Rethinking Label Consistency of In-Context Learning: An Implicit Transductive Label Propagation Perspective

Haoyang Chen[1,2], **Richong Zhang**[2,3 *], **Junfan Chen**[1,2]
[1]School of Software, Beihang University, Beijing, China
[2]CCSE, School of Computer Science and Engineering, Beihang University, Beijing, China
[3]Zhongguancun Laboratory, Beijing, China
Cauchy20011214@buaa.edu.cn,{zhangrc, chenjf}@act.buaa.edu.cn

Abstract

Large language models (LLMs) perform in-context learning (ICL) with minimal supervised examples, which benefits various natural language processing (NLP) tasks. One of the critical research focus is the selection of prompt demonstrations. Current approaches typically employ retrieval models to select the top-K most semantically similar examples as demonstrations. However, we argue that existing methods are limited since the label consistency is not guaranteed during demonstration selection. Our cognition derives from the Bayesian view of ICL and our rethinking of ICL from the transductive label propagation perspective. We treat ICL as a transductive learning method and incorporate latent concepts from Bayesian view and deduce that similar demonstrations guide the concepts of query, with consistent labels serving as estimates. Based on this understanding, we establish a label propagation framework to link label consistency with propagation error bounds. To model label consistency, we propose a data synthesis method, leveraging both semantic and label information, and use TopK sampling with Synthetic Data (TopK-SD) to acquire demonstrations with consistent labels. TopK-SD outperforms original TopK sampling on multiple benchmarks. Our work provides a new perspective for understanding the working mechanisms within ICL.

Code — https://github.com/Cauchy2001/TopK_SD

Introduction

Large language models (LLMs) (Min et al. 2023; Liu et al. 2023) demonstrate remarkable generative capabilities and achieve superior performance across a wide spectrum of traditional NLP tasks (Mann et al. 2020), including sentiment analysis, text classification, and machine translation. The efficacy of these models is significantly enhanced by in-context learning (ICL) (Dong et al. 2022; Bai et al. 2024), which enables task execution through in-context inference with minimal supervised prompts. Compared to conventional fine-tuning approaches (Devlin 2018), ICL accomplishes comparable task performance while substantially reducing data annotation costs (Lester, Al-Rfou, and Constant 2021), primarily by eliminating the need for extensive parameter updates. For instance, by providing only 4-8 labeled examples as demonstrations, ICL allows models like GPT (Radford et al. 2019; Floridi and Chiriatti 2020; Achiam et al. 2023) and Llama (Touvron et al. 2023a,b; Dubey et al. 2024) to generalize to unseen tasks without explicit parameter updates and accomplishes comparable task performance while substantially reducing data annotation costs (Lester, Al-Rfou, and Constant 2021). For instance, by providing only 4-8 labeled examples as demonstrations, ICL allows models like GPT (Radford et al. 2019; Floridi and Chiriatti 2020; Achiam et al. 2023) and Llama (Touvron et al. 2023a,b; Dubey et al. 2024) to generalize to unseen tasks without explicit parameter updates and accomplishes comparable task performance while substantially reducing data annotation costs (Lester, Al-Rfou, and Constant 2021).

Query	Demonstrations	Consistency	Accuracy	Selection
P	PPPPPPNN	75%	↑	✓
	PPNNNNNN	25%	↓	✗

Table 1: On SST-2, for a query with the label "positive", our method selects demonstrations with the same label to improve ICL performance.

The performance of ICL critically hinges on demonstration selection (Sorensen et al. 2022; Zhang, Feng, and Tan 2022; Li and Qiu 2023), their ordering (Lu et al. 2022; Liu et al. 2024), and formatting (Kim et al. 2022; Hao et al. 2022); hence, selecting appropriate demonstrations for ICL in large models is paramount. Consequently, the research community has introduced a wealth of methods for selecting demonstrations while also selecting demonstrations at the corpus or instance level with key factors such as diversity and similarity (Luo et al. 2024; Dong et al. 2022). A common retrieval approach generates embeddings (Reimers 2019), calculates Euclidean distance or cosine similarity, and selects the top-K most similar demonstrations (Liu et al. 2021). The nearest-neighbor examples' labels enhance the generative model's final prediction. Subsequent work has extended TopK method, including MDL (Wu et al. 2022) and ConE (Peng et al. 2024). The effectiveness of these methods is dependent on TopK retrieval.

To understand the mechanisms of ICL, some studies have focused on influential factors, including data distri-

*Corresponding Author

bution (Wies, Levine, and Shashua 2024), training process (Ding et al. 2023), and diversity (An et al. 2023), among others. Currently, there is no clear explanation for why ICL works. Existing principled explanations for ICL include the Bayesian veiw (Xie et al. 2021; Wies, Levine, and Shashua 2023), which suggests that demonstrations of ICL are used to activate the latent concepts learned during pretraining, and the gradient descent view (Dai et al. 2022; von Oswald et al. 2023), which posits that the effects of ICL are equivalent to those of fine-tuning. The work (Wang et al. 2023) interpret ICL's working principle via transformers' attention mechanism, arguing that label words anchor demonstration information to form final predictions.

Based on the existing research on the interpretability of ICL, most studies consider that ICL is not conventional learning (Kossen, Gal, and Rainforth 2023). However, if we rethink from the perspective of learning paradigms, ICL should be considered transductive learning, rather than inductive learning. From Bayesian view, we introduce the latent concepts into the transductive optimization objective of ICL. Under the assumptions of Bayesian inference framework, we deduce the working principle of latent concepts in ICL. Latent concepts help map sentences to labels. Moreover, we derive that similar demonstrations can effectively guide the latent concepts corresponding to the query, and consistent labels can estimate whether the guided concepts related to the query. Therefore, we establish a transductive label propagation framework to explain how demonstrations in ICL propagate concepts to query and use label consistency as a lower bound for evaluating propagation error. The derivation reveals why label consistency is important.

