space is a subset of the test label space, while indirect in-context learning (Askari et al., 2025) assumes the reverse. However, the fully disjoint scenario has remained largely unexplored.

Therefore, we further define *Task-Related In-Context Learning (TRICL)* as the application of in-context learning using such task-related retrieval datasets. Specifically, TRICL first identifies examples in the mismatched retrieval dataset that are relevant to the test label space via a lightweight Bayesian probabilistic criterion, and uses them to form a related dataset. We then perform ICL on the related dataset to obtain preliminary predictions; finally, we use these intermediate predictions to reduce and transform the output space of the target task before final inference. In this work, we use the term *related task* in a broad sense to refer to a task whose predictions can provide informative intermediate evidence or constraints on the output space of the target task, even when the two tasks do not share the same label space. By separating these two stages, TRICL enables effective inference under the fully disjoint ICL scenario.

Our results show that effective ICL does not require strict label alignment between retrieval and test datasets; rather, what matters is whether retrieval data provide task-related signals that can constrain the target label space. This perspective reveals a previously underexplored failure mode of standard ICL under output-space mismatch and motivates TRICL as a general framework that decouples task-related signal discovery from final prediction in fully disjoint setting. TRICL performs strongly across three LLMs, three task types, and four datasets, reaching up to 71.0% accuracy and an MSE of 0.79. We also explore variants based on the TRICL framework, demonstrating that it can be extended to generative tasks with considerable performance. The main contributions of this work are summarized as follows.

- We formalize the *Task-Related Retrieval Dataset* (TRRD) setting, which characterizes in-context learning scenarios where the retrieval and test datasets differ in task type or output space. Moreover, through systematic empirical analysis, we show that directly applying standard ICL under this setting is often ineffective.
- We propose *Task-Related In-Context Learning* (TRICL), a simple and general framework that enables effective in-context learning under the fully disjoint setting by explicitly separating task-related signal discovery from final prediction and using intermediate predictions to reduce the target output space.
- Our experiments demonstrate that effective in-context learning does not require strict label alignment between the retrieval and test datasets; instead, task-related signals alone can suffice to improve inference, substantially extending the applicability of ICL beyond standard aligned settings.
- We show that TRICL is robust across a broad range of retrieval–test relationships in the fully disjoint setting, and can be naturally adapted to generative tasks, highlighting its flexibility as a plug-and-play ICL framework.

2 Bayesian Theory and Frequency Approximation of Probability

We adopt a simple Bayesian criterion to characterize task-related signals. For two s A and B, the condition $P(A \mid B) > P(A)$ indicates that observing B provides positive evidence for A(Bernardo and Smith, 2009). In TRICL, this criterion is used to identify labels in the retrieval dataset that are informative for the target task. In practice, the probabilities are estimated using empirical frequencies computed from a small random sample of the retrieval dataset, following standard frequency-based approximation (Devroye et al., 2013).

2.1 Problem Formulation

Task-related Retrieval Dataset and Related Dataset. In real-world scenarios, the retrieval dataset may differ from the test dataset in task type and output space, yet still provide useful evidence for inference. We call such a mismatched-but-informative retrieval dataset a *Task-Related Retrieval Dataset (TRRD)*. We further define a *Related Dataset* $\mathbf{D}_{\mathrm{r}}$ as the subset of the TRRD that contains only those demonstrations whose retrieval labels provide such evidence for the target task. Figure 2 illustrates this setting: the test task predicts animal categories, while the TRRD is labeled by attributes such as color or gender. Although these attribute labels are not aligned with the target labels, they can constrain the prediction space and facilitate inference; for example, a gray image strongly suggests *elephant*. Therefore, we collect color-labeled examples into the Related Dataset $\mathbf{D}_{\mathrm{r}}$, and summarize the notation in Table 8.

Definition 1 (Task-related Retrieval Dataset). *Let $\mathbf{D}_t$ denote a test dataset with label space $\{A_1, \ldots, A_{m_a}\}$, and let $\mathbf{D}_{tr}$ denote a retrieval dataset with label space $\{B_1, \ldots, B_{m_b}\}$. If there exists at least one test label A_i such that*

$$\exists\, B_j \text{ such that } P(A_i \mid B_j) > P(A_i), \quad (1)$$

then $\mathbf{D}_{tr}$ is a task-related retrieval dataset *(TRRD).*

Definition 2 (Related Dataset). *Related dataset $\mathbf{D}_r$ is defined as a subset of the $\mathbf{D}_{tr}$. It consists of those demonstrations from $\mathbf{D}_{tr}$ that can facilitate self-inference for the test dataset $\mathbf{D}_t$. Formally, for a demonstration (x_t^d, y_t^d) in $\mathbf{D}_{tr}$, if $y_t^d = B_j$ and the following condition holds $P(A_i \mid B_j) > P(A_i)$, then the pair (x_t^d, y_t^d) is included in $\mathbf{D}_r$. Formally:*

$$\text{If } y_t^d = B_j, \text{and } P(A_i \mid B_j) > P(A_i), \text{then } (x_t^d, y_t^d) \in \mathbf{D}_r. \quad (2)$$

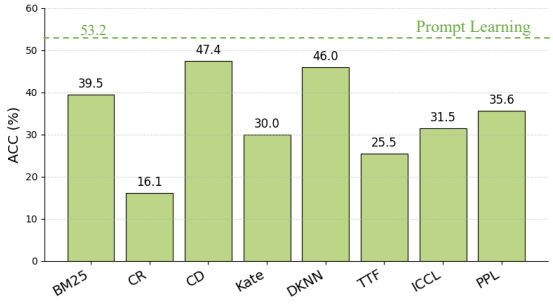


Figure 3: ICL performance on task-related retrieval datasets (TRRD) with 5 demonstrations. BM25, CR, CD, KATE, TTF, ICCL, and PPL are ICL methods.

2.2 Empirical Study

When the test and retrieval label spaces are disjoint, naive ICL often performs poorly. Following the TRRD setting in Table 1, we confirm this on Emotion with Qwen4-9B using k=5 demonstrations. As Figure 3 shows, ICL methods even underperform prompting learning (best ICL: 47.4% vs. prompting learning: 53.2%). These results suggest that simply using mismatched demonstrations is insufficient for TRRD, and effective inference in this setting requires exploiting task-related signals to constrain test label space (Li et al., 2024a; Ying et al., 2024).

3 Task-related In-context Learning

The mismatch between the test and retrieval datasets hinders the retrieval of useful demonstrations, leading to poor performance for naive ICL. We therefore propose **Task-Related In-Context Learning (TRICL)**. As shown in Figure 4, TRICL first identifies a *related dataset* from the TRRD and induces a corresponding *related task*. It then retrieves demonstrations and performs preliminary inference on the related task to obtain an intermediate prediction, which is finally used to reduce and transform the test output space for prediction (Liu et al., 2024a).

3.1 Related Dataset and Related Task

As Figure 4 shows, TRICL identifies demonstrations in TRRD whose labels provide positive evidence for the test task, forming a *related dataset*. Consider a classification test task T_A (e.g.,species classification) with label space $\{A_1, \ldots, A_{m_a}\}$ (e.g.,Goat, Polar bear,and Elephant) and a TRRD $\mathbf{D}_{\mathrm{tr}}$ with label space $\{B_1, \ldots, B_{m_b}\}$ (e.g.,White, Gray, Male, and Female). We regard a label B_j as *task-related label* if it provides positive evidence for any test label A_i. Specifically, after observ-

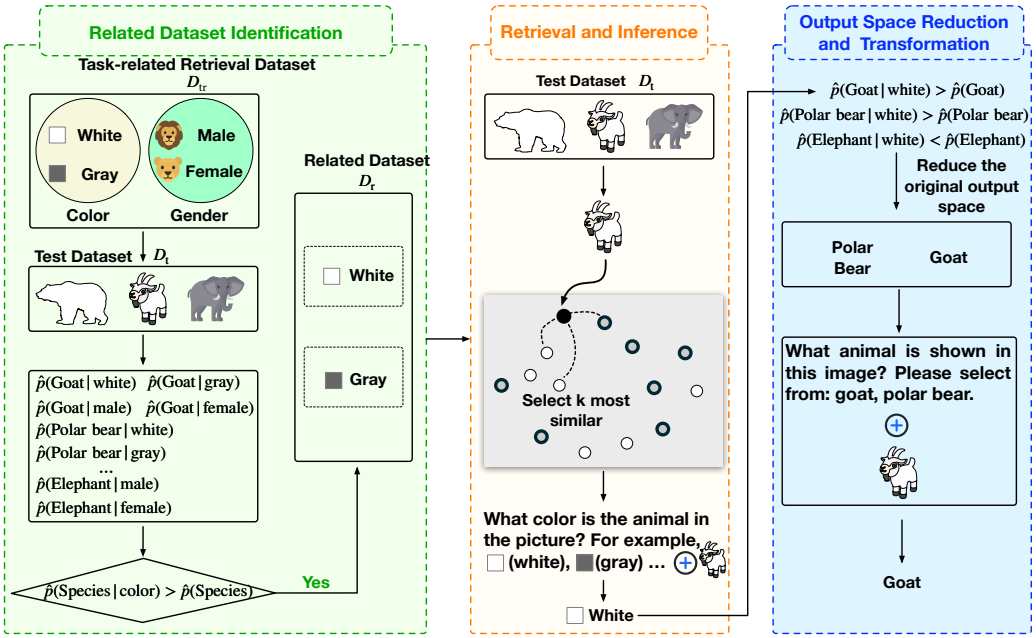


Figure 4: Overview of Task-Related In-Context Learning (TRICL). Left: TRICL identifies a related subset $\mathbf{D}_{\mathrm{r}} \subset \mathbf{D}_{\mathrm{tr}}$ by testing which retrieval labels (e.g., *color*) provide positive evidence for target labels (e.g., *species*). Middle: it retrieves top-k demonstrations from $\mathbf{D}_{\mathrm{r}}$ and performs in-context inference on the induced related task to obtain an intermediate prediction $\hat{y} \in \{C_w\}$. Right: $\hat{y}$ is treated as task-related evidence to reduce the target label space to a small candidate set, followed by final prediction via zero-shot prompting within this reduced space. If the related-task prediction provides no informative constraint, TRICL degenerates to standard zero-shot inference.

ing B_j, the predicted probability of at least one test label A_i increases relative to its prior, i.e., $P(A_i \mid B_j) > P(A_i)$. For example, observing a *gray* prediction raises the posterior probability of the image belonging to the *Elephant* label in Figure 4.

