CLL-RetICL: Contrastive Linguistic Label Retrieval-based In-Context Learning for Text Classification via Large Language Models

Anonymous ACL submission

Abstract

Recent research has delved into Retrieval-based In-Context Learning (RetICL), leveraging the power of large language models (LLMs) for text classification. Despite its promise, a persistent challenge lies in effectively retrieving relevant demonstrations from a support set. Many existing approaches have overlooked the essential role of linguistic label information in guiding this retrieval process. To bridge this gap, we present Contrastive Linguistic Label Retrieval-based In-Context Learning (CLL-RetICL), a novel framework designed to identify the most relevant and impactful sentences without altering the model parameters. Our approach uniquely integrates sentence-query similarity with sentence-label similarity, enabling a more nuanced and comprehensive evaluation of relevance. We tested CLL-RetICL across diverse text classification tasks and evaluated its performance on various LLMs. Experimental results demonstrate that CLL-RetICL consistently outperforms previous retrieval methods that do not incorporate linguistic label information. These findings highlight the critical importance of linguistic label-aware selection in enhancing text classification accuracy.[1]

1 Introduction

Recently, researchers have begun exploring few-shot in-context learning (ICL) using LLMs for text classification tasks. (Luo et al., 2024; Yu et al., 2023; Chae and Davidson, 2023; Rouzegar and Makrehchi, 2024). A significant advantage of ICL is particularly valuable in scenarios where fine-tuning is impractical, such as when access to model parameters is restricted, computational resources are limited, or available data is insufficient. (Loukas et al., 2023; Cahyawijaya et al., 2024; Wang et al., 2024; Milios et al., 2023). Instead of selecting static, pre-defined demonstration sets for ICL, RetICL adopts a dynamic, context-sensitive approach. At its core, adaptive demonstration selection leverages a specialized retriever to intelligently curate tailored demonstrations for each task input. RetICL has gained popularity because prior research suggests that context-insensitive demonstrations can limit the full potential of LLMs (Luo et al., 2024; Wu et al., 2022). Despite RetICL consistently surpassing approaches based on random or static demonstrations, it still remains an open challenge to retrieve relevant demonstrations.

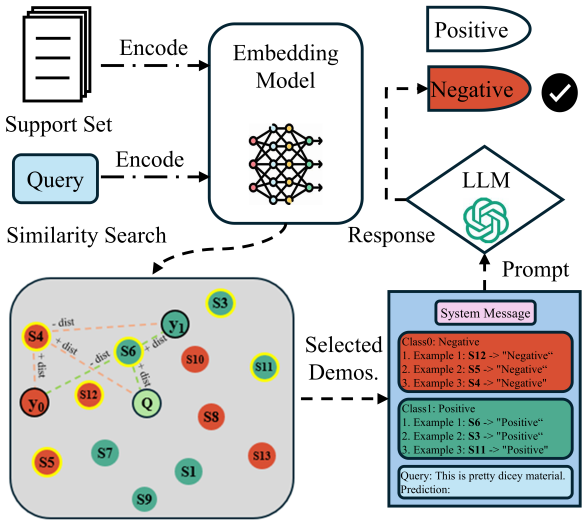
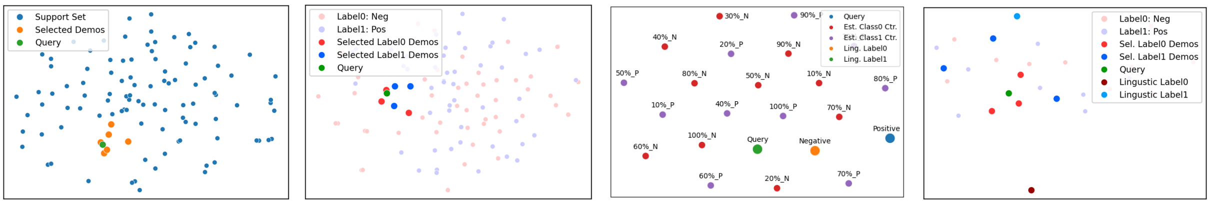


Figure 1: An illustration of CLL-RetICL with $N = 2$ and $k = 3$, demonstrating a prediction between Positive and Negative classes. Here, y_0 and y_1 represent the vector representations of the linguistic labels "Negative" and "Positive", respectively, in a pre-trained sentence embedding model. Similarly, $s_0, s_1, \ldots$ represent the vector representations of the sentences in a support set within the same pre-trained sentence embedding model.