Since we do not have sufficient prior knowledge of the query labels during demonstration selection, and existing methods do not incorporate labels into the selection, we designed a data synthesis method to synthesize new embeddings with label information. We sample from the synthesized embeddings with TopK method called TopK with Synthetic Data (TopK-SD). This method can better select demonstrations with consistent labels. Compared with TopK with original embeddings, TopK-SD shows significant improvement on multiple benchmarks. The reason is that TopK-SD maintains semantic similarity and greatly improves label consistency. Some methods based on TopK also conducted ablation studies, replacing the TopK module with the TopK-SD module, and it demonstrates the effectiveness of TopK-SD. It has verified the importance of label consistency and corroborated derivation of Bayesian transduction.

Briefly, the contribution of this study can be summarized as follows: (1) We rethink ICL as a form of transductive learning, derive the roles of demonstrations and labels through Bayesian view, model transductive label propagation framework, and highlight the importance of label consistency. (2) We propose a data synthesis method using semantic and label information, apply TopK-SD sampling to get consistent-label demonstrations. Experiments show that TopK-SD outperforms TopK with higher accuracy.

Related Works

In-Context Learning As an emerging learning paradigm, In-Context Learning (ICL) (Dong et al. 2022; Bai et al. 2024) has attracted significant attention in the field of natural language processing (NLP) in recent years (Mann et al. 2020). ICL enables models to perform inference directly using a small number of labeled examples (i.e., demonstrations) provided in the context, without requiring large-scale parameter updates (Devlin 2018). Compared to traditional fine-tuning methods, ICL achieves comparable task performance while substantially reducing the cost of data annotation (Lester, Al-Rfou, and Constant 2021). It makes ICL particularly advantageous in low-resource scenarios.

Interpretability of ICL Despite ICL's remarkable performance in practice, its underlying mechanisms remain unclear. Some studies have approached ICL from Bayesian view, suggesting that demonstrations activate latent concepts learned during pretraining (Xie et al. 2021; Wies, Levine, and Shashua 2023). Others have explained ICL through the view of gradient descent, positing that its effects are equivalent to fine-tuning (Dai et al. 2022; von Oswald et al. 2023). The work (Wang et al. 2023) have explored the attention mechanisms of transformers, arguing that labels aggregate demonstration information for label prediction. While these studies offer different perspectives on understanding mechanisms, many questions still remain open for investigation.

Demonstration Selection of ICL ICL critically hinges on the demonstration selection (Sorensen et al. 2022), their ordering (Lu et al. 2022), and formatting (Kim et al. 2022). Base above factors, demonstration selection methods can generally be divided into corpus-level and instance-level approaches. Corpus-level methods include Votek (Su et al. 2022),Q-Learning (Zhang, Feng, and Tan 2022), and others. However, corpus-level methods often perform worse than instance-level methods, which is why most researchers focus more on the latter. Instance-level methods include TopK, which select the top-K most similar demonstrations by calculating cosine similarity (Liu et al. 2021). Subsequent work has extended the TopK method, such as MDL (Wu et al. 2022) and ConE (Peng et al. 2024). These methods significantly improve ICL's performance by optimizing the retrieval process and demonstration selection strategies. However, their effectiveness heavily relies on TopK retrieval.

Problem Formulation

In-context learning is a training-free paradigm that enables LMs to learn downstream tasks using only a few demonstrations. Therefore, ICL can also be regarded as few-shot learning in such scenarios (Mann et al. 2020). Formally, for an input query x, there is a label candidate set $\mathcal{Y} = \{y_1, \cdots, y_M\}$. The likelihood of a candidate label y is derived from a scoring function f computed by LM within the context $C = \{x_1, y_1, \ldots, x_k, y_k\}$ (Dong et al. 2022).

$$P(y \mid x) \triangleq f_{LM}(y \mid C, x) = f_{LM}(y \mid X, Y, x) \quad (1)$$

Context C is a prompt constructed via k-shot learning, where each pair (x_i, y_i) represents the sentence and label of the i-th selected example. The sentences of demonstrations

in context is $X = \{x_i\}_{i=1}^{k}$ and the labels $Y = \{y_i\}_{i=1}^{k}$. Obviously, $C = \{X, Y\}$. The final predicted label $\hat{y}$ is the label with highest probability.

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} P\left(y \mid x\right) \tag{2}$$

The task is to select demonstrations $\{(x_i, y_i)\}_{i=1}^{k}$ that enhances ICL and improves the prediction performance.

Methodology

In this section, we rethink ICL from the perspective of transductive learning and argue that it can be modeled as label propagation. We explain the role of labels within ICL. Building on these insights, we introduce a data-synthesis method that leverages both demonstration semantics and label information to synthesis new embeddings and employ TopK for sampling on them. This method is referred to as TopK with Synthesis Data (TopK-SD).

Rethink Learning Paradigms of ICL

ICL, as a capability of LMs, can learn from only a few examples. While the learning mechanism behind it is still not fully understood by researchers, many views have been proposed to view ICL, such as the Bayesian view (Xie et al. 2021; Wies, Levine, and Shashua 2023) and gradient descent view (Dai et al. 2022; von Oswald et al. 2023). In existing research work, it is considered that ICL is not conventional learning (Kossen, Gal, and Rainforth 2023). From the perspective of learning paradigms, we update the view that ICL is regarded as transductive learning, rather than inductive learning in traditional sense.