Since these probabilities are unknown, we estimate them using empirical frequencies over a small sampled subset of $\mathbf{D}_{\mathrm{tr}}$. We randomly sample up to 100 demonstrations (or all demonstrations if fewer than 100) and apply prompting learning to predict labels in $\{A_1, \ldots, A_{m_a}\}$. The probabilities of $P(A_i \mid B_j)$ and $P(A_i)$ are estimated as:

$$\hat{P}(A_i) = \frac{N(A_i)}{N_{\text{total}}}, \quad \hat{P}(A_i \mid B_j) = \frac{N(A_i, B_j)}{N(B_j)}, \quad (3)$$

where N_{total} is the number of sampled demonstrations, $N(A_i)$ counts how many of them are predicted as label A_i, and $N(A_i, B_j)$ counts how many demonstrations both have retrieval label B_j and are predicted as A_i. If $\hat{p}(A_i \mid B_j) > \hat{p}(A_i)$ (e.g., $\hat{p}$("Goat" | "White") $> \hat{p}$("Goat")), then all demonstrations in the TRRD $\mathbf{D}_{\mathrm{tr}}$ with label B_j are considered *task-related* to the test task, and thus comprise the related dataset $\mathcal{D}_r$. We denote the $(x_i^{\mathrm{d}}, y_i^{\mathrm{d}})$ as the demonstration in $\mathbf{D}_{\mathrm{tr}}$, formally:

$$\forall\, y_i^{\mathrm{d}} = B_j, \text{ if } \hat{p}(A_i \mid B_j) > \hat{p}(A_i), \text{ then } (x_i^{\mathrm{d}}, y_i^{\mathrm{d}}) \in \mathbf{D}_{\mathrm{r}}. \quad (4)$$

After evaluating all labels $\{B_1, B_2, \ldots, B_{m_b}\}$ in the $\mathbf{D}_{\mathrm{tr}}$, we obtain the related dataset $\mathbf{D}_{\mathrm{r}}$ (e.g., images with labels White and Gray) associated with the task-related labels $\{C_1, \ldots, C_{m_c}\}$ (e.g., White and Gray), where each $C_w \in \{B_1, \ldots, B_{m_b}\}$ and satisfies the condition $\hat{p}(A_i \mid C_w) > \hat{p}(A_i)$. Therefore, the related dataset $\mathbf{D}_{\mathrm{r}}$ can be denoted as:

$$\mathbf{D}_{\mathrm{r}} = \{(x_1^{\mathrm{r}}, y_1^{\mathrm{r}}), (x_2^{\mathrm{r}}, y_2^{\mathrm{r}}), \ldots, (x_h^{\mathrm{r}}, y_h^{\mathrm{r}})\}, \quad (5)$$

s.t. $(x_i^{\mathrm{r}}, y_i^{\mathrm{r}}) \in \mathbf{D}_{\mathrm{tr}}$, and $y_i^{\mathrm{r}} \in \{C_1, C_2, \ldots, C_{m_c}\}$. The $(x_1^{\mathrm{r}}, y_1^{\mathrm{r}})$ is the demonstration of $\mathbf{D}_{\mathrm{r}}$. Therefore, the test task involves classifying a given input into one of the labels in $\{A_1, \ldots, A_{m_a}\}$, whereas the corresponding **related task** reformulates the problem as classification into one of the labels in $\{C_1, \ldots, C_{m_c}\}$.

3.2 Retrieval and Related-Task Inference

TRICL follows standard semantic retrieval practices widely adopted in prior ICL methods, such as VICL-rerank (Zhou et al., 2024) and KATE. Unlike standard ICL, retrieval is performed *within the related dataset* $\mathbf{D}_{\mathrm{r}}$ identified in the previous

stage. Specifically, given a test input x_i^{t}, we retrieve the top-k semantically most similar demonstrations from $\mathbf{D}_{\text{r}}$ and order them by descending similarity to construct the in-context $\mathbf{D}_{\text{de}}^i = \{(x_l^{\text{de}}, y_l^{\text{de}})\}_{l=1}^k$, where $y_l^{\text{de}} \in \{C_1, \ldots, C_{m_c}\}$ are labels from the related task. Detailed retrieval implementations follow standard practice and are provided in Appendix C.

In the inference phase, TRIC first predict the label of the test input x_i^{t} from the related label set $\{C_1, \cdots, C_{m_c}\}$ corresponding to the related task:

$$\hat{y}_i^{\text{r}} = f_{\text{LLM}}(x_i^{\text{t}}, \mathbf{D}_{\text{de}}^i) \in \{C_1, \ldots, C_{m_c}\}. \quad (6)$$

This intermediate prediction provides task-related evidence that is used in the next stage to reduce and transform the output space of the target task.

3.3 Output Space Reduction and Transformation

The prediction in Equation 6 lies in the output space $\{C_1, \ldots, C_{m_c}\}$ of the related task, while the final decision must be made in the original label space $\{A_1, \ldots, A_{m_a}\}$ of the test task. To bridge this mismatch, we adopt a two-stage output space transformation: we first use the related-task prediction $\hat{y}_i^{\text{r}}$ to constrain the original label space, and then perform zero-shot prompting within the resulting reduced space.

Figure 4 illustrates the process with a three-class animal classification example (*goat*, *polar bear*, *elephant*), where a related dataset is annotated by color. We first predict the color (*white*, *gray*) and then restrict the original label space accordingly: *goat*, *polar bear* for *white* and *elephant* for *gray*. Importantly, the eliminated labels are not removed arbitrarily, but correspond to those that are substantially less likely under the task-related evidence.

Since $C_w \in \{B_1, \ldots, B_{m_b}\}$ and $\hat{p}(A_i \mid B_j)$ is estimated in Equation 3, when $C_w = B_j$ we have $\hat{p}(A_i \mid C_w) = \hat{p}(A_i \mid B_j)$. Given the predicted related-task label C_w, we keep A_i as a candidate if $\hat{p}(A_i \mid C_w) > \hat{p}(A_i)$, i.e., task-related evidence increases its likelihood. We denote the retained labels by $A_i^{C_w}$ and define the reduced label set

$$\{A_1^{C_w}, A_2^{C_w}, \ldots, A_{m_d}^{C_w}\}, \quad \text{s.t. } A_i^{C_w} \in \{A_1, A_2, \ldots, A_{m_a}\}, \; \hat{p}(A_i^{C_w} \mid C_w) > \hat{p}(A_i^{C_w}), \quad (7)$$

where $m_d < m_a$ (proof in Section L). For a test input x_i^{t} predicted as C_w in Equation 6, the LLM is prompted to select the final label from the reduced set: "`Classify the text input into one of the labels:` $A_1^{C_w}, A_2^{C_w}, \ldots, A_{m_d}^{C_w}$." We denote f_{prompt} as the prompt learning predictor, and the final prediction y_i^{t} of x_i^{t} is

$$\hat{y}_i^{\text{r}} = f_{\text{prompt}}(x_i^{\text{t}}) \in \{A_1^{C_w}, A_2^{C_w}, \ldots, A_{m_d}^{C_w}\}. \quad (8)$$

In the extreme case where C_w provides no informative constraint, TRICL recover to standard prompt learning with no additional error.

Other downstream tasks. Although introduced for classification, TRICL extends to other task types whenever task-related signals can constrain the target output space (e.g., animal weight→species recognition, human height→weight estimation). For tasks with *continuous score outputs*, we discretize the continuous score range $[q, p]$ into u uniform bins of width $\Delta = \frac{p-q}{u}$. The i-th interval is

$$[q+i\Delta,\ q+(i+1)\Delta], \quad i = 0, 1, \cdots, u-1. \quad (9)$$

Index i are treat as the corresponding class label.

4 Experiment

4.1 Experiment Setup

Tasks, data, and LLMs. We evaluate TRICL on three NLP tasks over four datasets: text classification (SST5, Emotion (Saravia et al., 2018)), NLI (SNLI (Bowman et al., 2015)), and STS (STSB). The corresponding TRRD and related dataset are summarized in Table 1. Main results use three open-source LLMs: GLM4-9B, LLaMA 3.1-8B, and Qwen2.5-7B. We further report results on larger and smaller models (LLaMA 3.3-70B, Qwen3 32B and Qwen3-4B), as well as on ChatGPT, in additional experiments. LLM URLs are provided in Table 19.

Metrics and baselines. We report accuracy for classification and NLI task, and mean squared error (MSE) for STS task. We evaluate ICL baselines on the same task-related retrieval datasets as baselines, including TTF , CR (Li and Qiu, 2023b), KATE, CD (Naik et al., 2023), ICCL (Liu et al., 2024b), PPL (Webson and Pavlick, 2022), CEIL (Ye et al., 2023), MoD (Wang et al., 2024)).

Other setup. We use T5 (Raffel et al., 2020a) as the retrieval encoder and report 5-shot/10-shot results unless otherwise stated. For STSB, we discretize scores in $[0, 5]$ into $[0, 2.5)$ and $[2.5, 5]$ to form a classification-like setting. As in Table 1, we use one dataset for testing and the other three as task-related retrieval datasets. Due to computational constraints, main results are from a single run.

Table 1: Specific description of the test dataset and the task-related retrieval dataset. Let ODTTR denote the output space of the test dataset and the task-related retrieval dataset. **We further discuss in Section P the setting where the related dataset consists of multiple datasets . In this case, TRICL still achieves SOTA performance.**

Test dataset	Task-related retrieval dataset	Related dataset	ODTTR
SST5 (output: sentiment label | input: a text)	Emotion, SNLI, STSB	Emotion	Label to Label
Emotion (output: emotion label | input: a text)	SST5, SNLI, STSB	SST5	Label to Label
STSB (output: similarity score | input: two texts)	SST5, Emotion, SNLI	SNLI	Score to Label
SNLI (output: relation label | input: two texts)	SST5, Emotion, STSB	STSB	Label to Score

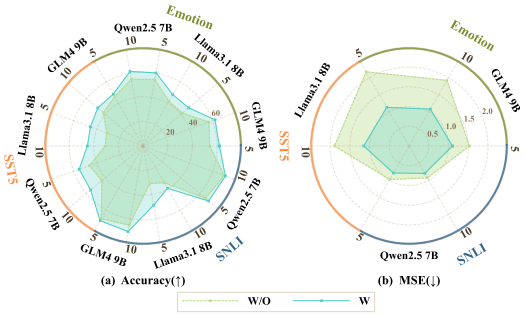


Figure 5: TRICL ablation: with vs. without related dataset extraction.

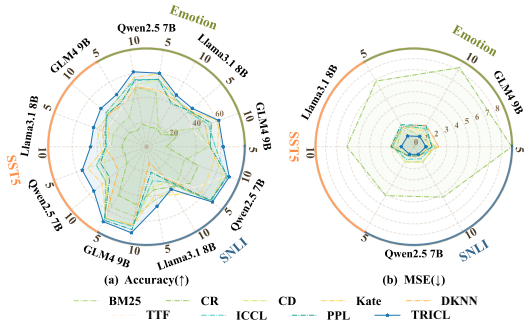


Figure 6: TRICL ablation: sensitivity to different in-context retrieval methods.

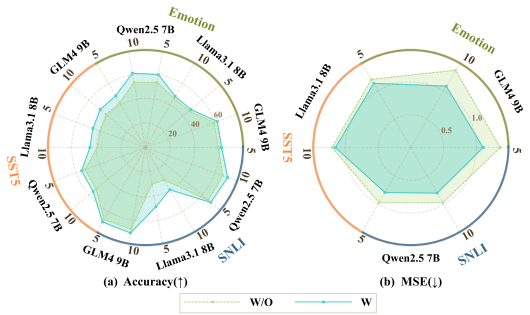


Figure 7: TRICL ablation: with vs. without reducing the output space.