To address the problem, previous researchers have proposed various strategies, including k-nearest neighbors (KNN), NwayKshot, and clustering-based RetICL (Li et al., 2024; Pecher et al., 2024; Zhang et al., 2022a). However, these methods suffer from various challenges, as shown in Figure 2. To identify the most effective demon-

[1]Our code is available: http://acl-org.github.io/ACLPUB/formatting.html



(a) KNN (b) NwayKshot (c) Clustering-Based (d) Ours

Figure 2: A comparison of four different approaches to RetICL strategies. (a) KNN suffers from two key weaknesses: the copying effect and misleading by similarity. (b) NwayKshot always ignores any linguistic cues conveyed through the labels. (c) Clustering-based approaches are hindered by the difficulty in estimating category centers and the neglect of query similarity. (d) Our method avoids the copying effect, prevents misleading similarity, incorporates linguistic label information, utilizes fixed label category centers, and integrates query similarity.

strations, we analyzed failure cases. Our investigation revealed that always existing a particular combination of demonstrations can enable LLMs to classify accurately. Additionally, our analysis uncovered that failure cases are error-prone: they often lie closer to the linguistic representation of an opposing label or near the center of an incorrect label cluster, despite their similarity to the query. In contrast, when the demonstrations are correctly combined, they align more closely with the intended label. A detailed discussion of these findings is presented in Section 3.

Building on these observations, we present a novel RetICL framework, CLL-RetICL (Contrastive Linguistic Label Retrieval-based In-Context Learning) as illustrated in Figure 1. Our approach introduces a trade-off method that computes a relevance score by integrating both sentence–query and sentence–label similarities, thereby effectively leveraging label information. Furthermore, to optimize the effectiveness of CLL-RetICL, we developed a universal *N-way K-shot* prompt structure applicable to all text classification tasks. This prompt design mitigates the copying effect and prevents LLMs from being misled by overly similar examples. Moreover, we demonstrate that the sentence embeddings of linguistic labels can serve as clustering centers—generated by a pre-trained sentence embedding model—to address the challenge of estimating clustering centers. Additionally, we initiate four variations for integrating the linguistic label style into RetICL and evaluate their effectiveness on four text classification datasets. Finally, to assess the generalizability of CLL-RetICL, we conduct experiments using Gemini (Team et al., 2024), Llama (Dubey et al., 2024), and Mistral (Jiang et al., 2024). Empirical experiments show that CLL-RetICL consistently outperforms both previous RetICL baselines and other variants across multiple datasets and LLMs. Ablation studies further reveal several key findings: (1) Effectiveness across variations: CLL-RetICL maintains strong performance across different k-shot settings, various pre-trained sentence embedding models, and multiple similarity functions. (2) Component dependency: The proposed method relies on the original component responsible for calculating sentence-query similarity; omitting this component degrades performance. (3) Impact of hyperparameters: Trade-off hyperparameters have a minor influence on the final classification accuracy. The following summarizes our main contributions:

- We present a novel perspective in which sentence embeddings of linguistic labels serve as highly accurate clustering centers, free from the biases introduced by limited support data and independent of data-driven constraints.
- We propose an innovative method, CLL-RetICL, which employs a rigorous relevance scoring metric that leverages linguistic label information to select high-quality demonstrations for improving LLMs in text classification tasks. Our approach does not require fine-tuning the pre-trained weights of either the sentence embedding models or LLMs.
- We conduct extensive experiments to evaluate the proposed method, achieving better performance on most datasets compared to existing RetICL methods.

2 Related Work

Text Classification via LLMs. Text classification via LLMs has recently demonstrated exceptional generalizability and reasoning capabilities, attracting significant research interest in their application to text classification tasks (Zhang et al., 2024; Wang et al., 2024; Fields et al., 2024). Existing methods can be broadly divided into two groups, depending on whether they involve adapting the parameters of LLMs or not. The first group concentrates on fine-tuning the parameters of LLMs to excel in custom text classification tasks (Chae and Davidson, 2023; Zhang et al., 2024; Yu et al., 2023; Jin et al., 2023). However, this approach generally demands significant computational resources to load the full LLM model parameters, and fine-tuning these models can often diminish their generalizability. The other category is known as ICL, or prompt engineering (Guo et al., 2024; Luo et al., 2024; Fan et al., 2024). While this method avoids the need to update LLM model parameters, it heavily depends on well-designed prompts, making it challenging to guide LLMs to consistently meet human expectations (Shi et al., 2023; Mavromatis et al., 2023; Edwards and Camacho-Collados, 2024).

RetICL. RetICL can generally be divided into two categories: approaches that retrain or fine-tune a retriever for specific text classification tasks, and approaches that utilize pre-trained language models without additional fine-tuning. An intuitive strategy for RetICL involves directly selecting a few similar sentences, leveraging readily available demonstration retrievers like those based on sentence embeddings. Existing methods include KATE (Liu et al., 2021), Z-ICL (Lyu et al., 2022) and ICL-ML (Milios et al., 2023). However, recent research has shown that selecting the most similar demonstrations can lead to the copying effect and misleading by similarity, degrading performance in text classification tasks (Olsson et al., 2022; Zhang et al., 2022b). To mitigate the issue of homogeneity in retrieval, clustering retrieval approaches ensure the selection of a diverse and representative set of demonstrations, which is critical to its effectiveness (Luo et al., 2024). Several methods exist, including NwayKshot (Li et al., 2024), Votek (Su et al., 2022) and SelfPrompt (Li et al., 2022). While these approaches leverage label information and offer improvements, accurately estimating the clustering center for each category remains challenging. This difficulty arises because clustering center estimation is a data-driven process that depends on a support set.