Transductive Learning Transductive learning uses the training set $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{n}$ and the unlabeled test set $\mathcal{X}_{\text{test}} = \{x_j\}_{j=1}^{m}$ to directly infer test labels without learning a generalizable function. It focuses on local optimization rather than generalization, unlike inductive learning which aims to predict unseen data. Mathematically:

$$\hat{y}_{1:m} = \arg\max_{y_{1:m} \in \mathcal{Y}^m} P\left(y_{1:m} \mid \mathcal{D}_{\text{train}}, \mathcal{X}_{\text{test}}\right) \tag{3}$$

Revisiting Equation 1, it's significant that the goal of demonstration selection should be to optimize the function f_{LM}. And while $m = 1$, the form of Equation 4 is identical to that of the function f_{LM}. Accordingly, we provide the optimization of ICL with the definition of transductive learning.

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} P\left(y \mid X, Y, x\right) \tag{4}$$

Reconsidering the nature of ICL, it should be fundamentally regarded as a transductive learning paradigm.

Bayesian Transduction

We derive the transductive learning objective of ICL. Express the transductive learning optimization Function 4 in terms of conditional probabilities as Equation 5. It shows that $P\left(y \mid X, Y, x\right)$ is proportional to $P\left(Y, y \mid X, x\right)$.

$$P\left(y \mid X, Y, x\right) = \frac{P\left(Y, y \mid X, x\right)}{P\left(Y \mid X, x\right)} \propto P\left(Y, y \mid X, x\right) \tag{5}$$

Combining the Bayesian inference framework hypothesis, make further deduction for ICL transductive learning.

Bayesian Inference Framework In the Bayesian framework, ICL is explained as implicit Bayesian inference shown as Equation 6 (Xie et al. 2021).

$$\begin{aligned} P\left(y \mid X, Y, x\right) &= \int_{\theta} p(y \mid \theta, X, Y, x)\, p(\theta \mid X, Y, x)\, \mathrm{d}\theta \\ &= \int_{\theta_0, \theta_1, \ldots, \theta_k} p(y \mid x_1, y_1, \theta_1, \ldots, x_k, y_k, \theta_k, x, \theta_0) \\ &\quad \cdot p(\theta_0, \theta_1, , \ldots, \theta_k \mid X, Y, x)\, \mathrm{d}\theta_0\, \mathrm{d}\theta_1\, \cdots\, \mathrm{d}\theta_k \end{aligned} \tag{6}$$

The role of demonstrations is to activate the latent concept θ which characterizes the relationship between sentences and labels and helps LM to better understand the task intentions. Further research indicates that there is a latent concept θ_i related to i^{th} demonstration, θ_0 is raleted to query x. Moreover, θ represents the shared concept of all θ_i and θ_0 and characterizes the overall mapping relationship in ICL. The mechanism of ICL is to activate the knowledge concepts learned during the pre-training stage based on the demonstrations in the context, and to generate answers y for the problem x according to these concepts.

Incorporating the Bayesian inference framework, latent concept θ is introduced to determine the mapping $f_\theta : X \to Y$. The mapping f_θ is also applicable from x to y, which is referred to as ICL. So $P\left(Y, y \mid X, x\right)$ characterizes such a mapping f_θ. Applying the Bayesian inference framework to $P\left(Y, y \mid X, x\right)$, Equation 7 can be described by introducing the latent concept θ as follows:

$$P\left(Y, y \mid X, x\right) = \int_{\theta} p\left(Y, y \mid \theta, X, x\right) p\left(\theta \mid X, x\right)\, \mathrm{d}\theta \tag{7}$$

Latent Optimization Objective Substituting Equation 7 into Equation 5 yields Equation 8. Since the denominator is a normalization constant, while the numerator represents the exploration of the latent concept θ, the fundamental objective of ICL should be to find suitable θ related to query x and its label y.

$$P\left(y \mid X, Y, x\right) = \frac{\int_{\theta} p\left(Y, y \mid \theta, X, x\right) p\left(\theta \mid X, x\right)\, \mathrm{d}\theta}{\int_{\theta} \int_{y} p\left(Y, y \mid \theta, X, x\right) p\left(\theta \mid X, x\right)\, \mathrm{d}y \mathrm{d}\theta} \tag{8}$$

Label Propagation Framework

Based on Bayesian transduction, the learning mechanism of ICis elaborated and a label propagation framework is established to explain ICL.

Bayesian Belief for Concept θ From Bayesian view, the selection of demonstrations X is an update to the concept θ. Equation 9 express $p\left(\theta \mid X, x\right)$ in the form of Bayes' theorem. Here, it reveals that LM has a prior probability $p\left(\theta \mid x\right)$ regarding θ while x is known. As X is introduced as evidence for the likelihood estimation of θ, LM updates the posterior probability $p\left(\theta \mid X, x\right)$. Since query x is known, and demonstrations X need to be selected, if X is highly similar to x, it can help the model better estimate the shared

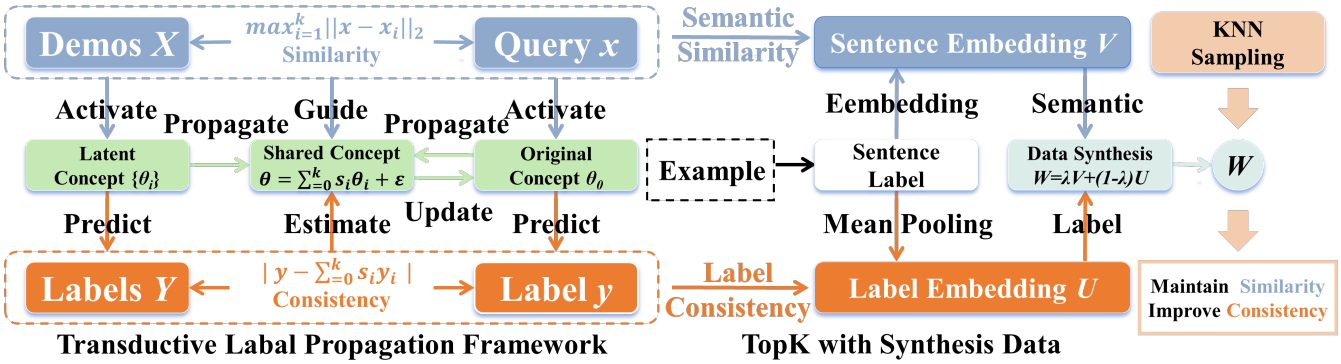


Figure 1: TopK sampling is performed on the synthesized embeddings with semantic and label information.