5 Analysis and Discussion

5.1 Main Results

We compare TRICL with ICL baselines under the task-related retrieval dataset setting. As shown in Table 2, TRICL consistently outperforms all baselines across all datasets and three LLMs, achieving up to 71.00% accuracy on classification/NLI tasks and 0.79 MSE on STS task. In contrast, directly applying ICL with task-related but label-mismatched demonstrations is often ineffective, highlighting the importance of extracting task-related signals and simplifying the output space when retrieval and test labels are not aligned. **We further discuss in Appendix Section P the setting where the related dataset consists of multiple datasets (e.g., for Emotion, the related datasets consisted of SST5 and IMDB datasets). In this case, TRICL still achieves SOTA performance.**

5.2 Ablation Study

Effect of Related Dataset Extraction. Removing the related dataset extraction step and directly applying TRICL to the TRRD leads to a clear performance drop. (see Fig. 5). For example, accuracy on Emotion with GLM4-9B decreases from 61.85% to 54.00%, indicating that identifying a compact related subset is critical under label-space mismatch. *Beyond accuracy, related dataset extraction also improves efficiency:* filtering the TRRD into a smaller subset reduces the retrieval pool size and can lower retrieval time in practice.

Sensitivity to the Retrieval Module. We replace only the retrieval component with alternative ICL retrievers (BM25, CR, CD, KATE, Random, Static, ICCL, and PPL), while keeping the remaining components unchanged. As shown in Fig. 6, performance generally degrades; for instance, BM25 reduces accuracy on Emotion and GLM4-9B LLM from 61.85% to 50.00%. This highlights the importance of retrieving demonstrations that better support task-related inference.

Role of Output Space Reduction. TRICL constrains second-stage prediction to a reduced candidate test label space. Removing this constraint

Table 2: Comparison of TRICL with other ICL methods. The best results are highlighted in bold, while the second-best results are underlined.

Data	Emotion (Accuracy%↑)						SST5 (Accuracy%↑)					
LLMs	GLM4 9B		Llama3.1 8B		Qwen2.5 7B		GLM4 9B		Llama3.1 8B		Qwen2.5 7B	
Shots	5	10	5	10	5	10	5	10	5	10	5	10
BM25	39.5	33.3	19.3	20.6	15.5	25.7	27.6	29.2	22.4	20.6	36.5	35.8
CR	16.1	16.5	16.3	18.7	22.4	20.1	14.8	13.8	16.4	18.3	15.2	15.4
CD	47.4	25.5	34.8	33.5	31.8	47.8	35.3	20.8	29.2	23.1	16.6	13.6
Kate	30.0	12.8	25.3	23.8	18.8	17.8	29.4	22.6	20.1	17.4	14.3	15.9
DKNN	46.0	19.3	36.1	35.5	17.0	27.8	32.2	24.5	32.4	16.3	13.5	11.7
TTF	25.5	25.5	36.1	43.8	47.7	43.5	39.9	26.3	35.1	29.4	29.4	31.0
CEIL	44.1	34.5	34.0	37.7	38.2	39.5	33.1	29.0	30.2	26.9	26.7	22.4
MoD	39.7	29.7	30.5	34.6	35.8	32.5	32.8	26.2	23.1	27.6	28.8	25.0
ICCL	31.5	13.1	23.5	25.7	20.4	18.2	28.3	22.2	23.8	24.6	11.7	16.8
PPL	35.6	11.6	25.7	26.9	18.2	15.7	29.1	20.4	21.4	16.7	19.1	15.5
TRICL	**61.9**	**62.1**	**48.2**	**48.1**	**60.2**	**61.4**	**48.5**	**48.1**	**45.5**	**45.2**	**55.2**	**55.5**
	SNLI (Accuracy%↑)						STSB (MSE↓)					
BM25	52.0	51.1	23.4	25.5	16.0	18.2	3.82	3.82	4.51	4.54	3.95	3.91
CR	33.6	31.6	23.1	23.8	27.2	29.1	8.09	8.21	7.93	7.23	8.30	8.79
CD	53.7	56.0	33.4	25.1	53.4	50.0	1.51	1.62	4.54	4.55	1.97	1.98
Kate	55.4	54.6	29.0	28.3	53.8	55.7	1.33	1.16	4.54	4.58	1.56	1.84
DKNN	53.7	52.9	28.1	27.4	52.2	54.0	1.29	1.13	4.40	4.44	1.51	1.78
CEIL	42.1	40.2	28.7	30.9	41.3	34.3	5.20	4.66	5.01	4.55	3.34	2.42
MoD	52.4	45.0	26.6	33.7	53.0	39.7	1.40	1.43	2.45	2.33	1.34	1.11
TTF	57.2	53.4	18.6	18.7	49.8	56.0	1.73	1.56	3.99	4.00	1.52	1.36
ICCL	53.0	53.8	29.7	25.2	52.1	51.4	1.47	1.40	4.41	4.66	1.62	1.96
PPL	55.2	52.8	22.1	21.8	20.6	54.7	1.52	1.35	4.23	4.32	1.50	1.68
TRICL	**69.8**	**70.5**	**48.8**	**39.6**	**69.3**	**71.0**	**1.10**	**1.08**	**1.13**	**1.15**	**0.79**	**0.80**

and predicting over the full label space degrades performance (see Fig. 7); on Emotion (GLM4-9B), accuracy drops from 61.85% to 56.15%. These results confirm that output-space reduction is a key driver of TRICL's gains.

Robustness Analysis. TRICL involves several design choices (e.g., retrieval encoder, similarity metric, demonstration number and ordering, and interval partition). We conduct perturbation tests by varying one component at a time. Overall, TRICL is robust: changing the encoder/metric or ordering direction causes only minor fluctuations, with no consistently dominant configuration. Increasing the demonstration number shows non-monotonic gains, **while fewer intervals generally perform better**. Full results are in Section D. and a discussion on the limitations of similarity-based dataset construction is included in Section O of the appendix.

6 Discussion

6.1 Experiments on Additional LLMs

To further assess the generality of TRICL, we additionally evaluate it on a broader set of LLMs from 1B-70B and GPT-4o. As shown in Table 3, TRICL consistently outperforms the corresponding baselines, achieving a peak accuracy of 67.4%.

6.2 Comparison with Prompting Learning and CoT

We further compare TRICL with prompt learning and Chain-of-Thought (CoT). As Table 6 shows, TRICL performs best overall, outperforming the second-best method by 5.4% accuracy points on

Table 3: 10-shot results on Emotion with additional LLMs. LLM 1–5 denote LLaMA 3.1 1B, Qwen 4B, Qwen3 32B, LLaMA 3 70B, GPT-4o.

LLMs	LLM 1	LLM 2	LLM 3	LLM 4	LLM 5
BM25	26.8	35.5	36.5	36.8	32.8
CR	11.3	19.2	30.7	32.5	36.4
CD	31.0	27.2	47.7	45.7	44.6
Kate	20.8	16.9	41.9	40.1	35.6
DKNN	30.4	22.3	48.4	47.8	30.9
TTF	17.2	27.2	48.2	52.4	58.9
CEIL	29.0	35.3	46.8	49.4	49.5
MoD	26.1	30.9	43.4	46.6	45.7
ICCL	21.8	17.2	40.2	41.9	36.9
PPL	23.6	14.7	38.8	40.2	32.3
TRICL	**33.4**	**56.7**	**63.7**	**65.8**	**67.4**

Table 4: Plug-and-play results with TRICL.

	LLMs	A	B	C
W/o TRICL	CR	16.5	18.7	20.1
	CD	25.5	33.5	47.8
With TRICL	CR	56.8	51.0	59.4
	CD	59.9	48.7	52.1

average for classification/NLI and reducing MSE by 0.11% on average for regression.

6.3 Plug-and-Play Integration

TRICL can be used as a lightweight, plug-and-play inference module on top of existing ICL methods. We integrate TRICL's output-space compression into two representative ICL baselines, CR and CD, and evaluate them on Emotion with 10 demonstrations using GLM4 9B. As shown in Table 4, applying TRICL improves the average accuracy by 27.6%. These results suggest that TRICL complements prior ICL approaches and yields consistent gains without modifying their core components.

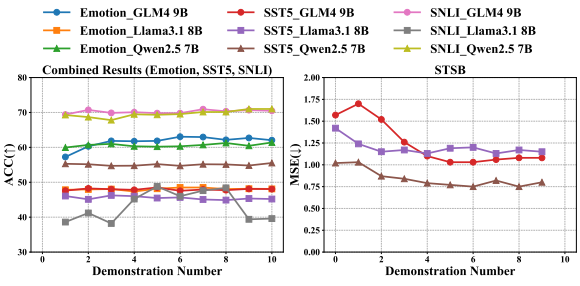


Figure 8: Performance of TRICL under varying numbers of demonstrations.

Table 5: 10-shot results on Emotion for two stages.

LLMs	GLM4 9B		LLaMA3.1 8B		Qwen2.5 7B	
Shots	5	10	5	10	5	10
1st-stage	94.0	93.1	88.2	88.6	93.8	94.0
2nd-stage	61.9	62.1	48.2	48.1	60.2	61.4

6.4 Effect of the Demonstration Number

Figure 8 shows the performance of TRICL under different numbers of demonstrations. Overall, the performance remains relatively stable across different tasks and models, indicating that TRICL is robust to the choice of demonstration number. For the classification and NLI tasks, including Emotion, SST5, and SNLI, the accuracy generally improves from the low-shot setting to a moderate number of demonstrations, and then tends to plateau with only minor fluctuations. Among them, SNLI consistently achieves strong performance across all three LLMs. For the regression task STSB, the MSE generally decreases as the number of demonstrations increases, especially in the low-shot regime, and then becomes stable after a moderate number of demonstrations. These results suggest that TRICL is not overly sensitive to the exact number of demonstrations, and that a moderate number of demonstrations is usually sufficient to achieve strong performance.

6.5 Experiments with the Related Dataset Setting

To provide a fairer comparison, we also allow other ICL methods to retrieve demonstrations from the *related dataset* identified by TRICL. In other words, all compared methods use the same related retrieval pool for in-context learning. As shown in Table 20, TRICL still achieves state-of-the-art performance, indicating that its advantage does not come merely from improved retrieval data, but from the overall TRICL framework.

6.6 Two-Stage Accuracy Analysis

We report the accuracies of both stages in TRICL's two-stage inference. As shown in Table 5, the first stage achieves a high average accuracy of 92.0%, providing a reliable basis for output-space reduction.

6.7 Variants of TRICL for Generation Tasks

TRICL also extends to **generation**. We evaluate summarization with STS-B (STS) as the retrieval

Table 6: Comparison of TRICL with prompt learning and CoT. Best results are in bold and second-best are underlined. LLM A/B/C denote GLM4 9B, LLaMA 3.1 8B, and Qwen2.5 7B, respectively.