The second category of RetICL involves fine-tuning or retraining a retriever model to rank relevant sentences using either in-domain or out-of-domain datasets for text classification tasks. There are established methods, such as PEFT(Tunstall et al., 2022), UDR(Li et al., 2023) and Ambig-ICL(Gao et al., 2023). These methods utilize label information and feedback to optimize model parameters, highlighting the essential role of labeled data in yielding valuable insights for text classification tasks. However, they often demand substantial computational resources and considerable time to construct a retriever.

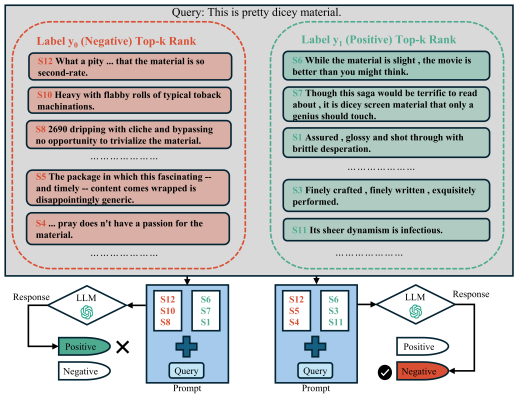


Figure 3: A comparison of the correct and incorrect demonstration combinations is presented. On the left, NwayKshot retrieves the top-*k* sentences most similar to the query from each group; however, this approach fails to classify the query correctly. In contrast, on the right, RetICL does not rely solely on proximity to the query, resulting in an accurate classification.

3 Linguistic Label Retrieval Hypothesis

Previous studies have shown that retrieving sentences closest to the query and applying a clustering-based selection method can enhance the diversity of demonstrations while mitigating the risk of misleading results due to similarity (Li et al., 2020; Luo et al., 2024). Therefore, a question arises: are the clustering centers reliable? To explore this further, we analyze the distribution of clustering centers, as shown in Figure 2 and Appendix C. By varying the proportion of fully supported data from 10% to 100%, we observe that the clustering center distribution shifts based on the number of sentences in the support set. Notably, negative-labeled clustering centers tend to be less distinct within a certain range compared to positive-labeled ones. These findings suggest that clustering center estimation is inherently data-driven and prone to bias, making it difficult to accurately identify true clustering centers. On the other hand, by analyzing failure cases, we find that, for a given query, there is often an optimal combination of demonstrations that can effectively guide LLMs to classify the query correctly. However, relying solely on the top-ranked closest demonstrations retrieved does not always yield accurate results. An example of this limitation is illustrated in Figure 3. To further investigate, we compared cases where the top-*k* closest demonstrations led to incorrect results versus cases where randomly selected demonstrations produced correct outcomes. We provide five examples of such instances in Appendix C. We found that incorrect nearest-neighbor demonstrations often exhibit an error-prone tendency, being either closer to the linguistic representation of an opposite label, closer to the center of an incorrect label cluster, or both—despite being similar to the query. Conversely, in correct combinations, the selected demonstrations exhibit a stronger alignment with the correct tendency. For example, sentences with a Negative label tend to show higher similarity to the linguistic word "Negative" and the same holds for "Positive" label. Although most correct demonstrations align closely with their respective cluster centers, we observe exceptions where a correct output contains sentences that are nearer to the center of an incorrect label cluster. Furthermore, even sentences closest to their correct cluster centers can still lead to classification errors due to inaccurate estimation of those centers.

Based on these observations, we hypothesize that the vector representations of linguistic labels should be explicitly incorporated into the retrieval process rather than relying on cluster center estimation. Compared to traditional clustering center estimation, this approach offers two advantages: (1) Independence from data Bias – The linguistic label clustering center is not data-driven, preventing bias introduced by the support set. (2) Leveraging linguistic information – Linguistic labels play a crucial role in zero-shot ICL, as LLMs rely entirely on these labels for text classification tasks.

4 Our Method: CLL-RetICL

Preliminary. Let the query set Q represent a task, where $q \in Q$ denotes a sample query for which we aim to find an answer via an LLM. In the context of RetICL, multiple demonstrations $(d_1, \dots, d_k)$ are retrieved from a support set C. Each demonstration d_i consists of a sentence and its label, $(s_i, y_i) \in C$, where y_i belongs to the label set Y.

Overview. We present CLL-RetICL, a novel RetICL approach leveraging information extraction between demonstrations and linguistic labels to predict the correct label for a given query input q_i (Wang et al., 2023). Unlike earlier methods (Liu et al., 2021; Su et al., 2022; Li et al., 2022; Milios et al., 2023) that create input-label pairs by retrieving sentences closest to a given query, CLL-RetICL selects demonstrations that balance a trade-off by augmenting the corresponding label while penalizing others.

CLL-RetICL involves three key steps, as illustrated in Figure 1: (1) Retrieving more relevant sentences by integrating sentence-query similarity with sentence-label similarity (detailed in Section 4.1), (2) Forming demonstrations by organizing the retrieved demonstrations into an N-way K-shot format (discussed in Section 4.2), and (3) Making inferences through ICL (explained in Section 4.3).