concept θ related to x. This explains from Bayesian belief why similarity is important for ICL (Liu et al. 2021).

$$p(\theta \mid X, x) = \frac{p(X \mid \theta, x)\, p(\theta \mid x)}{p(X \mid x)} \tag{9}$$

In addition, $p(Y, y \mid \theta, X, x)$ in Equation 7 can be regarded as the likelihood estimate of the shared concept θ given demonstrations' labels Y and query's label y, while X and x are known. When Y and y are as consistent as possible, $p(Y, y \mid \theta, X, x)$ can offer a better estimate of θ. If the consistency between Y and y is higher, it can better reflect that the shared concept θ is more strongly associated with x. From this view, the goal of ICL should be intrinsically linked to the concept associated with the query,and we introduce label propagation to explain the update of concepts.

Transductive Label Propagation Transductive label propagation is a semi-supervised learning approach that uses the labels of labeled data to infer the labels of unlabeled data. Based on Equation 7, the similarity of demonstrations for guiding latent concepts and the consistency of labels for estimating latent concepts align with the smoothness assumption of label propagation: similar data tend to have the same label. Base on this, we propose the view that ICL is a label propagation framework to update the concept θ.

Update of Concept θ The feature propagation in label propagation, as shown in Equation 10, represents the update of original concept θ_0 of x with the incorporation of $\{\theta_i\}_{i=1}^{k}$, resulting in shared concept θ. s_i represents propagation coefficient, which is related to factors such as the semantic similarity between x and x_i, the position of the x_i in order of the context and other relevant factors. Therefore, Equation 10 can be regarded as further updating information of input query by incorporating in-context demonstrations.

There are apparently errors, so it is necessary to introduce an error term into the feature propagation Equation 10. ϵ denotes the propagation error within the LM in ICL.

$$\theta = \sum_{i=0}^{k} s_i\theta_i + \boldsymbol{\epsilon} \tag{10}$$

The update of θ also signifies the update of the label y, in an effort to characterize the propagation error using y.

L-Lipschitz Constraint The mapping of ICL $f_\theta : X \to Y$ is smooth and satisfies the L-Lipschitz constraint as follow:

$$|\mathbf{y}_1 - \mathbf{y}_2| = |f_\theta(\mathbf{x}_1) - f_\theta(\mathbf{x}_2)| \leq L\,\|\mathbf{x}_1 - \mathbf{x}_2\|_2 \tag{11}$$

The propagation error can be characterized by the constraint.

$$\left|\mathbf{y} - \sum_{i=0}^{k} s_i\mathbf{y}_i\right| \leq L\|\boldsymbol{\epsilon}\|_2 + o\left(\max_{i=1}^{k}\|\mathbf{x} - \mathbf{x}_i\|_2\right) \tag{12}$$

Estimation of Propagation Error The L-Lipschitz constraint provides a lower bound for the error estimate. $L\|\boldsymbol{\epsilon}\|_2$ represents the noise caused by propagation error ϵ. Noise $L\|\boldsymbol{\epsilon}\|_2$ is proportional to the magnitude of error ϵ, and physical significance can be understood as an amplification of error ϵ. $o(\max_{i=1}^{k}\|\mathbf{x} - \mathbf{x}_i\|_2)$ represents the asymptotic upper bound of the maximum difference between x and x_i, reflecting semantic similarity. In addition, $|\mathbf{y} - \sum_{i=0}^{k} s_i\mathbf{y}_i|$ describes the discrepancy between the label obtained through propagation and the ground-truth label. It is worth noting that y_0 describes the initial label assigned to x by LM.

Noise's lower bound is determined by both semantic similarity and label consistency. Combining Equation 7 and Equation 12, the similarity demonstrations are intended to reduce the propagation error, while label consistency is aimed at estimating the error. This provides the two guiding principles for demonstration selection. The process should take into account both of these factors. In other words, base on the label propagation framework, ICL should focus more on demonstrations, which should have a high semantic similarity to the query and share the same label.

Label Consistency of Data Synthesis

In ICL, no substantial prior knowledge is available for the labels of input queries and most methods do not consider the utilization of label information of demonstrations. To address this, we design a data synthesis method based on semantic similarity measured by embedding models, which facilitates sampling demonstrations with high semantic similarity and the same label from the synthesized embeddings.

Embedding Interpolation To obtain demonstration with consistent label, we employ an interpolation method to force

vectors of the same category to converge towards a central vector. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ denote candidate demonstration set and $\mathcal{Y}$ as the label set. $\mathbf{V}_i$ denote the embedding of the i^{th} demonstration's sentence. $\mathbf{U}_k$ denote the center vector for category k, which is defined as the mean-pooling of all embedding vectors belonging to category k as follow:

$$\mathbf{U}_k = \frac{\sum_{i=1}^{n} \mathbf{V}_i \cdot \mathbb{I}\left[y_i = k\right]}{\sum_{i=1}^{n} \mathbb{I}\left[y_i = k\right]} \tag{13}$$

$\mathbf{U}_k$ is the semantic center of category k, that is, the label embedding representing the characteristics of the category. $\mathbb{I}[A]$ is an indicator function such that when Boolean expression A is true, $\mathbb{I} = 1$; otherwise, $\mathbb{I} = 0$. We interpolate to synthesize embedding using sentence and label embedding.

$$\mathbf{W}_i = \lambda \mathbf{V}_i + (1 - \lambda)\, \mathbf{U}_{y_i} \tag{14}$$

The synthesized embeddings can better conform to the smoothness theory. In this way, our method can utilize label information to enhance the original data features.

