Label Unreliability in In-Context Learning

Anonymous ACL submission

Abstract

In-context learning (ICL) enables large language models (LLMs) to solve downstream tasks with a small set of labeled demonstrations. A central problem in ICL is how to select these demonstrations, and most methods approach it empirically through similarity or diversity. However, to design effective selection strategies, it is important first to understand what ICL is actually learning, particularly how it depends on the relationship between inputs and labels. To study this, we unify prior studies under the Label Unreliability framework, which captures how unreliable labels can provide imperfect supervision. Viewing ICL as implicitly performing transductive label propagation, we establish a bridge between selection strategies and label unreliability which reveals a key insight: similarity-based selection is highly sensitive to label unreliability, whereas diversity-based selection offers robustness. Effective selection therefore requires balancing similarity, to capture meaningful sample representations, with diversity, to mitigate the effects of imperfect supervision.

1 Introduction

In-context learning (ICL) enables large language models (LLMs) to solve downstream NLP tasks by conditioning on a small set of labeled demonstrations placed in the prompt (Mann et al., 2020; Min et al., 2023; Liu et al., 2023). A central practical question is how to select these demonstrations. Previous works (Bai et al., 2024; Dong et al., 2024) show that most selection strategies are heuristic and typically emphasize either similarity or diversity of the demonstrations. Specifically, similarity-based selection retrieves examples that closely match the query in semantics, syntax and lexical patterns, thereby reducing distribution shift between demonstrations and the test instance (Liu et al., 2021; Wu et al., 2022; Peng et al., 2024). This often improves accuracy when the query aligns well with the retrieved examples, but can overfit to narrow local patterns and generalize poorly to unseen structures. Diversity-based selection, in contrast, seeks demonstrations that are mutually different (and sometimes different from the query), aiming to cover a wider range of label-relevant structures and reasoning patterns (Su et al., 2022; Zhang et al., 2022; Levy et al., 2023). While diversity can help with novel compositions, it may weaken the performance if the query well aligns with the demonstrations.

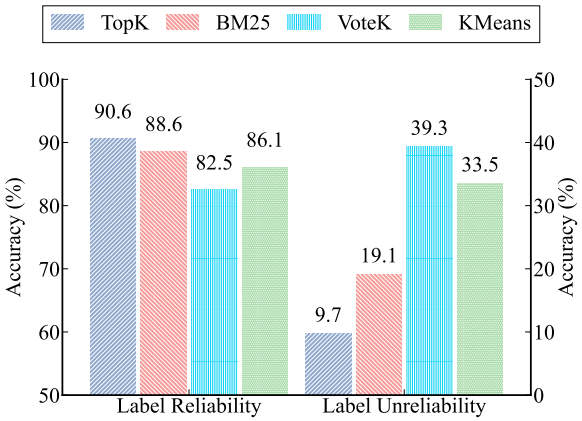


Figure 1: On AGNews, while the similarity-based method achieves higher performance with reliable labels, the diversity-based method demonstrates greater robustness with unreliable labels.

However, before that, it is necessary to understand what ICL actually learns in order to adopt an appropriate selection strategy. Existing work introduces random labels as label uncertainty and finds that ICL is not sensitive to label correctness, with incorrect labels having a limited impact on ICL performance (Min et al., 2022). Subsequent work highlights that ICL's sensitivity to unreliable labels—including both label uncertainty and label corruption—is influenced by the number of demonstrations (Kossen et al., 2024) as well as the model size (Wei et al., 2023). To summarize, we frame

the phenomena of both label uncertainty and label corruption under the concept of label unreliability, where unreliable labels provide imperfect supervision for ICL demonstrations.

This work establishes a bridge between selection strategies and label unreliability from the perspective of label propagation. This perspective is motivated by previous study that interprets ICL as an implicit form of transductive label propagation (Chen et al., 2025) within a Bayesian inference framework (Xie et al., 2021; Wies et al., 2023), which explains why similarity is crucial for transductive inference in ICL. By deriving from this mathematical hypothesis, we show that similar demonstrations are highly sensitive to label unreliability, whereas diverse demonstrations exhibit robustness to label unreliability. To this end, we design experiments with different levels of label unreliability, including label uncertainty and label corruption, to evaluate the performance of similarity-based and diversity-based selection strategies, and conclude that with low unreliability, similar demonstrations perform better, whereas with high unreliability, diverse ones are more robust. Therefore, in studies on ICL, the trade-off between similarity and diversity is of critical importance.