Data	Emotion(Accuracy % ↑)			SST5(Accuracy % ↑)			SNLI(Accuracy% ↑)			STSB(MSE ↓)		
LLMs	LLM A	LLM B	LLM C	LLM A	LLM B	LLM C	LLM A	LLM B	LLM C	LLM A	LLM B	LLM C
Prompt	53.2%	42.5%	53.9%	41.9%	39.2%	52.3%	67.3%	30.2%	66.6%	1.36	1.26	0.97
COT	**64.6%**	38.0%	54.3%	45.5%	33.2%	53.6%	69.7%	27.5%	67.7%	1.24	1.18	0.94
TRICL	62.1%	**48.1%**	**61.4%**	**48.1%**	**45.2%**	**55.5%**	**70.5%**	**39.6%**	**71.0%**	**1.08**	**1.15**	**0.80**

Table 7: The average token usage and time cost on the emotion dataset across three LLMs.

	BM25	CR	CD	Kate	DKNN	TTF	CEIL	MoD	ICCL	PPL	TRICL
Time (s)	3.950	3.679	0.260	0.266	0.351	0.215	0.314	0.321	0.357	0.325	0.455
Token	195.9	194.0	202.3	189.3	203.8	211.3	203.8	213.8	196.2	178.6	209.6

dataset: we generate multiple candidate summaries and use TRICL+STS to select the best one. As shown in Table 21, it achieves SOTA performance (ROUGE-1 $= 0.244$); details are in Appendix Section K.

6.8 Cost

TRICL issues two queries per test input, introducing additional time and token overhead. As shown in Table 11, its processing time and token consumption remain modest compared to other methods. We argue that this overhead is justified by the significant performance gains.

7 Conclusion

In this paper, we find that the retrieval dataset does not need to resemble the test dataset, but only needs to be related to the test task. Based on this insight, we propose TRICL, a method for identifying the related dataset and related task, and extend it to a wide range of downstream tasks. Additionally, we develop TRICL variants tailored to traditional ICL and task-unrelated settings, leading to improved inference performance in both cases.

8 Limitations

TRICL adds an extra preprocessing step to identify task-related datasets, increasing one-time compute and latency relative to standard ICL. In our setting, this overhead is amortized across many queries and is outweighed by the consistent performance gains, but it may be less attractive under strict real-time or extremely large-scale retrieval constraints.

References

Hadi Askari, Shivanshu Gupta, Terry Tong, Fei Wang, Anshuman Chhabra, and Muhao Chen. 2025. Unraveling indirect in-context learning using influence functions. *CoRR*.

José M Bernardo and Adrian FM Smith. 2009. *Bayesian theory*, volume 405. John Wiley & Sons.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Peng Chen, Fangjun Huang, and Chao Huang. 2026. Dyc-clip: Dynamic context-aware multi-modal prompt learning for zero-shot anomaly detection. *Pattern Recognition*, page 113215.

Chen Cheng, Xinzhi Yu, Haodong Wen, Jinsong Sun, Guanzhang Yue, Yihao Zhang, and Zeming Wei. 2024. Exploring the robustness of in-context learning with noisy labels. *arXiv preprint arXiv:2404.18191*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Luc Devroye, László Györfi, and Gábor Lugosi. 2013. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference*

on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net.

Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. 2023. Mitigating label biases for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14014–14031.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Karan Gupta, Sumegh Roychowdhury, Siva Rajesh Kasa, Santhosh Kumar Kasa, Anish Bhanushali, Nikhil Pattisapu, and Prasanna Srinivasa Murthy. 2023. How robust are llms to in-context majority label bias? *arXiv preprint arXiv:2312.16549*.

Zheng Yi Ho, Siyuan Liang, Sen Zhang, Yibing Zhan, and Dacheng Tao. 2024. Novo: Norm voting off hallucinations with attention heads in large language models. *arXiv preprint arXiv:2410.08970*.

Fanding Huang, Guanbo Huang, Xiao Fan, Yi He, Xiao Liang, Xiao Chen, Qinting Jiang, Faisal Nadeem Khan, Jingyan Jiang, and Zhi Wang. 2026. Semantic-space exploration and exploitation in rlvr for llm reasoning. *Preprint*, arXiv:2509.23808.

Fanding Huang, Jingyan Jiang, Qinting Jiang, Hebei Li, Faisal Nadeem Khan, and Zhi Wang. 2025. Cosmic: Clique-oriented semantic multi-space integration for robust clip test-time adaptation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9772–9781.

Xiaonan Li and Xipeng Qiu. 2023a. Finding support examples for in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6219–6235.

Xiaonan Li and Xipeng Qiu. 2023b. Mot: Pre-thinking and recalling enable chatgpt to self-improve with memory-of-thoughts. *CoRR*.

Xiaoxia Li, Siyuan Liang, Jiyi Zhang, Han Fang, Aishan Liu, and Ee-Chien Chang. 2024a. Semantic mirror jailbreak: Genetic algorithm based jailbreak prompts against open-source llms. *arXiv preprint arXiv:2402.14872*.

Xiping Li, Xiangyu Dong, Xingyi Zhang, Kun Xie, Yuanhao Feng, Bo Wang, Guilin Li, Wuxiong Zeng, Xiujun Shu, and Sibo Wang. 2025. Chi-square wavelet graph neural networks for heterogeneous graph anomaly detection. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 1565–1576.

Xiping Li and Jianghong Ma. 2026. Aim-cot: Active information-driven multimodal chain-of-thought for vision-language reasoning. *Preprint*, arXiv:2509.25699.

Xiping Li, Jianghong Ma, Kangzhe Liu, Shanshan Feng, Haijun Zhang, and Yutong Wang. 2024b. Category-based and popularity-guided video game recommendation: a balance-oriented framework. In *Proceedings of the ACM Web Conference 2024*, pages 3734–3744.

Xiping Li, Aier Yang, Jianghong Ma, Kangzhe Liu, Shanshan Feng, Haijun Zhang, and Yi Zhao. 2026. Cpgrec+: A balance-oriented framework for personalized video game recommendations. *ACM Transactions on Information Systems*, 44(3):1–44.

Ziqian Lin and Kangwook Lee. 2024. Dual operating modes of in-context learning. *ICLR*. Available at https://openreview.net/forum?id=HkAtRP9cOY.

Xinwei Liu, Xiaojun Jia, Jindong Gu, Yuan Xun, Siyuan Liang, and Xiaochun Cao. 2023. Does few-shot learning suffer from backdoor attacks? *arXiv preprint arXiv:2401.01377*.

Xinwei Liu, Xiaojun Jia, Yuan Xun, Siyuan Liang, and Xiaochun Cao. 2024a. Multimodal unlearnable examples: Protecting data against multimodal contrastive learning. *arXiv preprint arXiv:2407.16307*.

Yinpeng Liu, Jiawei Liu, Xiang Shi, Qikai Cheng, Yong Huang, and Wei Lu. 2024b. Let's learn step by step: Enhancing in-context learning ability with curriculum learning. *arXiv preprint arXiv:2402.10738*.

Shuang Luo, Linjun Shou, Shuming Ma, Jiawei Liu, and Zhiwei Zhang. 2024. In-context learning with retrieval: A survey. Preprint. arXiv:2401.06247.

Xinxi Lyu, Sewon Min, Iz Beltagy, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Z-ICL: Zero-shot in-context learning with pseudo-demonstrations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2304–2317.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064.

Ranjita Naik, Varun Chandrasekaran, Mert Yuksekgonul, Hamid Palangi, and Besmira Nushi. 2023. Diversity of thought improves reasoning abilities of large language models.

Jongho Park, Jaeseung Park, Zheyang Xiong, Nayoung Lee, Jaewoong Cho, Samet Oymak, Kangwook Lee, and Dimitris Papailiopoulos. 2024. Can mamba learn how to learn? a comparative study on in-context learning tasks. *arXiv preprint arXiv:2402.04124*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark,

Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020a. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020b. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Hsuan Chen. 2018. Carer: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop on BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9840–9855.

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

Wenqiang Wang and Yangshijie Zhang. 2025. Incomplete in-context learning. *arXiv preprint arXiv:2505.07251*.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Jerry Wei, Le Hou, Andrew Kyle Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, and Quoc V Le. 2023. Symbol tuning improves in-context learning in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 968—979.

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1423–1436.

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

Zonghao Ying, Aishan Liu, Tianyuan Zhang, Zhengmin Yu, Siyuan Liang, Xianglong Liu, and Dacheng Tao. 2024. Jailbreak vision language models via bi-modal adversarial prompt. *arXiv preprint arXiv:2406.04031*.

Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. 2024. Visual in-context learning for large vision-language models. *arXiv preprint arXiv:2402.11574*.

Yucheng Zhou, Tao Shen, Xiubo Geng, Chongyang Tao, Can Xu, Guodong Long, Binxing Jiao, and Daxin Jiang. 2023. Towards robust ranker for text retrieval. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5387–5401. Association for Computational Linguistics.

Overview of the Appendix

This appendix provides supplementary materials and detailed analyses to support the main paper. The content is organized as follows:

- **Section A** presents the table of notations used throughout the paper.
- **Section B** provides an extended discussion of related work.
- **Section C** details the implementation of retrieval and related-task inference.
- **Section D** investigates the robustness of TRICL against various perturbations.
- **Section E** examines the impact of interval numbers on performance.
- **Section F** presents ablation studies on task-related dataset retrieval strategies.
- **Section G** explores additional downstream task scenarios, including Score-to-Score and Label-to-Text settings.
- **Section H** details the specific experimental setup for the results reported in Table 2.
- **Section I** evaluates the scalability of TRICL across partial/disjoint retrieval and traditional ICL scenarios.
- **Section J** provides comprehensive definitions of tasks, datasets, metrics, and baselines.
- **Section K** discusses experiments on the text summarization task.
- **Section L** provides the mathematical proof for the output space reduction proposition.
- **Section M** details the procedure for identifying task-related datasets.
- **Section O** discusses the limitations of semantic similarity-based retrieval.
- **Section P** analyzes performance when using multiple related datasets.
- **Section Q** includes the statement regarding the use of large language models.

A Symbol Table

The notations employed throughout this paper are summarized in Table 8.

B Related Work

In-context learning (ICL) has emerged as a powerful paradigm enabling large language models (LLMs) to perform downstream tasks using a few input-output demonstrations without gradient-based fine-tuning. Its flexibility has led to extensive research into factors influencing performance, such as demonstration selection (Min et al., 2022; Cheng et al., 2024; Fei et al., 2023; Gupta et al., 2023; Li and Qiu, 2023a; Lyu et al., 2023; Wei et al., 2023), demonstration formatting (Min et al., 2022), ordering (Wu et al., 2023), and label assignment (Wang et al., 2023). A growing body of work has explored integrating retrieval mechanisms into ICL, forming retrieval-augmented ICL (RetICL) (Luo et al., 2024), where demonstrations are retrieved from external corpora to improve prediction. Retrieval strategies include sentence embedding similarity, diversity-based selection, and contrastive learning-based retriever training. Luo et al. provide a comprehensive RetICL survey, emphasizing its role in enhancing robustness and generalization (Luo et al., 2024). However, most existing work assumes aligned output spaces between retrieved demonstrations and test queries, an assumption that often fails in real-world tasks like cross-domain sentiment classification or multilingual topic detection. To address this gap, we introduce the Disjoint Output Spaces In-Context Learning (DOSICL) setting, which evaluates LLM generalization when the retrieval and test sets have disjoint label spaces.