For each input query, interpolation is performed following equation 14. Since query lacks label information, we select an appropriate estimation equation to minimize the error between the estimated value and ground-truth value. Define $\mathbf{U}$ as the reference vector, which be obtained through the mean-pooling of all $\mathbf{U}_k$.

$$\mathbf{U} = \frac{\sum_{k \in \mathcal{Y}} \mathbf{U}_k}{|\mathcal{Y}|} \tag{15}$$

Selecting reference vectors without any semantic information can reduce bias towards any specific category. We perform the following interpolation transformation based on $\mathbf{U}$.

$$\mathbf{W}_i = \lambda \mathbf{V}_i + (1 - \lambda)\, \mathbf{U} \tag{16}$$

By employing such data synthesis method, the synthetic embeddings of demonstrations from the same category can achieve better clustering performance, and the nature of the synthetic data better conforms to the smoothness theory.

Demonstration Selection TopK method samples K-Nearest Neighbors (KNN) on the original embeddings to serve as demonstrations for LM to perform ICL. The advantage of this approach is that it allows for the selection of the K most semantically similar samples as demonstrations, but the label consistency between demonstrations and the query is not very good. We propose a new method, TopK with data synthesis (TopK-SD), which selects KNN from the synthesized embeddings to serve as demonstrations for ICL. This method not only maintains semantic similarity but also improves label consistency. The resulting labels from sampling exhibit significantly improved consistency. The propagation error of ICL can be better reflected through the labels of the selected demonstrations.

Experiment

Datasets and Experiment Setting

Dataset We use widely-used datasets, comprising six classification tasks and three natural language inference (NLI) tasks (Sun et al. 2023). The data includes SST-2 (Socher et al. 2013), SST-5(Socher et al. 2013), AGNews (Zhang, Zhao, and LeCun 2015), TREC (Voorhees and Tice 2000), CR (Hu and Liu 2004), Subj (Pang and Lee 2004), MNLI (Williams, Nangia, and Bowman 2017), QNLI (Rajpurkar 2016) and RTE (Rajpurkar 2016).

Experiment Setting We employ a sentence-transformer models as the retrieval model all-roberta-large-v1 (Reimers 2019), along with the inference models GPT-j-6b (Wang and Komatsuzaki 2021), LlaMA2-7b (Touvron et al. 2023a), deepseek-llm-7b-base (Bi et al. 2024) and LlaMA3-8b (Grattafiori et al. 2024). We performed 8-shot experiments using A800 with this training set for demonstration selection, randomly sampling 1,000 test instances thrice to compute the average accuracy for three times.

Baseline Methods

We use TopK-SD method to generate embedding interpolations and sample KNN to select demonstrations and compare it with directly using prompt templates, evaluating their impact on ICL.

Prompting Prompting without in-context demonstrations.

Random Selecting demonstrations randomly.

Votek (Su et al. 2022) Selection methods that are enlightening for ICL at the corpus level with voting method.

BM25 (Robertson, Zaragoza et al. 2009) Using term frequency and inverse document frequency to evaluate document-query relevance and document length.

TopK (Liu et al. 2021; Gao, Fisch, and Chen 2020) Choosing the top-K most semantically similar demonstrations based on the embedding similarity.

In addition, for some methods based on the TopK approach, we replace the TopK module of the method with our TopK-SD method and verify the effectiveness of TopK-SD.

MDL (Wu et al. 2022) Adopting a framework that first ranks combinations of demonstrations based on the Minimum Description Length (MDL) principle, and chooses the combination with the best MDL scores.

ConE (Peng et al. 2024) Using a framework that ranks demonstrations by Conditional Entropy (ConE), and chooses the top k with the best ConE scores.

DPP (Ye et al. 2023) A method using Determinantal Point Processes (DPP) to model demonstration-input interactions and optimize via contrastive learning for selection.

Text Classification Results

Main Results In the experiments on LLaMA3, TopK-SD tested ten different values of λ (i.e., $\lambda = 0.0, 0.1, \ldots, 0.9$), and the value that achieved the highest accuracy is selected as the evaluation result. In the experiments on GPT-J, LLaMA2, and DeepSeek, the parameter λ for TopK-SD is set to 0.7. The results of the TopK-SD method compared to the baseline method are shown in Table 2. It demonstrates that in the experiments on the LLaMA3 model, selecting an appropriate λ for data synthesis enables TopK-SD to achieve the best performance on the validation set across different benchmarks. When λ is set to 0.9 on RTE, 0.8 on SST-2 and MNLI, 0.7 on SST-5 and Subj, 0.6 on TREC, 0.5 on AGNews, and 0.3 on QNLI. Compared with TopK, the average accuracy of TopK-SD has increased by 1.4%. Even on the