Briefly, we summarize our contributions as follows: (1) We unify label uncertainty and label corruption under the concept of label unreliability to systematically study the impact of imperfect supervision on ICL selection strategies. (2) By modeling ICL as a transductive label propagation process, we theoretically show that similarity-based demonstrations are intrinsically more sensitive to label unreliability, whereas diversity-based demonstrations offer greater robustness against such imperfections. (3) Empirically, we find that similarity-based selection achieves higher accuracy when label unreliability is low, while diversity-based selection exhibits superior robustness as label quality deteriorates.

2 Mathematical Derivation

2.1 Formulation of In-Context Learning

ICL is a training-free learning paradigm in which LLMs perform downstream tasks by selecting demonstrations $C = \{x_1, y_1, \cdots, x_k, y_k\}$ for a given query x (Dong et al., 2024; Luo et al., 2024). The LLM predicts the likelihood of each candidate label $y \in \mathcal{Y} = \{y_1, \cdots, y_M\}$.

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} P\left(y \mid C, x\right) \tag{1}$$

The objective is to select a set of demonstrations $\{(x_i, y_i)\}_{i=1}^{k}$ to boost prediction accuracy.

2.2 Transductive Label Propagation

Since ICL selects demonstrations based on the query x, it can be interpreted as an implicit form of transductive label propagation (Chen et al., 2025).

$$\mathbf{y} = \sum_{i=1}^{k} s_i \, y_i^* + s_0 y_0 \tag{2}$$

y_i^* denotes the ground-truth label of the i-th demonstration, and s_i denotes the similarity between the i-th demonstration and the query. Additionally, $s_0 y_0$ represents the model's prior knowledge, that is, the model's initial prediction y_0 along with its associated weight s_0. $\mathbf{y}$ denotes the prediction obtained from reliable demonstrations.

In previous work, it is explained that similarity is crucial for ICL (Dong et al., 2024; Liu et al., 2021; Chen et al., 2025). Under the L-Lipschitz constraint, selecting demonstrations that are highly similar to the query minimizes the feature discrepancy between the query x and the demonstrations $\{x_i\}_{i=1}^{k}$, thereby reducing prediction error due to input perturbations $\boldsymbol{\epsilon}$.

$$\|\mathbf{y} - \mathbf{y}^*\|_F \leq L\|\boldsymbol{\epsilon}\|_F + o\left(\max_{i=1}^{k} \|x - x_i\|_F\right) \tag{3}$$

$\mathbf{y}^*$ represents the ground-truth label of the query x. Consequently, similarity-based selection methods can effectively aggregate the most relevant information, enabling the model to construct precise, task-aligned internal representations while reducing errors caused by perturbations.

2.3 Label Unreliability

In practical applications, the labels of demonstrations used for ICL may be unreliable due to annotation errors or inherent ambiguity. In the label unreliability model, each observed label y is considered a potentially unreliable version of the ground-truth label y^*. The unreliability is modeled as a probabilistic mixture: with probability $1 - \eta$, the label is correct, and with probability η, it is replaced drawn from a distribution $p(y)$. Formally, this can be expressed as:

$$y = (1 - \eta)\, y^* + \eta\, p(y) \tag{4}$$

Where y^* is the ground-truth label, $p(y)$ represents a general distribution over labels capturing uncertainty or corruption, and $\eta \in [0, 1]$ quantifies the

label unreliability rate. A value of $\eta = 0$ indicates fully reliable labels, while $\eta = 1$ indicates labels entirely determined by $p(y)$. Intermediate values represent partially reliable labels, providing a unified framework to model annotation errors, inherent ambiguity, or unreliable demonstrations in ICL.