4.1 Linguistic Label Retriever

RetICL employs a retrieval mechanism to identify k examples from C that are most relevant to a given query q. This process is guided by a similarity function, sim, which quantifies the relationship between a sentence s_i and a query q. The corresponding formula is as follows:

$$score_{RetICL} = sim(q, s_i) \tag{1}$$

To build on this hypothesis, CLL-RetICL incorporates sentence-query similarity with sentence-label similarity. Rather than solely considering the similarity distance between a sentence s_i and the query q, CLL-RetICL employs the following formula:

$$\begin{aligned} score_{c\text{-}RetICL} &= sim(q, s_i) \\ &+ w_1 * log \frac{\exp^{sim(s_i, y_i)}}{\frac{1}{n-1}\sum_{y \in Y}^{y \neq y_i} \exp^{sim(s_i, y)}} \end{aligned} \tag{2}$$

where w_1 is a trade-off hyperparameter that balances the relative importance of the corresponding terms in the objective function.

CLL-RetICL considers the relationship between sentences and linguistic labels by utilizing a similarity function. It increases the score based on the similarity between a sentence and its assigned correct label (referred to as the positive label) while decreasing the score based on the similarity between the sentence and other labels (referred to as negative labels). Additionally, we propose several variations and evaluate their performance through experiments. These include Positive Label Augment (PLA), Negative Label Penalty (NLP), and Contrastive Label (CTL). The corresponding formulas are provided below:

$$score_{PLA} = sim(q, x_i) + w_1 * sim(x_i, y_i) \quad (3)$$

$$score_{NLP} = sim(q, x_i) - w_1 * \frac{1}{n-1} \sum_{y \in Y}^{y \neq y_i} sim(x_i, y) \quad (4)$$

$$\begin{aligned} score_{CTL} =& sim(q, x_i) + w_1 * sim(x_i, y_i) \\ & - w_2 * \frac{1}{n-1} \sum_{y \in Y}^{y \neq y_i} sim(x_i, y) \end{aligned} \quad (5)$$

where w_1 and w_2 are trade-off hyperparameters.

Our methods ensure that the selected sentences (1) maintain a safe distance from q to prevent the copying effect (Olsson et al., 2022; Zhang et al., 2022b), (2) incorporate the information between sentences and linguistic labels and (3) align closely with the requirements of the custom text classification task.

4.2 *N-way K-shot*

We adopt a clustering-based retrieval method, as prior research suggests that *N-way K-shot* effectively addresses the issue of homogeneity (Li and Qiu, 2023). Here, we partition all sentences into N sub-groups, aiming to cluster sentences that share the same label. Our retriever selects top K high demonstrations according to above score formula from each sub-group, resulting in a final set of $N \times K$ demonstrations.

4.3 Inference

Finally, CLL-RetICL constructs a prompt by concatenating *N-way K-shot* input-label pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k)$ for each N-way label, along with the query input q. This prompt is then fed into a LLM, which generates a prediction using $argmax_{y \in Y} P(y|prompt)$. The universal prompt template for each text classification task is outlined in Table 5 in Appendix B.

5 Experimental Analysis

5.1 Experimental Setup

We evaluate multiple LLMs to identify factors affecting classification accuracy across four tasks. Key results are summarized in the main text, with additional details presented in the Appendix D.

5.1.1 Datasets

We conduct experiments on four widely recognized text classification tasks: SST2 (Socher et al., 2013), CoLA (Warstadt et al., 2018), CARER (Saravia et al., 2018) and BBCnews (Greene and Cunningham, 2006). Similar to conventional text classification methodologies, we treat the training sets as support sets and the test sets as query sets, while disregarding development sets if they exist. The detailed data statistics are provided in Appendix A and summarized in Table 3.

5.1.2 Baselines

We compare CLL-RetICL with the zero-shot approach as well as various RetICL methods.

Zero-shot predicts $argmax_{y \in Y} P(y|q)$ without using any demonstrations (Radford et al., 2019; Brown et al., 2020). This method utilizes LLMs and linguistic label information to enhance text classification.

Z-ICL leverages physical neighbors to avoid selecting demonstrations that are overly similar to the query. Furthermore, it introduces the use of synonymous labels to mitigate the copying effect, highlighting the potential for effectively utilizing the linguistic meaning of labels (Lyu et al., 2022).

KATE employs a standard KNN approach to retrieve demonstrations, which remains the most widely used method in RetICL (Liu et al., 2021).

NwayKshot is a clustering-based retrieval method designed to tackle the challenge of homogeneity in demonstrations (Li et al., 2024).

SelfPrompt builds on NwayKshot but applies k-means clustering to identify the cluster centers. It then selects the demonstration closest to the center from each sub-group (Li et al., 2022).

Votek selects k representatives from N sub-groups through a voting mechanism to best represent the group (Su et al., 2022).

5.1.3 Experimental Details

LLMs. We conduct experiments using three LLMs: Gemini (Team et al., 2024), Llama (Dubey et al., 2024) and Mistral (Jiang et al., 2024). Specifically, we utilize fixed versions of these models, namely Gemini 1.5 Flash, Llama 3.2-90b-Vision, and Mistral Large. These recently developed models demonstrate strong performance and exceptional generalization across a variety of tasks.