Model	Method	SST-2	SST-5	AGNews	TREC	CR	Subj	MNLI	QNLI	RTE	Avg.	Δ
LlaMA3	Prompt	86.3	26.8	70.5	47.0	83.8	61.4	47.7	50.0	65.0	59.8	+21.3
	Random	95.7	47.4	82.4	66.4	89.4	90.9	56.8	58.4	72.9	73.4	+7.8
	Votek	96.5	50.3	85.8	64.2	83.2	94.3	63.0	53.1	70.4	73.4	+7.7
	BM25	95.5	47.4	93.0	**93.2**	91.8	95.7	62.6	61.8	**74.4**	79.5	+1.6
	TopK	96.0	52.7	93.5	91.0	91.5	**96.7**	63.4	62.7	70.4	79.8	+1.4
	TopK-SD	**96.5**	**53.7**	**93.9**	92.6	**92.8**	96.3	**64.7**	**67.0**	72.6	**81.1**	-
GPT-j	Random	91.2	41.6	70.8	49.6	81.4	72.0	39.7	51.4	55.2	61.4	+12.6
	TopK	93.8	49.2	90.1	88.6	89.6	92.0	43.0	52.8	53.8	72.5	+1.5
	TopK-SD	**94.6**	**50.1**	**90.6**	**91.6**	**90.7**	**92.3**	**44.2**	**55.3**	**57.0**	**74.0**	-
LlaMA2	Random	92.1	46.1	82.6	63.2	90.7	47.1	50.3	56.5	69.7	66.5	+10.6
	TopK	94.9	53.6	91.3	89.0	93.4	82.5	**52.2**	59.5	**67.1**	75.9	+1.2
	TopK-SD	**95.3**	**53.7**	**92.1**	**92.2**	**93.6**	**85.9**	51.3	**63.1**	66.8	**77.1**	-
DeepSeek	Random	95.5	44.4	80.9	61.2	93.1	74.9	45.4	54.2	59.9	67.7	+9.2
	TopK	95.8	51.8	92.1	87.8	93.4	93.5	**50.6**	58.4	**66.1**	76.6	+0.3
	TopK-SD	**95.8**	**52.6**	**92.1**	**90.4**	**93.9**	**94.1**	48.4	**61.1**	63.5	**76.9**	-

Table 2: TopK-SD is compared with various baseline methods across four models and nine datasets. Best results are **bold**.

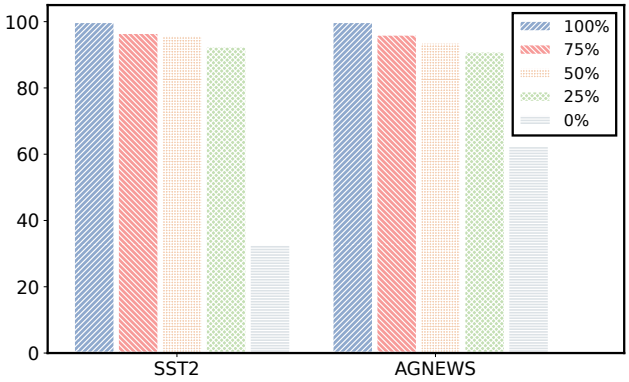


Figure 2: Investigate the relationship between label consistency and ICL accuracy on different datasets.

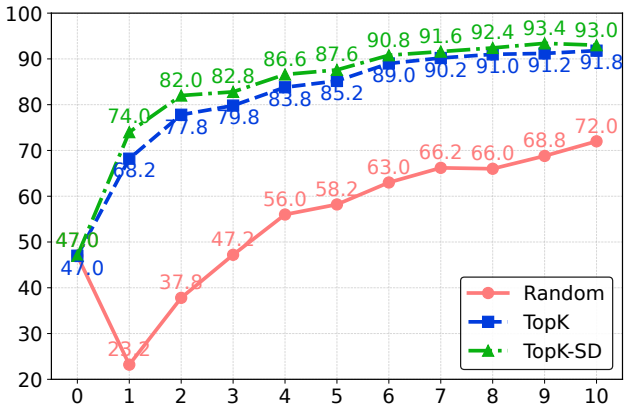


Figure 3: Using different numbers of demonstrations, compare TopK-SD ($\lambda = 0.7$) and other methods on TREC.

other three models, where λ is set to only 0.7, most benchmarks and the average accuracy are improved. On three different models, the average accuracy of TopK-SD was improved by 1.5%, 1.2%, and 0.3% compared to TopK, respectively. If appropriate values of λ are selected for the datasets, the accuracy of TopK-SD can be further improved.

Ablation Study Results In MDL, ConE, and DPP methods, retrieval involves two stages: Stage 1 uses TopK or TopK-SD ($\lambda = 0.7$) to narrow down to 30 candidates, and Stage 2 employs three strategies to select 8 final demonstrations. We replaced TopK with TopK-SD and tested on LLaMA3. Results in Table 3 show that TopK-SD outperformed TopK with average accuracy gains of 0.9%, 0.6%, and 0.4% across strategies, though the improvement is not significant, possibly due to the effective strategy in Stage 2 and the only choice of $\lambda = 0.7$. Even so, it demonstrates that the narrowed candidate set by TopK-SD enhances label consistency, thereby improving performance across various scenarios.

Analysis of Label Consistency in ICL

Label Consistency Since the propagation coefficient s_i is hard to characterize, label consistency is therefore quantified by the ratio of demonstrations with the same label. As depicted in Figure 2, the experimental outcomes demonstrate a positive correlation between label consistency and accuracy: as label consistency increases, the accuracy of ICL also improves. Furthermore, when no demonstration labels are consistent with y, it implies a lack of guidance on how to extract the latent concept θ that maps x to y. Consequently, the accuracy significantly decreases. It is evident that label consistency contributes to ICL.