To account for this, We capture the label unreliability between the ground-truth label y^* and the observed label y using a transition matrix $T \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|}$, where $|\mathcal{Y}|$ is the number of classes. Each entry of the matrix is defined as

$$T_{a,b} = P(y = b \mid y^* = a), \tag{5}$$

The transition matrix T characterizes the relationship between ground-truth labels and their corresponding observed labels under unreliable supervision, specifying the likelihood that the ground-truth label a is observed as label b. Under this equation, the observed label $y_i = Ty_i^*$ is generated from the ground-truth label y_i^* through the unreliability process induced by T. Consequently, ICL operates on demonstrations whose labels provide imperfect and unreliable supervision, governed by the transition matrix T.

$$\begin{aligned}\tilde{\mathbf{y}} &= \sum_{i=1}^{k} s_i\, y_i + s_0\, y_0 \\ &= \sum_{i=1}^{k} s_i\, T y_i^* + s_0\, y_0\end{aligned} \tag{6}$$

The unreliable labels $\{y_i\}_{i=1}^k$ causes a prediction error $e = \tilde{\mathbf{y}} - \mathbf{y}^*$, deviating from the true result $\mathbf{y}^*$.

2.4 Unreliability Sensitivity

To characterize how prediction error e responds to label unreliability captured by the transition matrix T, we measure the sensitivity of e with respect to T by computing the gradient $\nabla_T e$ and taking its Frobenius norm as a scalar measure of sensitivity.

$$\|\nabla_T e\|_F = \left\| \sum_{i=1}^{k} s_i\, y_i^* \right\|_F \tag{7}$$

The rate of change of e with respect to T is defined as the sensitivity of the demonstration to label unreliability. Thus, the sensitivity depends on both the similarity s_i and the ground-truth label y_i^*.

To investigate the factors influencing unreliability sensitivity, we normalize the similarity in Equation 7 as $\tilde{s}_i = \frac{s_i}{\sum_{j=1}^{k} s_j}$. As shown below, the sensitivity is mainly determined by semantic similarity and label diversity.

$$\left\| \sum_{i=1}^{k} s_i\, y_i^* \right\|_F = \underbrace{\left(\sum_{i=1}^{k} s_i \right)}_{\text{Similarity}} \cdot \underbrace{\left\| \sum_{i=1}^{k} \tilde{s}_i\, y_i^* \right\|_F}_{\text{Diversity}} \tag{8}$$

Similarity. Obviously, the sensitivity is proportional to the sum of similarities $\sum_{i=1}^{k} s_i$. Similar to traditional label propagation algorithms, samples with higher similarity transmit label information more strongly, which makes unreliable labels more likely to influence predictions for similar samples. When a similarity-based TopK (Liu et al., 2021) sampling method is used, the sampling bias induced by the concentrated distribution of demonstrations can amplify the effect of label unreliability, and boundary samples are particularly prone to include highly similar demonstrations with unreliable labels.

Diversity. $\left\| \sum_{i=1}^{k} \tilde{s}_i y_i^* \right\|_F$ reflects the label diversity of the demonstrations. It is regarded as the weighted expectation with weights $\tilde{s}_i$, and by introducing the variance of y_i^* with weights $\tilde{s}_i$, the following identity can be obtained:

$$\underbrace{\left\| \sum_{i=1}^{k} \tilde{s}_i\, y_i^* \right\|_F^2}_{\text{Expectation}} + \underbrace{\sum_{i=1}^{k} \tilde{s}_i \left\| y_i^* - \sum_{j=1}^{k} \tilde{s}_j y_j^* \right\|_F^2}_{\text{Variance}} \equiv 1 \tag{9}$$

For a detailed proof of the identity, see the Appendix A. This identity elucidates a negative correlation between the expectation and the variance: a higher expectation corresponds to a lower variance, implying greater consistency among the sample labels and increased sensitivity to unreliability; conversely, greater label diversity is associated with enhanced robustness. Therefore, diversity-based sampling methods, such as VoteK (Su et al., 2022) and KMeans (Zhang et al., 2022), select demonstrations by accounting for coverage in the semantic space, which in turn enhances the robustness.