LLM	Zero-shot		Z-ICL		KATE		SelfPrompt		Votek		Nwaykshot		CLL-RetICL	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
SST2														
Gemini	93.29	0.933	92.31	0.923	94.17	0.941	94.93	0.950	94.16	0.942	94.67	0.947	**95.17**	**0.952**
Llama	94.83	0.948	**96.21**	**0.962**	94.78	0.948	93.61	0.936	94.77	0.948	90.82	0.908	95.06	0.951
Mistral	90.08	0.901	90.72	0.906	93.78	0.938	94.88	0.949	94.34	0.943	94.34	0.943	**95.60**	**0.956**
Avg.	92.73	0.927	93.08	0.930	94.24	0.942	94.47	0.945	94.42	0.944	93.27	0.933	**95.28**	**0.953**
CoLA														
Gemini	68.26	0.663	60.21	0.583	70.08	0.641	80.32	0.765	81.43	0.783	82.74	0.795	**83.60**	**0.801**
Llama	61.74	0.585	52.34	0.511	68.36	0.650	71.62	0.711	61.42	0.607	74.52	0.686	**77.66**	**0.742**
Mistral	74.30	0.697	71.52	0.666	78.71	0.752	84.29	0.811	84.48	0.821	85.23	0.816	**85.52**	**0.828**
Avg.	68.10	0.648	61.36	0.587	72.38	0.681	78.74	0.762	75.78	0.737	80.83	0.766	**82.26**	**0.790**
CARER														
Gemini	59.20	0.493	65.85	0.607	70.85	0.621	61.65	0.533	59.95	0.541	66.25	0.596	**72.65**	**0.669**
Llama	56.75	0.488	65.70	0.594	61.95	0.537	57.35	0.499	59.50	0.526	64.25	0.579	**69.15**	**0.635**
Mistral	56.50	0.506	67.10	0.617	68.89	0.601	60.25	0.515	58.75	0.498	72.10	0.670	**76.85**	**0.717**
Avg.	57.48	0.495	66.22	0.606	67.23	0.586	59.75	0.516	59.40	0.521	67.53	0.615	**72.88**	**0.674**
BBCNews														
Gemini	87.00	0.869	87.70	0.872	90.99	0.909	85.30	0.850	86.20	0.858	88.60	0.884	**91.50**	**0.912**
Llama	94.89	0.948	93.43	0.933	94.70	0.946	93.60	0.935	96.00	0.960	96.10	0.960	**96.80**	**0.967**
Mistral	91.70	0.915	90.60	0.903	**92.99**	**0.929**	83.10	0.826	83.00	0.825	87.20	0.872	92.10	0.919
Avg.	91.20	0.910	90.57	0.902	92.89	0.928	87.33	0.870	88.40	0.881	90.63	0.905	**93.46**	**0.932**

Table 1: Text classification results evaluated on four datasets using three LLMs. **Bold** indicates the best result and underline indicates the result worse than the best result.

Method	Gemini		Llama		Mistral		Avg.	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1
SST2								
Baseline	94.67	0.947	90.82	0.908	94.34	0.943	93.27	0.932
PLA	**95.44**	**0.954**	93.46	0.934	94.34	0.943	94.41	0.943
NLP	95.38	**0.954**	92.31	0.922	**96.37**	**0.963**	94.68	0.946
CTL	**95.44**	**0.954**	91.65	0.916	95.11	0.951	94.06	0.940
Ours	95.17	0.952	**95.06**	**0.951**	95.60	0.956	**95.28**	**0.953**
CoLA								
Baseline	82.74	0.795	64.52	0.586	85.23	0.816	77.49	0.732
PLA	83.31	0.798	73.53	0.656	85.31	**0.832**	80.72	0.762
NLP	82.45	0.791	64.05	0.579	85.04	0.823	77.18	0.731
CTL	82.74	0.794	62.79	0.579	85.04	0.824	76.86	0.732
Ours	**83.60**	**0.801**	**77.66**	**0.742**	**85.52**	0.828	**82.26**	**0.790**
CARER								
Baseline	66.25	0.596	64.25	0.579	72.10	0.670	67.53	0.615
PLA	65.75	0.598	61.65	0.556	65.55	0.596	64.32	0.583
NLP	67.35	0.619	64.40	0.583	70.00	0.644	67.25	0.615
CTL	66.90	0.605	65.40	0.586	67.80	0.615	66.70	0.602
Ours	**72.65**	**0.669**	**69.15**	**0.635**	**76.85**	**0.717**	**72.88**	**0.673**
BBCNews								
Baseline	88.60	0.884	96.10	0.960	87.20	0.872	90.63	0.905
PLA	89.40	0.891	96.70	0.966	**89.50**	**0.895**	91.86	0.917
NLP	89.00	0.889	96.40	0.964	88.40	0.883	91.20	0.875
CTL	**90.30**	**0.901**	96.50	0.964	89.40	0.893	**92.06**	**0.919**
Ours	89.50	0.892	**96.80**	**0.967**	88.10	0.879	91.47	0.912

Table 2: A comparative analysis of various linguistic label retrieval methods across four datasets.