Parallel to ICL, visual in-context learning (VICL) extends these ideas to vision-language tasks. VICL consists of Visual Demonstration Retrieval, Intent-Oriented Image Summarization, and Demonstration Composition. Demonstrations are retrieved based on both visual features and textual descriptions using image encoders such as ViT (Dosovitskiy et al., 2021), with semantic relevance refined via cross-modal reranking using vision-language models like CLIP (Radford et al., 2021). This ensures contextual alignment between query and demonstration (Zhou et al., 2023). VICL builds on the emergent abilities of LLMs to adapt through prompting without parameter updates (Radford et al., 2019; Raffel et al., 2020b; Wei et al., 2022; Fu et al., 2023; Li et al., 2025, 2026; Li and Ma, 2026; Li et al., 2024b; Huang et al., 2026, 2025; Chen et al., 2026).

Furthermore, theoretical perspectives model ICL as implicit meta-learning or Bayesian infer-

Table 8: Symbol Table for Task-related In-context Learning

Symbol	Definition
D_t	Test dataset: $\{x_1^t, x_2^t, \ldots, x_n^t\}$
D_d	Disjoint retrieval dataset: $\{(x_i^d, y_i^d)\}_{i=1}^m$
y_i^t	Ground-truth label of test input x_i^t
$\hat{y}_i^t$	Predicted label of test input x_i^t
A_i	Labels in the output space of the test task
B_j	Labels in the output space of the disjoint retrieval dataset
C_w	Labels in the task-related retrieval dataset
D_{de}	Demonstration set used in in-context learning
D_{tr}	Task-related retrieval dataset
f_{LLM}	The large language model function used for inference
f_{pre}	Pre-trained encoder for feature extraction
e_i^t	Embedding of test input x_i^t
e_j^{tn}	Embedding of task-related example x_j^{tn}
s_{ij}	Cosine similarity between e_i^t and e_j^{tn}
N_{total}	The total Number of samples
$p(A_i)$	Marginal probability of A_i
$p(A_i\|B_j)$	Conditional probability of A_i given B_j
τ	Similarity threshold between test and retrieval inputs
k	Number of demonstrations used for in-context learning
Δ	Interval length used in discretization of regression outputs

ence, where models either recognize known data-generating functions or learn new ones in context. This view supports ICL's capacity to generalize beyond memorized mappings. Empirical studies also examine context length scaling, demonstrating improved performance with larger context windows, though long-range dependencies remain challenging. Distinctions between task recognition and task learning have been proposed (Lin and Lee, 2024), highlighting different modes of ICL adaptation. Recent advances include many-shot ICL, hybrid architectures like MambaFormer (Park et al., 2024), and low-resource prompting for instruction-following. Despite significant progress, challenges remain, particularly under disjoint label conditions. Our DOSICL benchmark addresses this by providing a new lens for evaluating LLM robustness to task-level distributional shifts.

Despite progress, most approaches assume aligned label spaces between retrieved and test samples. Our work addresses this by proposing the disjoint retrieval dataset setting, which evaluates generalization when label spaces differ across retrieval and test sets.

C Detailed Retrieval and Related-Task Inference

This section provides implementation-level details of the retrieval and related-task inference module used in TRICL. The overall procedure follows standard semantic retrieval practices widely adopted in prior ICL methods (e.g., KATE and VICL-rerank (Zhou et al., 2024)), with the key distinction that retrieval is restricted to the related dataset $\mathbf{D}_{\text{r}}$ identified in Section 3.1.

C.1 Semantic Representation

Given a test instance $x_i^{\text{t}} \in \mathbf{D}_{\text{t}}$ and the related dataset $\mathbf{D}_{\text{r}} = \{x_j^{\text{r}}\}_{j=1}^{|\mathbf{D}_{\text{r}}|}$, we first compute dense semantic representations using a pre-trained encoder f_{pre}. Specifically, the test input and each candidate instance are encoded as

$$\mathbf{e}_i^{\text{t}} = f_{\text{pre}}(x_i^{\text{t}}), \quad \mathbf{e}_j^{\text{r}} = f_{\text{pre}}(x_j^{\text{r}}), \quad x_j^{\text{r}} \in \mathbf{D}_{\text{r}}. \tag{10}$$

The encoder f_{pre} is fixed during inference and shared across all experiments. In our implementation, f_{pre} is instantiated using a standard sentence-level encoder.

C.2 Similarity Computation

We measure semantic similarity between the test instance and each candidate in $\mathbf{D}_{\text{r}}$ using cosine

similarity:

$$s_{ij} = \frac{\mathbf{e}_i^{\mathrm{t}} \cdot \mathbf{e}_j^{\mathrm{r}}}{\|\mathbf{e}_i^{\mathrm{t}}\| \, \|\mathbf{e}_j^{\mathrm{r}}\|}. \tag{11}$$

Cosine similarity is chosen for its widespread use in prior ICL retrieval methods and its robustness to vector magnitude variations. No additional normalization or re-ranking heuristics are applied unless otherwise stated.

C.3 Top-k Retrieval and Ordering

Given the similarity scores $\{s_{ij}\}_{j=1}^{|\mathbf{D}_{\mathrm{r}}|}$, we select the top-$k$ instances with the highest similarity values. The selected instances are then ordered in descending order of s_{ij} to form the demonstration set:

$$\mathbf{D}_{\mathrm{de}} = \{(x_l^{\mathrm{de}}, y_l^{\mathrm{de}})\}_{l=1}^{k}. \tag{12}$$

Here, x_l^{de} denotes the l-th retrieved demonstration, and $y_l^{\mathrm{de}} \in \{C_1, C_2, \ldots, C_{m_c}\}$ is its corresponding label from the related task. If multiple candidates have identical similarity scores, ties are broken arbitrarily. In all experiments, the value of k is fixed within a task and reported in the main text.

C.4 Related-Task In-Context Inference

Using the ordered demonstration set $\mathbf{D}_{\mathrm{de}}$, TRICL performs an initial in-context inference on the related task. Conditioned on $\mathbf{D}_{\mathrm{de}}$, the LLM predicts a related-task label for the test input:

$$\hat{y}_i^{\mathrm{r}} = f_{\mathrm{LLM}}(x_i^{\mathrm{t}}, \mathbf{D}_{\mathrm{de}}) \in \{C_1, C_2, \ldots, C_{m_c}\}. \tag{13}$$

This prediction serves as intermediate, task-related evidence that is subsequently used to reduce and transform the output space of the target task.

D Robustness Analysis

TRICL involves several design choices in its experimental pipeline, including the *pre-trained encoder* for retrieval representations, the *similarity metric*, the *number and ordering of demonstrations*, and the *interval discretization* for numerical outputs. To assess robustness, we perform one-factor-at-a-time perturbation tests: we vary one component while keeping all other settings fixed.

The effects of varying semantic similarity metrics, pre-trained encoders, and demonstration orders in TRICL are illustrated in Figure 9 and Tables 12, 13, and 14, respectively. These factors contribute to stochastic fluctuations in the experimental results.

Encoder / similarity / ordering perturbations (STSB). We evaluate three retrieval encoders: T5 (Raffel et al., 2020a), BERT (Devlin et al., 2019), and CLIP (CLIP is a multimodal model that can provide text embeddings). We also compare cosine similarity, Euclidean distance, and Manhattan distance, and test both ascending and descending similarity orderings. For STSB, we report **Acc** as the **coarse interval prediction accuracy** in the intermediate stage (predicting the score bin), and **MAE** on the **final continuous score prediction**. All numbers below are computed by first taking the **test-set mean** for each LLM and then **averaging over the three LLMs**.[1]

Overall, these choices only introduce *minor* fluctuations. For example, averaged over the three LLMs, replacing the retrieval encoder with T5, CLIP, and T5+CLIP yields Acc values of 55.7%, 56.6%, and 56.5%, with MAE values of 1.01, 1.00, and 1.01, respectively. The maximum Acc difference across these encoders is 0.9 points (56.6% vs. 55.7%), and the MAE difference is 0.01 (1.00 vs. 1.01), suggesting that TRICL does not hinge on a specific embedding backbone. Similarly, averaged over the three LLMs, cosine, Euclidean, and Manhattan metrics yield Acc values of 55.7%, 56.1%, and 56.1%, and MAE values of 1.01, 1.01, and 1.00, where the Acc range is within 0.4 points and MAE varies by at most 0.01. For demonstration ordering, averaged over the three LLMs, ascending vs. descending similarity results in Acc of 55.7% vs. 55.9% (a 0.2-point difference), and MAE of 1.01 vs. 1.02 (a 0.01 difference). Taken together, we do not observe a consistently dominant choice among encoder/metric/ordering configurations.

Demonstration budget. We vary the number of demonstrations from 1 to 10. As shown in Figure 9, performance can be non-monotonic: accuracy on text classification and NLI fluctuates with different budgets, while STSB error decreases initially and then plateaus with small variations. This indicates diminishing (and sometimes inconsistent) gains from further increasing the context budget, which aligns with the known sensitivity of ICL to context composition.

Interval discretization for numerical outputs. For score-based tasks, we discretize the score range

[1] Most robustness configurations are evaluated with a single run due to compute constraints; when multiple runs are available, we report mean±std in the corresponding additional experiments.

Variant	Acc	MAE
(A) Retrieval encoder		
T5	55.7%	1.01
CLIP	56.6%	1.00
(B) Similarity metric		
Cosine	55.7%	1.01
Euclidean	56.1%	1.01
Manhattan	56.1%	1.00
(C) Demonstration ordering		
Ascending	55.7%	1.01
Descending	55.9%	1.02

Table 9: **Robustness of TRICL on STSB under encoder/metric/ordering perturbations.** Acc denotes the intermediate-stage interval prediction accuracy (predicting the score bin), and MAE is computed on the final continuous score prediction. All values are obtained by first computing test-set means per LLM and then averaging over the three LLMs.

into u intervals and treat interval indices as labels for the intermediate decision used in output-space reduction. We test $u \in \{2, 3, 4\}$ and find that fewer intervals tend to yield better performance, as detailed in Section E. This is consistent with the intuition that coarser discretization yields a more reliable intermediate decision. Throughout the paper, we report STSB performance using the original regression metric (MSE) for the final prediction.

Takeaway. Overall, TRICL maintains stable performance under reasonable perturbations of retrieval representations, similarity metrics, demonstration ordering/budget, and interval granularities, indicating that its gains are not driven by a single fragile design choice.