Based on the above derivation and smoothing theory, since similar samples share the same label, we can conclude that a similarity-based selection strategy, due to the concentration of the sampling region, is more sensitive to label unreliability, which in turn leads to imperfect supervision from the demonstrations. In contrast, diversity-based selection considers the global distribution, resulting in a more dispersed sampling region. The selected

Model	Method	SST2	SST5	AGNews	Subj	CR	MNLI	QNLI	Avg.	Δ
GPT-j-6b	Random	90.9(1.8)	44.7(2.6)	70.7(2.1)	73.1(1.3)	81.1(3.1)	41.2(2.0)	49.3(2.7)	64.4	-
	TopK	94.2(1.3)	**51.7(1.6)**	**86.7(1.2)**	**88.6(2.6)**	**89.3(0.8)**	**45.4(2.1)**	53.0(2.5)	**72.7**	↑8.3
	BM25	93.7(1.7)	47.8(3.2)	82.7(2.7)	84.3(2.4)	87.3(1.3)	42.8(3.7)	52.1(2.0)	70.1	↑5.7
	VoteK	**95.3(1.5)**	50.6(4.7)	73.7(4.2)	72.7(7.0)	70.8(5.0)	42.1(5.3)	**54.8(4.1)**	65.7	↑1.3
	KMeans	94.7(1.7)	42.9(5.6)	73.9(3.2)	80.3(9.3)	77.6(11.0)	40.4(6.2)	50.9(4.3)	65.8	↑1.4
LlaMA3-8b	Random	95.7(0.9)	45.5(4.5)	84.0(2.1)	89.8(1.7)	89.3(0.9)	57.0(3.6)	55.0(2.0)	73.8	-
	TopK	96.6(0.8)	**51.8(2.6)**	**90.6(1.3)**	**95.5(1.0)**	**92.3(1.3)**	57.8(2.5)	**59.0(1.6)**	**77.7**	↑3.9
	BM25	95.7(0.8)	48.9(2.3)	88.6(1.2)	93.2(0.5)	92.1(1.7)	**58.3(2.5)**	56.9(1.8)	76.2	↑2.4
	VoteK	**96.9(1.5)**	51.6(2.6)	82.5(6.6)	94.9(2.2)	86.0(3.8)	55.9(7.3)	53.7(5.8)	74.5	↑0.7
	KMeans	96.7(0.8)	45.2(4.9)	86.1(4.8)	87.3(16.0)	90.2(3.1)	54.9(4.2)	57.6(6.6)	74.0	↑0.2

Table 1: Evaluation of methods with **Ground-truth Labels**, reporting performance, where Δ denotes the comparison with the Random method. The highest accuracy is highlighted in **bold** and the second is underlined.

labels are more diverse, making the strategy relatively more robust to label unreliability and the resulting imperfect supervision. Overall, while the similarity-based method performs better when label unreliability is low, the diversity-based method demonstrates greater robustness under high label unreliability and the associated imperfect supervision.

3 Experiment Setting

3.1 Setting Up

Datasets. We use seven datasets, comprising five classification tasks and two natural language inference (NLI) tasks (Sun et al., 2023), including SST-2 (Socher et al., 2013), SST-5(Socher et al., 2013), AGNews (Zhang et al., 2015), CR (Hu and Liu, 2004), Subj (Pang and Lee, 2004), MNLI (Williams et al., 2017) and QNLI (Rajpurkar, 2016).

Models. The main experiments use LlaMA3-8b (Grattafiori et al., 2024) and GPT-j-6b (Wang and Komatsuzaki, 2021) as inference models and all-roberta-large-v1 (Reimers, 2019) as the embedding model to evaluate the performance of different methods. For details on other models, see the Appendix C.

Setting Up. In the experiments, each method selects eight demonstrations and each run uses 1,000 samples as the candidate set and 300 samples as the validation set. Each experiment is repeated five times on an A100 GPU, with the mean accuracy and standard deviation reported.

3.2 Methods

The following five selection strategies are adopted in the experiments.

Random Selecting demonstrations randomly.

TopK (Liu et al., 2021; Gao et al., 2020) From the perspective of semantic similarity, the top-K most semantically similar demonstrations are selected using an embedding model.