Similarity function. We define a similarity function, sim, as the cosine similarity between two sentence embeddings. These embeddings are generated using the all-MiniLM-L6-v2 model from the SBERT (Reimers and Gurevych, 2019).

Implementation details. For all LLMs, we use two random seeds and report the average results. Unless otherwise specified, we set the default number of demonstrations k as 3 for per class for all experiments. We adopt the typical prompt design methodology proposed by (Luo et al., 2024). To ensure accurate and consistent results in text classification tasks, we employ fixed hyperparameters for LLMs, thereby minimizing variability and limiting creative outputs. Further details are provided in Appendix B.

5.2 Experimental Results

5.2.1 Main results

Table 1 presents the results obtained using various retrieval strategies across three LLMs. The zero-shot approach, which does not rely on retrieving relevant demonstrations from the support set, leverages only the semantic understanding of labels. This strategy enables LLMs to achieve a baseline level of accuracy without additional context. Although Z-ICL mitigates the Copying Effect by leveraging physical neighbors and synonym labels, it only marginally outperforms the zero-shot baseline. However, it lags behind other methods, likely due to the inherent complexity and challenges associated with selecting appropriate synonym labels. KATE achieves better performance than zero-shot and Z-ICL by utilizing the most similar demonstrations to the query. However, it is susceptible to errors caused by misleading similarities. As a result, KATE still struggles to perform well on the CoLA and CARER datasets. To mitigate the effects of misleading similarities, NwayKshot generally outperforms KATE in most scenarios. However, as noted earlier, NwayKshot still struggles to identify an optimal combination of demonstrations. VoteK attempts to further select more effective and relevant demonstrations from the support set. However,

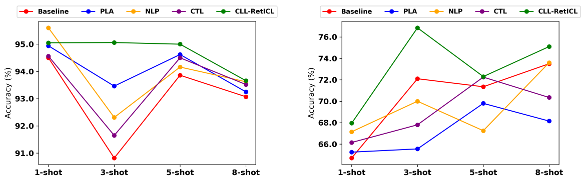


Figure 4: A comparison of the performance of various shot configurations is presented across a baseline and four linguistic label retrieval strategies. Evaluations for the SST2 task (using Llama) are on the left, while results for the CARER task (using Mistral) appear on the right.

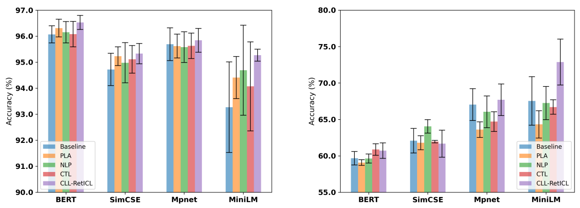


Figure 5: A comparison of the performance of various sentence embedding models is presented, with evaluations conducted on SST2 on the left and CARER on the right.

this method still fails to utilize label information effectively. On the other hand, SelfPrompt leverages label information from a distributional perspective but does not account for the linguistic meaning of the labels. While both VoteK and SelfPrompt show improvements in accuracy for certain tasks, they fall short in addressing a fundamental issue: the importance of linguistic label meaning in text classification tasks. This oversight leads to inconsistent performance and highlights their inherent weaknesses. Finally, our proposed method, CLL-RetICL, significantly outperforms all baseline approaches. On average, CLL-RetICL improves RetICL's performance by an absolute margin of 2–15% over the zero-shot strategy and by 0.57–13.48% over existing RetICL-based methods. These results demonstrate consistent performance gains across all datasets and LLMs by effectively leveraging the relationships between linguistic labels and their corresponding sentences.

Comparison to Variants of Label-Related RetICL. We use the NwayKshot method as our baseline, a retrieval-based approach that does not utilize linguistic label information. To enhance performance, we evaluate four proposed strategies that incorporate linguistic label related retrieval methods, with the results summarized in Table 2. All four strategies outperform the baseline across all datasets and LLMs, demonstrating the benefits of leveraging label information. Among these, CLL-RetICL consistently delivers the best performance, achieving an average absolute improvement of 0.8–5.3% over the NwayKshot method. While PLA, NLP, and CTL also surpass the baseline, they show minor performance drops on certain tasks. In contrast, CLL-RetICL not only outperforms these methods in most tasks but also achieves consistent gains in classification accuracy.

5.3 Ablation Study

We conduct detailed ablation studies to analyze the significance of each component in CLL-RetICL. In our ablation study, the NwayKshot approach serves as the baseline, as shown in the following tables and figures.

Effect of the number of shots. The number of shots significantly impacts the performance of LLMs. We explore experiments comparing four different shot configurations for each label class: 1-shot, 3-shot, 5-shot, and 8-shot. Figure 4 presents partial results, while the complete results are provided in Appendix D.1. The results in Figure 4 demonstrate that CLL-RetICL consistently outperforms the baseline methods across different values of k. While some alternative strategies occasionally achieve better performance than CLL-RetICL, they lack robustness and often fall short of both CLL-RetICL and the baselines. This indicates that CLL-RetICL delivers more stable performance across a range of scenarios. Based on the experimental results, we selected $k = 3$ as the hyperparameter for the number of shots, as CLL-RetICL demonstrated higher improvement with a 3-shot configuration.