E The Results of Interval Number

The output space of the STSB dataset ranges from 0 to 5. In the original study, this space is evenly divided into two intervals: $[0, 2.5)$ and $[2.5, 5]$. To explore the effect of different interval partitioning strategies on model performance, we extend this setup by increasing the number of intervals (referred to as the *Interval Number*) to 3 and 4. When the Interval Number is 3, the output space is divided into three equal intervals: $[0, 1.667)$, $[1.667, 3.333)$, and $[3.333, 5]$. When increased to 4, the output space is partitioned into four equal intervals: $[0, 1.25)$, $[1.25, 2.5)$, $[2.5, 3.75)$, and $[3.75, 5]$.

We use STSB as the test dataset and SNLI as the task-related retrieval dataset. As shown in Table 10, the performance of LLMs degrades as the Interval Number increases. Specifically, the MSE increases. For example, on the STSB dataset using the GLM4 9B model, as the Interval Number increases from 2 to 3 and subsequently to 4, the MSE increases from *1.08* to *1.12*, and then to *1.27*, indicating reduced inference performance.

Table 10: The results of different interval number metrics

	STSB (MSE↓)		
Interval Number	GLM4 9B	Llama3.1 8B	Qwen2.5 7B
2	1.08	1.15	0.80
3	1.12	1.19	0.84
4	1.27	1.23	0.96

F Ablation Study on Task-Related Dataset Retrieval

This section presents an ablation study on different strategies for retrieving task-related datasets. The goal is to evaluate whether alternative retrieval heuristics can replace or improve upon the probabilistic task-related inference used in TRICL.

Semantic similarity-based retrieval. As an alternative, we retrieve task-related datasets based on semantic similarity, measured by embedding-level similarity between datasets. Experimental results show that this strategy is effective primarily when task semantics are highly aligned. In practice, we observe clear gains only on sentiment-oriented benchmarks such as SST-5 and Emotion. However, when semantic overlap between tasks is weak or ambiguous, similarity-based retrieval fails to identify truly task-relevant datasets, leading to degraded performance.

Significance-based filtering on top of task-related inference. We further evaluate a variant that augments TRICL with a statistical significance test, selecting only significantly related datasets as context. Although this approach can filter out weakly related datasets, it requires additional cost to construct candidate related datasets and perform significance testing. More importantly, overly strict filtering may excessively compress the candidate label space. In TRICL, such over-compression can be harmful: if the retained datasets cover only a narrow or incorrect subset of labels, the second-stage

prediction may be constrained to an erroneous label range, resulting in misclassification.

Comparison with probabilistic task-related inference. Compared to the above alternatives, TRICL's end-to-end probabilistic inference consistently achieves superior performance across datasets. By softly weighting task-related evidence instead of making hard selection decisions, TRICL avoids prematurely discarding potentially useful labels and provides a more reliable foundation for the second-stage output space reduction. These results confirm that the proposed retrieval strategy is both more robust and more effective than semantic similarity- or significance-based alternatives.

G Downstream Task Scenarios

G.1 Score to Score Scenario

In prior experiments, we explored various configurations involving different output spaces for the test and task-related retrieval datasets. These configurations include: (1) both datasets having label-based output spaces (e.g., SST5 as the test dataset and Emotion as the retrieval dataset), (2) a setting where the test dataset has a label-based output space and the retrieval dataset has a score-based output space (e.g., SNLI as the test dataset and STSB as the retrieval dataset), and (3) the reverse configuration, in which the test dataset has a score-based output space and the retrieval dataset is label-based (e.g., STSB as the test dataset and SNLI as the retrieval dataset). We now extend our investigation to a setting where both the test and retrieval datasets feature score-based output spaces. Specifically, the test dataset is STSB, whose output space is [0,5], and the task-related retrieval dataset is WMT-en-cs, which includes translation quality annotations in the form of sentence pair scores ranging from 0 to 100. To ensure that the datasets remain disjoint, we select retrieval samples with scores between 10 and 100. We conduct experiments using the GLM4 9B, Llama3.1 8B and Qwen2.5 7B LLMs. As shown in Table 15, TRICL also achieves SOTA performance in this setting, attaining a best MSE of 1.03. Due to time constraints, we include only the BM25, CR, CD, Kate, Random, and Static methods as comparative baselines.

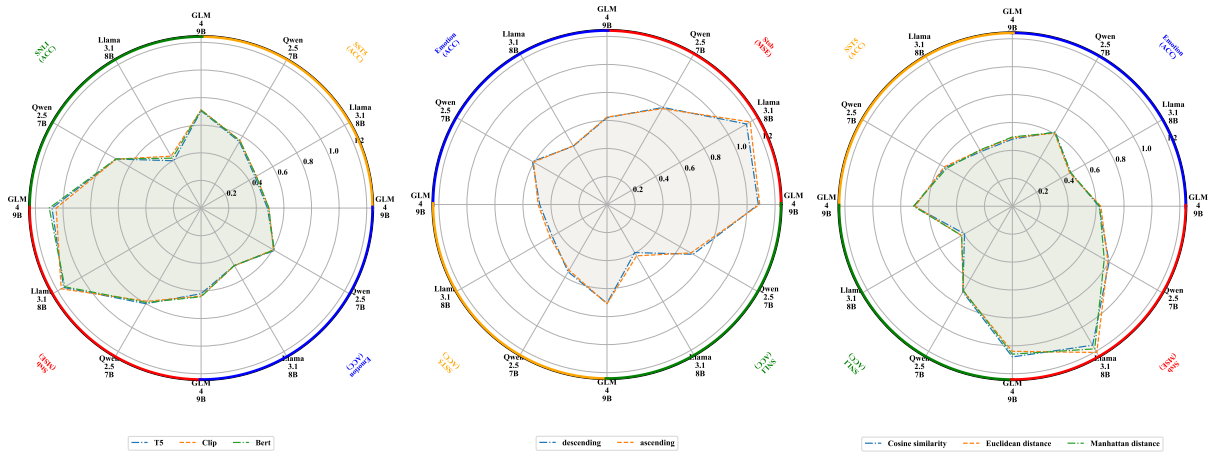
G.2 Label to Text Scenario

For the Label-to-Text scenario, we use SST5 as the test dataset and Gigaword-Tiny as the task-related retrieval dataset. Unlike previous settings, this scenario first requires collecting all instances from the test dataset. These instances are then summarized, and the resulting texts are encoded into dense representations using a pre-trained T5 model. The embeddings are subsequently clustered using the K-Means algorithm with the number of clusters set to 2. The cluster labels are then mapped to form a binary classification task. TRICL achieves an accuracy of 32.7% under this configuration. This performance is largely attributable to the fact that, during clustering, sentences with positive sentiment tend to be grouped into one cluster, while those with negative sentiment are grouped into another. Based on the cluster labels obtained from the summaries, the original five-class sentiment classification task—with labels *very negative*, *negative*, *neutral*, *positive*, and *very positive*—can be reformulated as a three-class negative sentiment task (i.e., *very negative*, *negative*, *neutral*) or a three-class positive sentiment task (i.e., *neutral*, *positive*, *very positive*). Due to time constraints, we include only the BM25, CR, CD, Kate, Random, and Static methods as comparative baselines, and conduct experiments solely on the GLM4 9B language model.

H Details of the Experimental Setup Reported in Table 2

SST5 is a five-class sentiment classification dataset with the labels *very negative*, *negative*, *neutral*, *positive*, and *very positive*. In comparison, the Emotion dataset comprises six sentiment categories: *joy*, *sadness*, *anger*, *fear*, *love*, and *surprise*. Given their thematic similarity, Emotion can serve as a task-related retrieval dataset for SST5, and vice versa.

STSB is a regression-based dataset designed to evaluate the semantic similarity between pairs of sentences, with similarity scores ranging from 0 to 5 (e.g., 2.78). SNLI, by contrast, is a large-scale benchmark dataset for sentence-level natural language inference (NLI), consisting of over 570,000 sentence pairs annotated with one of three labels: *entailment*, *contradiction*, or *neutral*. Sentence pairs labeled as *entailment* typically exhibit high semantic similarity, those labeled as *neutral* show moderate similarity, and those labeled as *contradiction* tend to have low similarity. Therefore, STSB and SNLI can be effectively used as task-related retrieval datasets for each other.



(a) (b) (c)

Figure 9: Subfigure (a) shows the results of different semantic similarity metrics. Subfigure (b) shows the results of different pre-trained encoders. Subfigure (c) shows the different orders of demonstrations in TRICL.

Table 11: The average token usage and time cost on the emotion dataset across three LLMs.

	BM25	CR	CD	Kate	DKNN	TTF	CEIL	MoD	ICCL	PPL	TRICL
Time (s)	3.950	3.679	0.260	0.266	0.351	0.215	0.314	0.321	0.357	0.325	0.455
Token	195.9	194.0	202.3	189.3	203.8	211.3	203.8	213.8	196.2	178.6	209.6

I Scalability of TRICL to Partial-and-disjoint Retrieval Dataset, traditional ICL, and Zero-shot Prompt Learning Scenarios

J Experiment Setup

J.1 Task

Text classification. Text classification is a fundamental task in Natural Language Processing (NLP) that involves assigning predefined categories or labels to textual data. It is widely applied in areas such as sentiment analysis, spam detection, topic labeling, and intent recognition. Performance is typically evaluated using metrics like accuracy, precision, recall, and F1 score.

Natural Language Inference. Natural Language Inference (NLI) is a core task in NLP that involves determining the logical relationship between a pair of sentences: a premise and a hypothesis. The goal is to classify this relationship into one of three categories: entailment (the hypothesis is definitely true given the premise), contradiction (the hypothesis is definitely false given the premise), or neutral (the hypothesis may or may not be true given the premise). NLI is essential for building systems that understand and reason about language, and it plays a key role in applications such as question answering, dialogue systems, and information retrieval. Modern approaches typically use deep learning models, especially transformer-based architectures like BERT, RoBERTa, and DeBERTa, which are fine-tuned on large-scale NLI datasets such as SNLI and MNLI. Despite its progress, NLI remains challenging due to the need for nuanced understanding of context, syntax, semantics, and world knowledge.

Semantic Textual Similarity. Semantic Textual Similarity (STS) is a fundamental task in Natural Language Processing (NLP) that focuses on quantifying the degree of semantic equivalence between two text segments, typically at the sentence level. Unlike traditional classification tasks, STS is formulated as a regression problem, where the objective is to assign a continuous similarity score—commonly on a scale from 0 (completely unrelated) to 5 (semantically equivalent)—reflecting the extent to which the two texts convey the same meaning. This task plays a crucial role in a wide range of downstream applications, including information retrieval, question answering, text summarization, and paraphrase identification. Accurate STS modeling requires a deep understanding of lexical semantics, syntactic structure, context, and often commonsense knowledge. Recent advances in STS have been driven by the

Table 12: The results of different semantic similarity metrics. These are the experimental results obtained with a context size of 10.