BM25 (Robertson et al., 2009) Term-level similarity between the query and demonstrations is evaluated using TF–IDF with length normalization.

VoteK (Su et al., 2022) At the corpus level, a query-based voting method is used to select demonstrations for ICL that are diverse.

KMeans (Zhang et al., 2022) Applying the KMeans algorithm to form clusters and selecting representative points from each cluster as demonstrations to enhance the diversity.

4 ICL with Ground-Truth Labels

In this section, we assess the performance of various selection strategies on ICL under a label unreliability rate of $\eta = 0$, corresponding to the use of ground-truth labels.

For demonstrations sampled using different selection strategies, labels were assigned based on ground-truth annotations. The experimental results, reported in Table 1, show that similarity-based methods, such as TopK and BM25, consistently outperform other methods under this condition. This finding aligns with the intuition that when labels are reliable, leveraging similar demonstrations is critical for constructing task-relevant representations and achieving accurate label prediction. In contrast, diversity-based methods, such as VoteK and KMeans, perform substantially worse under low unreliability, as they are less effective at capturing semantically coherent patterns.

Model	Method	SST2	SST5	AGNews	Subj	CR	MNLI	QNLI	Avg.	Δ
GPT-j-6b	Random	87.7(2.2)	39.1(2.9)	64.7(3.9)	56.5(3.6)	**75.6(0.9)**	42.0(2.6)	**50.3(1.3)**	59.4	-
	TopK	82.7(2.3)	36.6(2.1)	54.8(1.9)	58.9(1.8)	69.4(1.7)	39.1(3.0)	49.6(2.1)	55.9	↓3.5
	BM25	85.9(0.8)	39.3(3.7)	59.8(3.9)	54.7(1.9)	72.5(1.1)	37.8(4.4)	49.5(2.1)	57.1	↓2.3
	VoteK	87.7(4.1)	**41.6(3.4)**	63.6(3.5)	**60.4(4.7)**	70.5(4.5)	39.9(2.2)	49.2(2.4)	59.0	↓0.4
	KMeans	**89.4(2.3)**	38.5(3.2)	**64.9(3.8)**	59.9(2.3)	73.6(5.7)	**42.0(4.7)**	48.3(2.3)	**59.5**	↑0.1
LlaMA3-8b	Random	**85.2(2.8)**	**35.6(4.4)**	49.1(1.4)	62.9(2.6)	72.2(1.5)	**52.1(1.8)**	**54.7(3.1)**	**58.8**	-
	TopK	75.2(2.2)	31.7(3.2)	35.1(3.2)	65.8(5.0)	68.7(4.1)	48.3(2.7)	53.6(1.7)	54.1	↓4.7
	BM25	81.7(2.6)	33.3(1.6)	41.5(2.9)	63.5(2.7)	70.7(0.8)	50.9(3.1)	51.4(3.5)	56.1	↓2.7
	VoteK	80.7(4.4)	34.1(2.9)	**52.1(6.1)**	65.7(3.8)	71.3(3.1)	49.9(3.3)	51.5(0.4)	57.9	↓0.9
	KMeans	78.8(4.4)	35.5(3.5)	48.9(2.4)	**66.0(2.3)**	**76.1(2.9)**	50.8(3.7)	51.7(1.6)	58.3	↓0.5

Table 2: Evaluation of methods with **Uncertain Labels**, reporting performance, where Δ denotes the comparison with the Random method. The highest accuracy is highlighted in **bold** and the second is underlined.

A detailed analysis further clarifies this phenomenon. Similarity-based methods exploit highly similar demonstrations to guide LLMs in extracting task-relevant features and forming structured representations. By presenting demonstrations that share high similarity with the query, these methods enable better alignment between internal representations and the target label space, allowing LLMs to optimize from observed patterns and accurately predict labels. In contrast, diversity-based methods, which prioritize coverage over similarity, provide weaker and less signals for representation formation, and are therefore less effective even when labels are reliable.