Demonstrations with Consistent Label To ensure that the improvement in ICL due to enhanced label consistency is not from biased distribution of demonstration labels but from sufficient demonstrations providing necessary information for inference, we conducted ablation studies by re-

Stage1	Stage2	SST-2	SST-5	AGNews	TREC	CR	Subj	MNLI	QNLI	RTE	Avg.	Δ
TopK	MDL	96.7	54.1	87.4	87.0	93.6	93.6	**59.4**	63.7	62.5	77.6	+0.9
TopK-SD		**97.2**	**55.9**	**87.8**	**90.2**	**93.9**	**94.6**	58.8	**65.4**	**62.5**	**78.5**	-
TopK	ConE	95.8	46.0	**92.2**	93.6	91.5	**96.8**	**58.9**	**66.4**	67.1	78.7	+0.4
TopK-SD		**96.1**	**46.1**	91.9	**94.2**	**92.6**	96.5	58.7	65.8	**70.4**	**79.1**	-
TopK	DPP	95.4	49.9	**92.0**	86.0	91.0	94.2	53.2	57.6	**69.7**	76.5	+0.6
TopK-SD		**96.1**	**50.9**	91.7	**86.8**	**92.3**	**94.3**	**54.8**	**59.5**	67.5	**77.1**	-

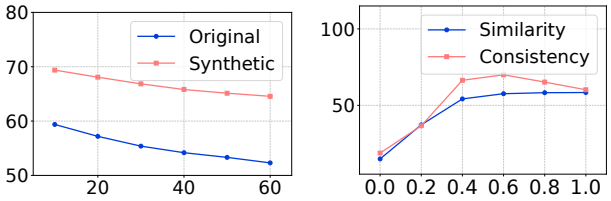
Table 3: Replace the TopK module in Stage 1 of three methods with TopK-SD and conduct comparisons. Best results are **bold**.

λ	Method	SST-2	SST-5	AGNews	TREC
0.0	Vote	49.5	17.6	25.0	18.8
	ICL	**68.4**	**18.4**	**25.4**	**18.8**
0.2	Vote	90.6	17.7	42.4	**37.6**
	ICL	**92.6**	**18.6**	**47.8**	37.2
0.4	Vote	91.5	33.4	82.7	75.0
	ICL	**95.6**	**43.3**	**90.4**	**89.2**
0.6	Vote	90.6	43.3	91.3	84.6
	ICL	**96.1**	**51.9**	**93.3**	**92.6**
0.8	Vote	89.0	45.1	91.8	80.8
	ICL	**96.5**	**53.3**	**93.3**	**92.6**
1.0	Vote	87.3	42.6	91.7	76.8
	ICL	**96.0**	**52.7**	**93.5**	**91.0**

Table 4: In ablation study, inference process of LM is eliminated. With the same demonstrations collected, ICL is compared to label voting method. Best results are **bold**.

moving inference module. For different λ values, KNN algorithm samples demonstrations on synthetic data. The vote method selects query labels by label-voting (without inference), while LM inferences for ICL based on the same demonstrations. Results in Table 4 show that ICL significantly outperforms the vote method. It indicates that more demonstrations with the same label enable ICL to better uncover query-related latent concepts for inference rather than just copying labels. Additionally, we also compared the performance of three methods with different numbers of demonstrations, as shown in Figure 3. With few demonstrations, random method in few-shot learning performs worse than zero-shot learning. It often samples demonstrations with the same label difficultly, making it hard for ICL to classify accurately. Meanwhile, TopK-SD outperforms TopK when demonstrations are three or fewer, showing that label consistency is crucial with limited demonstrations.

Role of Parameter λ in Data Synthesis The efficacy of the TopK-SD method is contingent upon the selection of λ. Optimal λ values yield demonstrations that excel in both semantic similarity and label consistency. To elucidate this relationship, we embarked on experiments to assess how varying λ influences these dual metrics as shown in Figure 4. In Figure (a), we executed KNN sampling across a spectrum of



(a) KNN Sampling (b) Selection of λ

Figure 4: Experiments (a) and (b) were both conducted on the TREC dataset. Figure (a): KNN sampling with different values of K; Figure (b): Synthetic embeddings using different values of λ, with only 8 demonstrations selected.

K values. Specifically, different values of K, utilizing both the original and synthetic embeddings ($\lambda = 0.6$). The synthetic embedding, when subjected to KNN sampling, outperforms the original embedding in procuring demonstrations with congruent labels. As shown in the figure, consistency has essentially improved by more than 10%. Figure (b) examines the average semantic similarity and label consistency of demonstrations from KNN sampling on synthetic embeddings with λ values from 0.0 to 1.0. $\lambda = 1.0$ is the original embedding. Semantic similarity, measured by cosine similarity ($[-1, 1]$ range, scaled by 100 for clarity), increases with λ. Label consistency peaks at $\lambda = 0.6$. At $\lambda = 0.6$, semantic similarity is high and stable, while label consistency surges. Our findings underscore the criticality of dataset-specific λ optimization. Further substantiating this point, Table 4 reveals that the quality of demonstrations selected via different λ values markedly impacts subsequent model inference.

Conclusion

This work rethinks the learning paradigm of ICL as transductive learning, establishing a transductive label propagation framework based on Bayesian view that highlights the importance of semantic similarity and label consistency to the shared concepts. We propose TopK-SD for demonstration selection. Experiments on multiple benchmarks show significant improvements. It validates the importance of label consistency and confirmed the derivation of Bayesian transduction. We hope this work not only provides a new perspective on the essence of ICL but also offers a robust mathematical foundation for future research.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. U23B2056 and No. 62306026), in part by the National Science and Technology Major Project under Grant 2022ZD0120202, in part by the Fundamental Research Funds for the Central Universities, and in part by the State Key Laboratory of Complex & Critical Software Environment.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

An, S.; Lin, Z.; Fu, Q.; Chen, B.; Zheng, N.; Lou, J.-G.; and Zhang, D. 2023. How Do In-Context Examples Affect Compositional Generalization? *arXiv preprint arXiv:2305.04835*.

Bai, Y.; Chen, F.; Wang, H.; Xiong, C.; and Mei, S. 2024. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in neural information processing systems*, 36.

Bi, X.; Chen, D.; Chen, G.; Chen, S.; Dai, D.; Deng, C.; Ding, H.; Dong, K.; Du, Q.; Fu, Z.; et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Dai, D.; Sun, Y.; Dong, L.; et al. 2022. Why can GPT learn in-context? Language models implicitly perform gradient descent as meta-optimizers. *arXiv preprint arXiv:2212.10559*.