	Emotion (Accuracy↑)			SST5 (Accuracy↑)		
	GLM4 9B	Llama3.1 8B	Qwen2.5 7B	GLM4 9B	Llama3.1 8B	Qwen2.5 7B
Cosine similarity	62.0%	48.1%	61.4%	48.0%	45.2%	55.5%
Euclidean distance	63.4%	47.5%	60.5%	49.1%	46.1%	56.4%
Manhattan distance	62.5%	48.2%	60.9%	49.6%	46.5%	54.3%
	SNLI (Accuracy↑)			STSB (MSE ↓)		
Cosine similarity	70.4%	39.5%	71.0%	1.08	1.15	0.80
Euclidean distance	69.6%	42.1%	70.3%	1.04	1.21	0.79
Manhattan distance	70.8%	41.9%	70.0%	1.06	1.18	0.76

Table 13: The results of different pre-trained encoders. These are the experimental results obtained with a context size of 10.

	Emotion (Accuracy↑)			SST5 (Accuracy↑)		
	GLM4 9B	Llama3.1 8B	Qwen2.5 7B	GLM4 9B	Llama3.1 8B	Qwen2.5 7B
T5	62.0%	48.1%	61.4%	48.0%	45.2%	55.5%
Clip	63.7%	48.2%	60.7%	49.2%	46.2%	56.4%
Bert	64.2%	48.1%	60.8%	49.0%	46.1%	56.3%
	SNLI (Accuracy↑)			STSB (MSE ↓)		
T5	70.4%	39.5%	71.0%	1.08	1.15	0.80
Clip	71.1%	43.4%	70.8%	1.05	1.17	0.78
Bert	70.9%	41.7%	71.1%	1.10	1.14	0.79

adoption of transformer-based pre-trained language models, such as BERT, RoBERTa, and Sentence-BERT, which produce dense semantic representations of text and allow for effective computation of similarity using distance metrics like cosine similarity. Despite significant progress, challenges remain in handling idiomatic expressions, domain variability, and subtle semantic nuances.

J.2 Data

Stanford Sentiment Treebank (SST5). The Stanford Sentiment Treebank (SST) is a widely used benchmark dataset for sentiment analysis, providing fine-grained sentiment annotations for sentences and phrases extracted from movie reviews. The **SST5** variant formulates sentiment classification as a five-class problem, where each sentence is labeled with one of five sentiment categories: *very negative*, *negative*, *neutral*, *positive*, or *very positive*. Unlike binary or ternary sentiment tasks, SST5 enables the evaluation of models on more nuanced sentiment distinctions, making it particularly valuable for studying subtle linguistic cues and gradations in emotional tone. The dataset is syntactically parsed and includes phrase-level annotations based on the constituency parse tree, allowing for both sentence-level and hierarchical sentiment analysis. Due to its complexity and granularity, SST5 is commonly used to benchmark the performance of advanced language models in fine-grained sentiment classification tasks.

Emotion Dataset. The *Emotion* dataset is a benchmark corpus for multi-class emotion classification, originally introduced by Saravia et al. (Saravia et al., 2018). It consists of 20,000 English-language tweets annotated with one of six basic emotion labels: *joy*, *sadness*, *anger*, *fear*, *love*, and *surprise*. The dataset is balanced across these classes and is derived from real-world social media content, making it well-suited for evaluating models in affective computing and emotion recognition tasks. Given the informal and diverse linguistic expressions found in tweets, this dataset presents challenges such as slang, abbreviations, and figurative language. It is widely used to assess the performance of deep learning and transformer-based models in capturing emotional nuance in short, noisy text.

Table 14: The results of different orders of demonstrations in TRICL

	Emotion (Accuracy↑)			SST5 (Accuracy↑)		
	GLM4 9B	Llama3.1 8B	Qwen2.5 7B	GLM4 9B	Llama3.1 8B	Qwen2.5 7B
descending	62.0%	48.1%	61.4%	48.0%	45.2%	55.5%
ascending	62.4%	48.3%	60.5%	49.2%	46.8%	54.2%
	SNLI (Accuracy ↑)			Stsb (MSE↓)		
descending	70.4%	39.5%	71.0%	1.08	1.15	0.80
ascending	70.8%	42.3%	69.1%	1.09	1.18	0.79

Table 15: The results of score to score scenario

	GLM4 9B	Llama3.1 8B	Qwen2.5 7B
BM25	1.40	**4.17**	3.56
CR	8.29	8.29	8.29
CD	1.07	4.33	3.80
Kate	1.52	4.75	4.78
Random	1.80	4.79	5.77
Static	1.55	4.62	5.09
TRICL	**1.03**	4.28	**3.44**

Semantic Textual Similarity Benchmark (STSB). The Semantic Textual Similarity Benchmark (STSB) is a widely used dataset designed to evaluate the ability of models to capture the semantic similarity between sentence pairs. Introduced as part of the GLUE benchmark (Wang et al., 2018), STSB consists of sentence pairs drawn from various sources such as news headlines, image captions, and forum discussions. Each pair is annotated with a similarity score ranging from 0 to 5, where 0 indicates no semantic similarity and 5 denotes complete semantic equivalence. Unlike classification tasks, STSB is formulated as a regression problem, requiring models to produce continuous-valued similarity predictions. Due to its diversity and fine-grained scoring, STSB serves as a crucial benchmark for assessing sentence embeddings, semantic understanding, and the generalization capabilities of pretrained language models.

Stanford Natural Language Inference (SNLI). The Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015) is one of the most influential benchmark corpora for evaluating natural language inference (NLI) models. It consists of 570,000 human-written English sentence pairs, each comprising a *premise* and a *hypothesis*, annotated with one of three inference labels: *entailment*, *contradiction*, or *neutral*. The dataset was constructed using image captions from the Flickr30k corpus as premises, with hypotheses written and labeled by crowdworkers. SNLI played a key role in advancing the development of supervised deep learning models for textual entailment, as its scale and quality enabled effective training of complex architectures. It remains a traditional benchmark for evaluating sentence representation learning and reasoning capabilities in NLP systems.

J.3 Metrics

Accuracy. Accuracy is one of the most commonly used evaluation metrics for classification tasks. It is defined as the ratio of correctly predicted instances to the total number of predictions made, formally expressed as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

where TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively. Accuracy provides an overall measure of a model's performance across all classes, making it particularly suitable for balanced datasets. However, it can be misleading in scenarios involving class imbalance, where high accuracy may be achieved by predicting the majority class. In such cases, complementary metrics such as precision, recall, or F1 score are often used to provide a more nuanced evaluation.

Mean Squared Error (MSE). Mean Squared Error (MSE) is a widely used evaluation metric for regression tasks, measuring the average squared difference between predicted and actual values. It is formally defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{15}$$

where n denotes the number of data points, y_i represents the true value, and $\hat{y}_i$ is the predicted

Table 16: The results of the traditional ICL scenario

	SNLI (Accuracy ↑)			SST5 (Accuracy ↑)			STSB (MSE ↓)		
	GLM4 9B	Llama3.1 8B	Qwen2.5 7B	GLM4 9B	Llama3.1 8B	Qwen2.5 7B	GLM4 9B	Llama3.1 8B	Qwen2.5 7B
BM25	56.7%	45.8%	45.6%	34.1%	35.9%	46.8%	1.54	4.03	1.55
CR	37.1%	32.2%	43.6%	18.4%	14.8%	13.3%	2.91	4.06	3.41
CD	76.4%	33.6%	84.3%	43.8%	36.4%	45.9%	0.85	**1.20**	0.99
Kate	75.1%	49.1%	83.5%	49.7%	35.7%	51.8%	0.66	1.28	0.63
Random	61.8%	28.0%	75.0%	42.9%	40.0%	42.7%	0.73	2.56	0.91
Static	64.4%	20.3%	37.9%	43.4%	33.0%	43.8%	1.05	2.72	0.70
ICCL	75.6%	47.4%	83.7%	48.3%	36.9%	52.4%	0.67	1.11	0.69
PPL	74.8%	49.5%	81.1%	46.9%	36.3%	48.6%	0.65	1.76	0.68
TRICL	**78.3%**	**51.2%**	**85.8%**	**51.8%**	**45.0%**	**57.9%**	**0.64**	1.34	**0.62**

Table 17: The results of the task-unrelated scenario

Data	SST5 (Accuracy ↑)					
LLMs	GLM4 9B		Llama3.1 8B		Qwen2.5 7B	
Shot	5	10	5	10	5	10
Prompt	41.8%	41.8%	39.1%	39.1%	52.3%	52.3%
BM25	14.9%	15.3%	13.5%	12.6%	20.4%	22.6%
Cluster kate	7.7%	8.3%	9.3%	10.2%	9.3%	9.6%
Cluster_static	21.6%	16.7%	17.3%	17.8%	15.2%	17.9%
Kate	20.1%	17.6%	14.9%	11.1%	15.3%	16.8%
Random	18.8%	17.9%	16.9%	15.6%	13.0%	11.9%
Static	20.8%	16.6%	19.8%	13.7%	18.0%	16.4%
ICCL	19.1%	17.5%	17.6%	13.4%	14.0%	16.2%
PPL	18.4%	16.1%	18.3%	14.5%	16.3%	15.8%
TRICL	46.4%	46.2%	42.7%	43.3%	54.6%	54.4%

value for the i-th instance. MSE penalizes larger errors more heavily due to the squaring operation, making it sensitive to outliers. It is particularly useful when the goal is to minimize the overall magnitude of prediction errors. However, in contexts where robustness to outliers is important, alternative metrics such as Mean Absolute Error (MAE) or Huber loss may be preferred.

J.4 Baselines

We evaluate a variety of in-context learning (ICL) strategies, encompassing several distinct prompting techniques. The *Zero-shot Prompt* approach operates without incorporating any demonstrations. The *Static* method selects the k demonstrations from a retrieval dataset. In contrast, the *Random* strategy draws demonstration examples arbitrarily from the same retrieval pool for each test dataset. *Clustering-retrieval* (Li and Qiu, 2023b) partitions all available examples into k distinct clusters, designed to group semantically similar instances, and selects a representative sample from each cluster to construct the final demonstration set. The *Kate* method retrieves samples that exhibit the highest similarity in terms of sentence-level embeddings. Meanwhile, *Cluster-Diversity* (Naik et al., 2023) applies clustering to organize all demonstrations into k groups and selects the exemplar closest to the centroid of each cluster.

J.5 Zero-shot Prompt Learning Scenario

We find that the *Output Space Reduction* component in TRICL is particularly effective in the Zero-shot Prompt Learning scenario. In this setting, we employ two LLMs to generate label predictions for each test dataset. For each test instance, two predicted labels are produced, and the reduced output space is subsequently determined according to the procedure defined in TRICL. Due to time constraints, we limit our experiments to the SST5 dataset. TRICL demonstrates strong performance in this scenario, achieving a maximum accuracy of 54.8%.