5 ICL with Uncertain Labels

In this section, we model label unreliability as label uncertainty. When $p(y)$ in Equation 4 is uniformly distributed over the candidate label set $\mathcal{Y}$, the resulting label unreliability can be viewed as a form of label uncertainty, corresponding to a soft-label setting. Therefore, the transition matrix T of label uncertainty is defined as follows:

$$T_{i,j} = \begin{cases} 1 - \eta + \dfrac{\eta}{|\mathcal{Y}|}, & i = j, \\ \dfrac{\eta}{|\mathcal{Y}|}, & i \neq j. \end{cases} \tag{10}$$

When $\eta = 1$, the demonstrations are equipped with random labels, and the experimental results are shown in Table 2. The experimental results show that similarity-based selection strategies consistently achieve slightly lower accuracy than diversity-based selection strategies, while the performance of diversity-based strategies is comparable to that of the Random method. It can be concluded that similarity-based selection strategies are more sensitive to label uncertainty, whereas diversity-based selection strategies are more robust.

To further investigate the label uncertainty rate η, we control the probability distributions used to generate soft labels under different uncertainty levels and adopt Monte Carlo sampling to compute the accuracy, likelihood and information entropy of different strategies. The experimental results are shown in Figure 2. With respect to the gradient of accuracy as uncertainty increases, similarity-based selection exhibits a steeper decline than diversity-based selection. With respect to the gradient of accuracy as uncertainty increases, similarity-based selection exhibits a steeper decline than diversity-based selection. Moreover, the likelihood assigned by similarity-based selection to the correct query label becomes increasingly closer to $\frac{1}{|\mathcal{Y}|}$, and the corresponding entropy rises from a lower value toward maximum. The results reveal that label uncertainty is more likely to induce uncertainty in ICL predictions through similarity, leading LLMs to exhibit lower confidence in their own outputs.

6 ICL with Corrupted Labels

In this section, we model label unreliability as label corruption. When $p(y)$ in Equation 4 is uniformly distributed over the label set $\{y \in \mathcal{Y} \mid y \neq y*\}$, the resulting label corruption can be viewed as a form of label unreliability corresponding to a noisy-label setting. Therefore, the transition matrix T of label corruption is defined as follows:

$$T_{i,j} = \begin{cases} 1 - \eta, & i = j, \\ \dfrac{\eta}{|\mathcal{Y}| - 1}, & i \neq j. \end{cases} \tag{11}$$