Effect of sentence embedding model. Pre-trained sentence embeddings play a crucial role in ICL. The objective is to evaluate the effectiveness of the proposed methods by comparing them against four off-the-shelf sentence embedding models. Figure 5 illustrates the average performance of three LLMs across two datasets. CLL-RetICL consistently outperforms the baseline and the other three strategies across all sentence embedding models, with the exception of SimCSE (Gao et al., 2021) in the CARER dataset. We attribute the relatively lower performance of our method with SimCSE to the fact that SimCSE has already employed contrastive learning to fine-tune the pre-trained sentence embedding model. This suggests that our approach is generally more effective for pre-trained sentence embeddings that do not utilize contrastive learning strategies. Compared to other sentence embedding models, MiniLM demon-

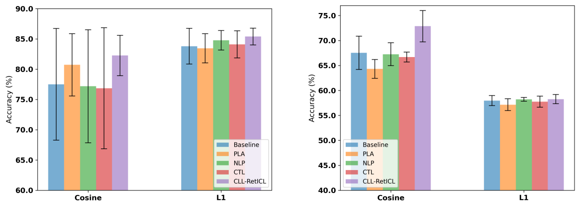


Figure 6: A comparison of the performance of various similarity functions is presented, with evaluations conducted on CoLA on the left and CARER on the right.

strates the greatest improvement over the baseline; therefore, we have chosen it as our default. Full results are presented in Appendix D.2.

Effect of similarity function. To evaluate the effect of the similarity function in our CLL-RetICL model, we compare its performance using another similarity function, L1, as described in (Winata et al., 2023). The results are presented in Figure 6 with detailed results provided in Appendix D.3.

CLL-RetICL performs effectively with both cosine and L1 similarity functions. However, experiments show that cosine similarity outperforms the L1 function, suggesting that it better leverages CLL-RetICL's potential. Consequently, we use cosine similarity as the default.

Because our proposed additional component can serve as a scoring criterion for selecting demonstrations, the question arises whether the similarity score between demonstrations and the query should be included in CLL-RetICL.

We evaluate the problem and present the results in Figure 7. Our findings indicate that the performance without the component addressing the similarity between queries and sentences is consistently lower than that of linguistically labeled RetICL. In fact, it performs even worse than the baseline. These results highlight that the similarity component between queries and sentences is an essential part of the retrieval process. Detailed results are presented in Appendix D.4.

Effect of trade-off hyperparameters.

Effect of w/o similarity between demonstration and query. We use a trade-off approach to balance the impact between sentences and their label set. Based on the results of the previous experiment, sentence-query similarity remains a crucial factor in selecting relevant demonstrations. This raises an important question: how should we trade off between the original method, which retrieves the closest demonstrations to the query, and our approach? To address this question, we evaluate the effects of various hyperparameter settings. Specifically, we focus on hyperparameters lower than 1.0, as previous research has consistently shown that closer demonstrations generally outperform those that are further away. We maintain the principle that proximity to the query remains a core factor in our approach. Based on these observations in Appendix D.5, we found that the trade-off hyperparameter has some influence on the final results. However, their impact on PLA, NLP, and CTL methods is relatively small. Interestingly, we observed that a trade-off hyperparameter value of 1.0 yields the best performance for our CLL-RetICL method. Consequently, we adopt 1.0 as the default hyperparameter.

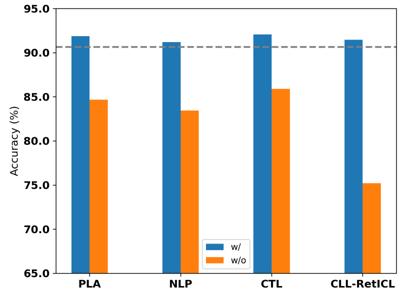


Figure 7: A comparison of the retrieval process with and without incorporating the similarity score between the query and the sentence is illustrated on BBCNews dataset. The baseline is represented by a dashed line.

6 Conclusion

This paper introduces a new paradigm Contrastive Linguistic Label Retrieval-based In-Context Learning. Unlike existing approaches that universally sample demonstrations without considering the linguistic label information, we propose a general framework for identifying more effective and relevant demonstrations. This framework enhances the capabilities of LLMs to produce more accurate text classification results. Additionally, we design a universal prompt that is adaptable to all text classification tasks. Empirical evaluation on four datasets demonstrates that CLL-RetICL significantly outperforms conventional practices in RetICL by incorporating the similarity between linguistic labels and sentences. This highlights the promising performance of CLL-RetICL and opens up several intriguing research opportunities for further methodological exploration.

7 Limitations

Requiring Semantic Labels. Our approach focuses exclusively on the semantic label text classification task. Certain text classification scenarios, however, may involve ambiguous label classes, such as class0, class1, ... Ambiguities in labeling could introduce additional challenges, and addressing these issues remains an area for future research.

Better Descriptive Labels In some classification tasks, explanations are provided for the meaning of each label. In this work, we did not utilize those explanations. Incorporating these explanations into the classification process is left as a direction for future work.