Table 18: The results of ChatGPT-4o

Test Data	Emotion	SST5	SNLI	STSB
Metric	Accuracy ↑			MSE ↓
LLMs	ChatGPT-4o			
Shot	10			
BM25	43.9%	33.3%	54.9%	2.84
CR	16.0%	29.4%	53.0%	1.70
CD	51.3%	46.9%	64.5%	1.77
Kate	56.7%	47.2%	62.4%	0.82
Random	42.2%	40.7%	63.2%	0.94
Static	48.6%	40.8%	62.5%	1.13
ICCL	22.9%	35.8%	52.2%	1.44
PPL	34.1%	37.7%	57.2%	1.50
TRICL	**73.0%**	**55.4%**	**77.1%**	**0.73**

K The Experiment of Text Summarization Task

The core idea of the TRICL method is to enhance prediction performance by reducing the output space. However, when the test task involves text generation—such as text summarization—this strategy cannot be directly applied. To overcome this limitation, we propose a variant of TRICL.

When a text summarization model produces high-quality summaries, the semantic similarity between the original text and the generated summary is typically high. Based on this observation, we use the Gigaword-Tiny dataset as the test dataset and STSB as the task-related retrieval dataset. ROUGE-1 serves as the evaluation metric for the text summarization task, where a higher ROUGE-1 score indicates better performance.

Due to time constraints, we conduct experiments using only the GLM4 9B LLM, and compare TRICL with several baseline methods, including Prompt, BM25, CR, and, CD. As shown in Table 21, TRICL achieves SOTA performance in this setting, attaining a ROUGE-1 score of 0.244.

L Proof

Proof. We prove this result by contradiction. Given the fact that

$$\hat{p}(A_i^{C_w} \mid C_w) > \hat{p}(A_i^{C_w}), \tag{16}$$

and $\{A_1^{C_w}, A_2^{C_w}, \ldots, A_{m_d}^{C_w}\}$, where $A_i^{C_w} \in \{A_1, A_2, \ldots, A_{m_a}\}$, is the possible label for every label $C_w \in \{B_1, B_2, \cdots, B_{m_b}\}$, there must be $m_d \leq m_a$. Thus we assume here that the predicted output $\hat{y}_i^{\mathrm{t}}$ does not reduce the output space, i.e., $m_d = m_a$. Then for each index $i \in \{1, 2, \cdots, m_a\}$, we obtain that

$$\begin{aligned}
\hat{P}(C_w) &= \sum_{i=1}^{m_a} \hat{P}(A_i^{C_w} \mid C_w) \\
&= \hat{P}(A_1^{C_w} \mid C_w) + \cdots + \hat{P}(A_{m_a}^{C_w} \mid C_w) \\
&> \hat{P}(A_1^{C_w}) + \hat{P}(A_2^{C_w}) + \cdots + \hat{P}(A_{m_a}^{C_w}) \\
&= \sum_{i=1}^{m_a} \hat{P}(A_i^{C_w}) = 1,
\end{aligned} \tag{17}$$

which contradicts the conclusion that $\hat{P}(C_w) \leq 1$ obviously. Hence our assumption must be false, and the predicted output $\hat{y}_i^t$ indeed reduces the output space. This proves the result $m_d < m_a$. □

M Details of Related Dataset Identification

This appendix provides the full formulation and implementation details for the related dataset identification procedure described in Section 3.1.

LLM-based Label Prediction. Let $\mathbf{D}_{\mathrm{tr}} = \{(x_1^{\mathrm{d}}, y_1^{\mathrm{d}}), \ldots, (x_m^{\mathrm{d}}, y_m^{\mathrm{d}})\}$ denote the task-related retrieval dataset, where $y_t^{\mathrm{d}} \in \{B_1, \ldots, B_{m_b}\}$. We randomly sample $N_{\mathrm{total}} \leq 100$ instances from $\mathbf{D}_{\mathrm{tr}}$. For each sampled instance x^{d}, we apply zero-shot prompting with the LLM to predict a target label:

$$\hat{y}_{\mathrm{pred}}(x^{\mathrm{d}}) = f_{\mathrm{LLM}}(x^{\mathrm{d}}) \in \{A_1, \ldots, A_{m_a}\}. \tag{18}$$

Empirical Probability Estimation. On the sampled subset, we define the following counts:

$$N(A_i, B_j) = \sum \mathbb{I}[y^{\mathrm{d}} = B_j \wedge \hat{y}_{\mathrm{pred}} = A_i], \tag{19}$$

$$N(A_i) = \sum \mathbb{I}[\hat{y}_{\mathrm{pred}} = A_i], \tag{20}$$

$$N(B_j) = \sum \mathbb{I}[y^{\mathrm{d}} = B_j], \tag{21}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function.

The marginal and conditional probabilities are estimated as:

$$\hat{P}(A_i) = \frac{N(A_i)}{N_{\mathrm{total}}}, \quad \hat{P}(A_i \mid B_j) = \frac{N(A_i, B_j)}{N(B_j)}. \tag{22}$$

Retrieval labels with $N(B_j) = 0$ are excluded from the relatedness test.

Related Label and Dataset Construction. A retrieval label B_j is considered task-related if:

$$\exists\, A_i \in \{A_1, \ldots, A_{m_a}\} \text{ s.t. } \hat{P}(A_i \mid B_j) > \hat{P}(A_i). \tag{23}$$

The set of task-related labels is thus defined as:

$$\{C_1, \ldots, C_{m_c}\} = \{B_j \in \{B_1, \ldots, B_{m_b}\} \mid \exists\, A_i \text{ such that } \hat{P}(A_i \mid B_j) > \hat{P}(A_i)\Big\}. \tag{24}$$

Finally, the related dataset is constructed as:

$$\mathbf{D}_{\mathrm{r}} = \{(x^{\mathrm{d}}, y^{\mathrm{d}}) \in \mathbf{D}_{\mathrm{tr}} \mid y^{\mathrm{d}} \in \{C_1, \ldots, C_{m_c}\}\}. \tag{25}$$

N Qualitative Analysis on Emotion Classification.

We provide a qualitative example to illustrate how TRICL reduces the output space and simplifies decision making in text classification. Given the input

sentence: *"I expected something exciting, but it turned out to be disappointing."*, zero-shot prompting over the full label space produces a relatively flat posterior distribution over {*neutral*, *sadness*, *anger*}, reflecting uncertainty among semantically close labels. TRICL first performs inference on a related task that captures coarse affective polarity, predicting a negative emotional state. Conditioning on this related-task evidence, the candidate label space is reduced to {*sadness*, *anger*}, eliminating the weakly correlated label *neutral*. Subsequent zero-shot inference within this reduced space yields the correct prediction. This example demonstrates that TRICL improves robustness by removing labels that contribute disproportionately to ambiguity rather than discriminative power.

O Limitations of Semantic Similarity-Based Retrieval for Related Data

Semantic similarity-based retrieval relies on surface-level semantic alignment between the retrieval and test tasks, and thus can fail when the task form or semantics are mismatched. For example, when evaluating on SNLI, it fails to identify STSB as a related dataset despite its relevance at the supervision-signal level, leading to missed useful retrieval data; as a result, its gains are mostly limited to strongly aligned cases.

P Multiple Related Datasets

We further examine a setting where the related dataset is composed of multiple datasets. Specifically, we construct related datasets by combining two or three emotion-related corpora and evaluate performance on the Emotion benchmark with GLM4 9B. Specifically, we use the Emotion dataset for evaluation. When the related dataset consists of two datasets, it is formed by SST-5 and IMDB. When it consists of three datasets, it is formed by SST-5, IMDB, and SST-2.

As shown in Table 22, TRICL consistently outperforms all baselines under both settings. Despite the increased heterogeneity in the related data, TRICL remains effective at identifying useful task-related signals and reducing the target output space accordingly. These results suggest that TRICL is robust to variations in the composition of related datasets and does not rely on a single carefully curated retrieval source.

Q Use of Large Language Models

In accordance with the ACL policy on the use of AI writing assistance, we declare that we utilized GPT-4 to refine the clarity and readability of the text in this paper. The use of the model was strictly limited to grammatical error correction and stylistic polishing. All scientific claims, experimental designs, and data analyses remain the original work of the authors. We have reviewed the entire text to ensure accuracy and assume full responsibility for the content.

Table 19: URLs for the LLMs and VLMs.

Model	URL
GPT-4o	https://platform.openai.com/docs/models/gpt-4o
Qwen2.5-7b	https://huggingface.co/Qwen/Qwen2.5-7B
Qwen3-4b	https://huggingface.co/Qwen/Qwen3-4B
Qwen3-8b	https://huggingface.co/Qwen/Qwen3-8B
Qwen3-32b	https://huggingface.co/Qwen/Qwen3-32B
GLM4 9B	https://huggingface.co/zai-org/glm-4-9b
LLAMA-3.1-8b	https://huggingface.co/meta-llama/Llama-3.1-8B
LLAMA-3.2-1b	https://huggingface.co/meta-llama/Llama-3.2-1B
LLAMA-3.2-3b	https://huggingface.co/meta-llama/Llama-3.2-3B
LLAMA-3.3-70b	https://huggingface.co/meta-llama/Llama-3.3-70B

Table 20: The results on the related dataset.

Data	Emotion (Accuracy% ↑)			SST5 (Accuracy% ↑)		
LLMs	GLM4 9B	Llama3.1 8B	Qwen2.5 7B	GLM4 9B	Llama3.1 8B	Qwen2.5 7B
BM25	44.3%	19.2%	48.5%	27.6%	28.5%	43.0%
CR	27.2%	17.0%	18.5%	17.8%	19.8%	20.5%
CD	56.2%	42.0%	50.5%	39.8%	37.1%	41.5%
Kate	56.8%	41.8%	55.0%	42.1%	30.8%	43.7%
DKNN	50.5%	44.3%	48.8%	40.7%	34.4%	29.9%
TTF	39.3%	42.9%	51.5%	34.1%	28.4%	38.3%
CEIL	39.6%	46.5%	41.2%	36.3%	33.1%	32.4%
MoD	38.6%	51.3%	50.5%	32.6%	37.4%	38.1%
ICCL	53.2%	45.8%	55.1%	42.7%	32.2%	43.3%
PPL	55.6%	40.3%	54.4%	41.4%	33.2%	38.6%
TRICL	**62.1%**	**48.1%**	**61.4%**	**48.1%**	**45.2%**	**55.5%**

Table 21: The results of the text summarization task in gigaword dataset and GLM4 9B LLM,

Metric	Prompt	BM25	CR	CD	DKNN	TTF	CEIL	MoD	ICCL	TRICL
ROUGE-1	0.124	0.238	0.023	0.026	0.022	0.045	0.169	0.204	0.188	**0.244**

Table 22: Results on related datasets composed of two and three datasets, evaluated on the Emotion dataset with the GLM4 9B LLM.

Scenario	BM25	CR	CD	Kate	DKNN	TTF	ICCL	PPL	TRICL
2 Datasets	37.6%	11.7%	44.2%	26.3%	31.4%	33.1%	30.2%	33.7%	**62.3%**
3 Datasets	36.4%	12.3%	44.9%	26.3%	33.9%	35.9%	31.5%	32.4%	**60.5%**