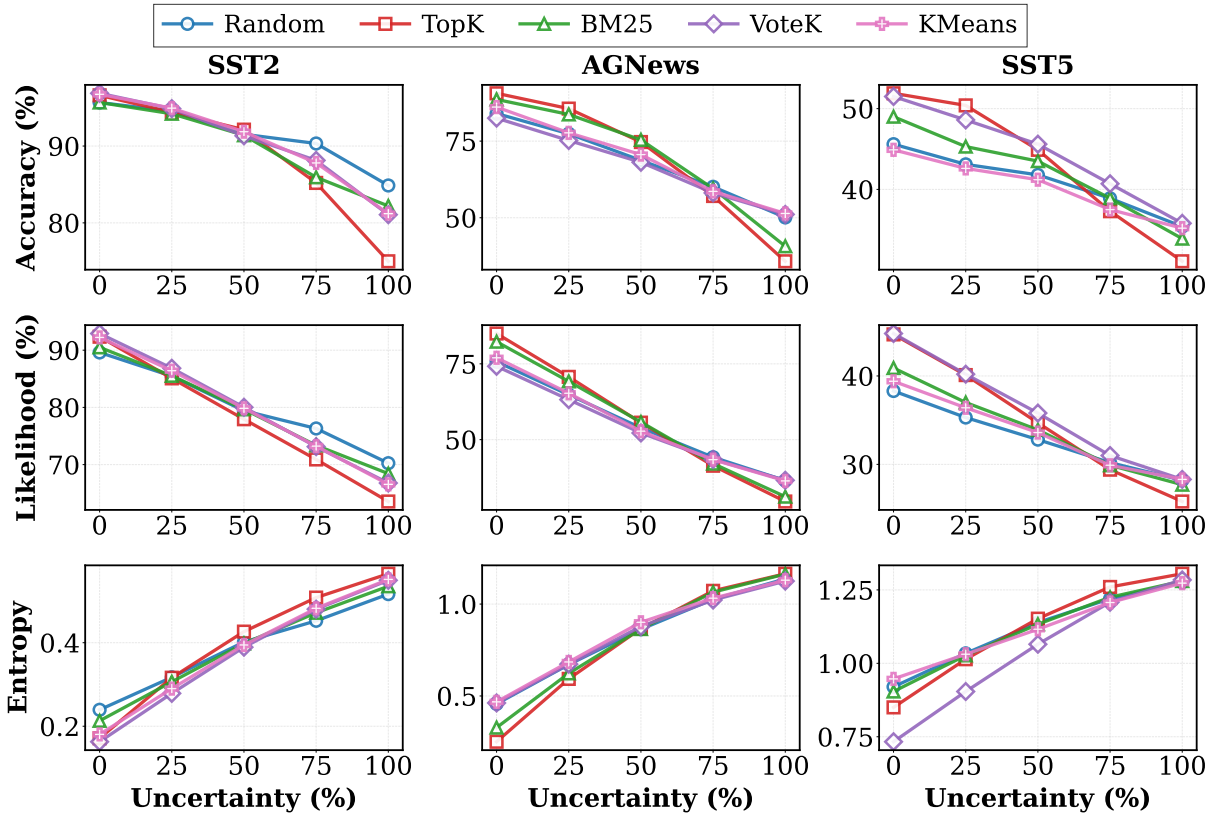


Figure 2: Under varying **Uncertainty Rate**, we examine and compare the trends of accuracy, likelihood, and information entropy across five strategies. The experiments are conducted on three datasets employing LLaMA3-8B.

Model	Method	SST2	SST5	AGNews	Subj	CR	MNLI	QNLI	Avg.	Δ
GPT-j-6b	Random	78.9(3.5)	**40.3(2.0)**	59.5(1.7)	43.9(3.4)	**62.0(3.7)**	39.7(2.5)	**49.7(2.7)**	**53.4**	-
	TopK	37.6(2.9)	32.5(4.2)	17.5(2.7)	13.5(1.6)	27.3(1.5)	37.1(1.8)	48.3(1.3)	30.5	↓22.9
	BM25	66.7(2.6)	35.6(3.8)	35.1(3.4)	24.7(3.3)	41.3(2.1)	37.5(3.0)	47.3(1.7)	41.2	↓12.2
	VoteK	**79.5(10.9)**	37.5(3.7)	**61.7(4.4)**	**48.2(3.9)**	35.5(1.9)	**41.1(3.5)**	48.8(2.3)	50.3	↓3.1
	KMeans	79.0(11.9)	36.5(3.8)	61.7(4.5)	35.7(7.9)	52.2(20.8)	41.0(1.8)	48.9(1.8)	50.7	↓2.7
LlaMA3-8b	Random	**54.9(1.1)**	**31.5(3.1)**	36.0(2.4)	**28.5(1.6)**	**48.2(4.1)**	47.3(2.8)	**50.9(4.5)**	**42.5**	-
	TopK	17.0(1.7)	24.7(2.1)	9.7(0.8)	8.8(1.7)	21.1(0.8)	44.2(2.6)	45.8(3.4)	24.5	↓18.8
	BM25	41.3(2.9)	27.9(1.9)	19.1(1.6)	13.2(1.9)	29.7(2.4)	46.1(3.0)	46.2(2.5)	31.9	↓10.6
	VoteK	30.1(15.4)	27.3(3.8)	**39.3(11.9)**	24.1(5.3)	35.5(1.9)	**48.3(1.9)**	48.9(2.2)	36.2	↓6.3
	KMeans	22.9(12.8)	30.8(4.9)	33.5(3.2)	24.0(13.4)	45.9(16.9)	46.9(6.3)	47.1(2.5)	35.9	↓6.6

Table 3: Evaluation of methods with **Corrupted Labels**, reporting performance, where Δ denotes the comparison with the Random method. The highest accuracy is highlighted in **bold** and the second is underlined.