Enhance prompt clarity. In previous work, researchers observed that well-crafted prompts can lead to better results. However, in this study, we did not compare the effects of different prompt formats. Determining how to construct optimal prompts to leverage the potential of our CLL-RetICL framework fully remains an open question and is left for future exploration.

References

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

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. Llms are few-shot in-context low-resource language learners. *arXiv preprint arXiv:2403.16512*.

Youngjin Chae and Thomas Davidson. 2023. Large language models for text classification: From zero-shot learning to fine-tuning. *Open Science Foundation*.

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

Aleksandra Edwards and Jose Camacho-Collados. 2024. Language models for text classification: Is in-context learning enough? *arXiv preprint arXiv:2403.17661*.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

John Fields, Kevin Chovanec, and Praveen Madiraju. 2024. A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive? how safe? *IEEE Access*.

Lingyu Gao, Aditi Chaudhary, Krishna Srinivasan, Kazuma Hashimoto, Karthik Raman, and Michael Bendersky. 2023. Ambiguity-aware in-context learning with large language models. *arXiv preprint arXiv:2309.07900*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Derek Greene and Pádraig Cunningham. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd international conference on Machine learning*, pages 377–384.

Qi Guo, Leiyu Wang, Yidong Wang, Wei Ye, and Shikun Zhang. 2024. What makes a good order of examples in in-context learning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14892–14904.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Hongpeng Jin, Wenqi Wei, Xuyu Wang, Wenbin Zhang, and Yanzhao Wu. 2023. Rethinking learning rate tuning in the era of large language models. In *2023 IEEE 5th International Conference on Cognitive Machine Intelligence (CogMI)*, pages 112–121. IEEE.

Guozheng Li, Peng Wang, Jiajun Liu, Yikai Guo, Ke Ji, Ziyu Shang, and Zijie Xu. 2024. Meta in-context learning makes large language models better zero and few-shot relation extractors. *arXiv preprint arXiv:2404.17807*.

Jing Li, Billy Chiu, Shanshan Feng, and Hao Wang. 2020. Few-shot named entity recognition via meta-learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(9):4245–4256.

Junlong Li, Jinyuan Wang, Zhuosheng Zhang, and Hai Zhao. 2022. Self-prompting large language models for zero-shot open-domain qa. *arXiv preprint arXiv:2212.08635*.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023. Unified demonstration retriever for in-context learning. *arXiv preprint arXiv:2305.04320*.

Xiaonan Li and Xipeng Qiu. 2023. Finding support examples for in-context learning. *arXiv preprint arXiv:2302.13539*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Lefteris Loukas, Ilias Stogiannidis, Prodromos Malakasiotis, and Stavros Vassos. 2023. Breaking the bank with chatgpt: few-shot text classification for finance. *arXiv preprint arXiv:2308.14634*.

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

Xinxi Lyu, Sewon Min, Iz Beltagy, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. Z-icl: Zero-shot in-context learning with pseudo-demonstrations. *arXiv preprint arXiv:2212.09865*.

Costas Mavromatis, Balasubramaniam Srinivasan, Zhengyuan Shen, Jiani Zhang, Huzefa Rangwala, Christos Faloutsos, and George Karypis. 2023. Which examples to annotate for in-context learning? towards effective and efficient selection. *arXiv preprint arXiv:2310.20046*.

Aristides Milios, Siva Reddy, and Dzmitry Bahdanau. 2023. In-context learning for text classification with many labels. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 173–184.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Branislav Pecher, Ivan Srba, Maria Bielikova, and Joaquin Vanschoren. 2024. Automatic combination of sample selection strategies for few-shot learning. *arXiv preprint arXiv:2402.03038*.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Hamidreza Rouzegar and Masoud Makrehchi. 2024. Enhancing text classification through llm-driven active learning and human annotation. *arXiv preprint arXiv:2406.12114*.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin 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.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, et al. 2022. Selective annotation makes language models better few-shot learners. *arXiv preprint arXiv:2209.01975*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*.

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. *arXiv preprint arXiv:2305.14160*.

Zhiqiang Wang, Yiran Pang, and Yanbin Lin. 2024. Smart expert system: Large language models as text classifiers. *arXiv preprint arXiv:2405.10523*.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Genta Indra Winata, Liang-Kang Huang, Soumya Vadlamannati, and Yash Chandarana. 2023. Multilingual few-shot learning via language model retrieval. *arXiv preprint arXiv:2306.10964*.

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2022. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. *arXiv preprint arXiv:2212.10375*.

Hao Yu, Zachary Yang, Kellin Pelrine, Jean Francois Godbout, and Reihaneh Rabbany. 2023. Open, closed, or small language models for text classification? *arXiv preprint arXiv:2308.10092*.

Yazhou Zhang, Mengyao Wang, Chenyu Ren, Qiuchi Li, Prayag Tiwari, Benyou Wang, and Jing Qin. 2024. Pushing the limit of llm capacity for text classification. *arXiv preprint arXiv:2402.07470*.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022a. Active example selection for in-context learning. *arXiv preprint arXiv:2211.04486*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022b. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.