Devlin, J. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ding, N.; Levinboim, T.; Wu, J.; Goodman, S.; and Soricut, R. 2023. CausalLM is not optimal for in-context learning. *arXiv preprint arXiv:2308.06912*.

Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Ma, J.; Li, R.; Xia, H.; Xu, J.; Wu, Z.; Liu, T.; et al. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Floridi, L.; and Chiriatti, M. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30: 681–694.

Gao, T.; Fisch, A.; and Chen, D. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hao, Y.; Sun, Y.; Dong, L.; Han, Z.; Gu, Y.; and Wei, F. 2022. Structured Prompting: Scaling In-Context Learning to 1,000 Examples. *arXiv preprint*, abs/2212.06713.

Hu, M.; and Liu, B. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168–177.

Kim, H. J.; Cho, H.; Kim, J.; Kim, T.; Yoo, K. M.; and goo Lee, S. 2022. Self-generated In-Context Learning: Leveraging Autoregressive Language Models as a Demonstration Generator. *arXiv preprint*, abs/2206.08082.

Kossen, J.; Gal, Y.; and Rainforth, T. 2023. In-context learning learns label relationships but is not conventional learning. *arXiv preprint arXiv:2307.12375*.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Li, X.; and Qiu, X. 2023. Finding Supporting Examples for In-Context Learning. *CoRR*, abs/2302.13539.

Liu, J.; Shen, D.; Zhang, Y.; Dolan, B.; Carin, L.; and Chen, W. 2021. What Makes Good In-Context Examples for GPT-3? *arXiv preprint arXiv:2101.06804*.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35.

Liu, Y.; Liu, J.; Shi, X.; Cheng, Q.; and Lu, W. 2024. Let's Learn Step by Step: Enhancing In-Context Learning Ability with Curriculum Learning. *arXiv preprint*, arXiv:2402.10738.

Lu, Y.; Bartolo, M.; Moore, A.; Riedel, S.; and Stenetorp, P. 2022. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8086–8098. Dublin, Ireland: Association for Computational Linguistics.

Luo, M.; Xu, X.; Liu, Y.; Pasupat, P.; and Kazemi, M. 2024. In-context learning with retrieved demonstrations for language models: A survey. *arXiv preprint arXiv:2401.11624*.

Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 1.

Min, B.; Ross, H.; Sulem, E.; Veyseh, A. P. B.; Nguyen, T. H.; Sainz, O.; Agirre, E.; Heintz, I.; and Roth, D. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2): 1–40.

Pang, B.; and Lee, L. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv preprint cs/0409058*.

Peng, K.; Ding, L.; Yuan, Y.; Liu, X.; Zhang, M.; Ouyang, Y.; and Tao, D. 2024. Revisiting demonstration selection strategies in in-context learning. *arXiv preprint arXiv:2401.12087*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Rajpurkar, P. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Reimers, N. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084*.

Robertson, S.; Zaragoza, H.; et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4): 333–389.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1631–1642.

Sorensen, T.; Robinson, J.; Rytting, C. M.; Shaw, A. G.; Rogers, K. J.; Delorey, A. P.; Khalil, M.; Fulda, N.; and Wingate, D. 2022. An information-theoretic approach to prompt engineering without ground truth labels. *arXiv preprint arXiv:2203.11364*.

Su, H.; Kasai, J.; Wu, C. H.; Shi, W.; Wang, T.; Xin, J.; Zhang, R.; Ostendorf, M.; Zettlemoyer, L.; Smith, N. A.; et al. 2022. Selective annotation makes language models better few-shot learners. *arXiv preprint arXiv:2209.01975*.

Sun, X.; Li, X.; Li, J.; Wu, F.; Guo, S.; Zhang, T.; and Wang, G. 2023. Text classification via large language models. *arXiv preprint arXiv:2305.08377*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

von Oswald, J.; Niklasson, E.; Randazzo, E.; Sacramento, J.; Mordvintsev, A.; Zhmoginov, A.; and Vladymyrov, M. 2023. Transformers Learn In-Context by Gradient Descent. In *International Conference on Machine Learning, ICML 2023*, volume 202 of *Proceedings of Machine Learning Research*, 35151–35174. Honolulu, Hawaii, USA: PMLR.

Voorhees, E. M.; and Tice, D. M. 2000. Building a question answering test collection. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 200–207.

Wang, B.; and Komatsuzaki, A. 2021. GPT-J-6B: A 6 billion parameter autoregressive language model.

Wang, L.; Li, L.; Dai, D.; Chen, D.; Zhou, H.; Meng, F.; Zhou, J.; and Sun, X. 2023. Label words are anchors: An information flow perspective for understanding in-context learning. *arXiv preprint arXiv:2305.14160*.

Wies, N.; Levine, Y.; and Shashua, A. 2023. The learnability of in-context learning. *Advances in Neural Information Processing Systems*, 36: 36637–36651.

Wies, N.; Levine, Y.; and Shashua, A. 2024. The learnability of in-context learning. *Advances in Neural Information Processing Systems*, 36.

Williams, A.; Nangia, N.; and Bowman, S. R. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Wu, Z.; Wang, Y.; Ye, J.; and Kong, L. 2022. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. *arXiv preprint arXiv:2212.10375*.

Xie, S. M.; Raghunathan, A.; Liang, P.; and Ma, T. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Ye, J.; Wu, Z.; Feng, J.; Yu, T.; and Kong, L. 2023. Compositional exemplars for in-context learning. In *International Conference on Machine Learning*, 39818–39833. PMLR.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Zhang, Y.; Feng, S.; and Tan, C. 2022. Active Example Selection for In-Context Learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9134–9148. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.