When $\eta = 1$, the demonstrations are equipped with flipped labels, and the experimental results are shown in Table 3. The results indicate that the accuracy of similarity-based selection strategies drops substantially, whereas that of diversity-based strategies decreases only marginally. It can be concluded that similarity-based selection is highly sensitive to label corruption, whereas diversity-based selection strategies are more robust. Across text classification tasks, LLMs are more sensitive to label corruption in demonstrations, as classification relies more on direct label propagation, whereas NLI tasks are more robust because models focus on the semantic and logical relationships between the premise and the hypothesis.

To further investigate the label corruption rate η, we flip the labels of demonstrations according to the probabilities given by the transition matrix

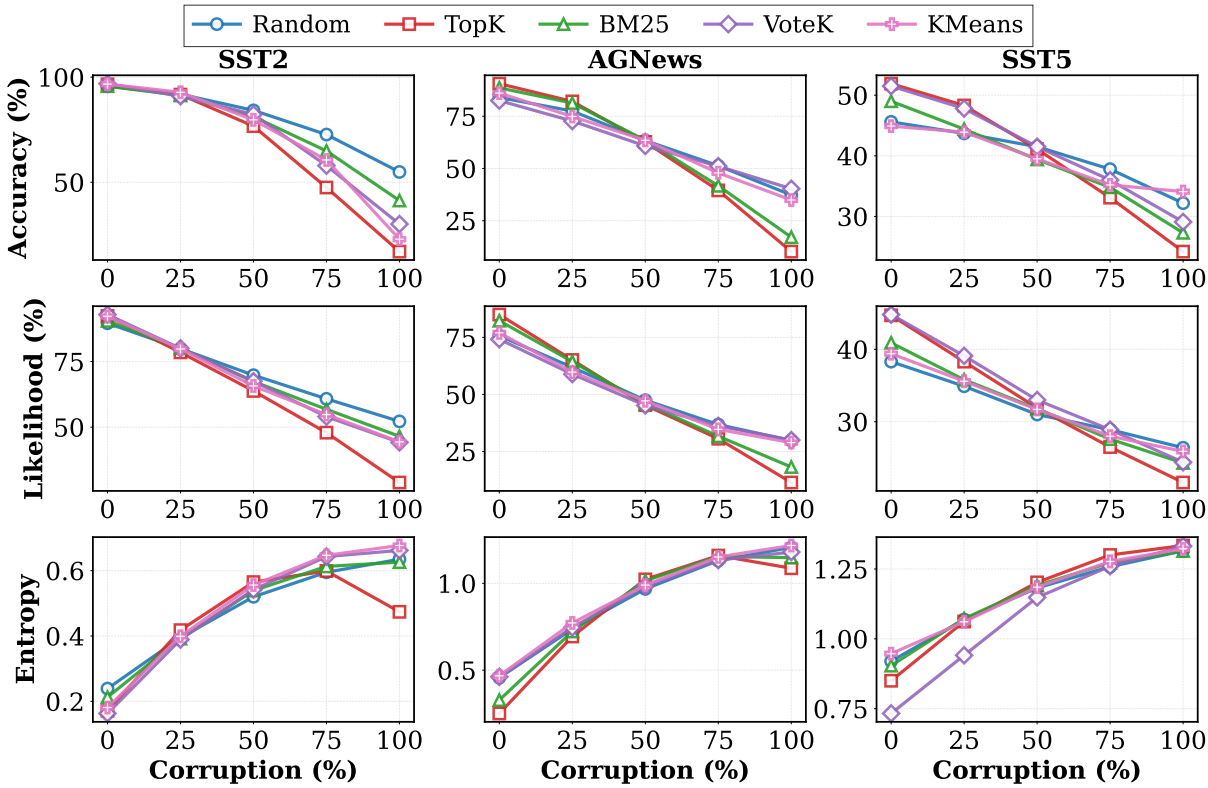


Figure 3: Under varying **Corruption Rate**, we examine and compare the trends of accuracy, likelihood, and information entropy across five strategies. The experiments are conducted on datasets employing LLaMA3-8B.

T and, under different levels of label corruption, compute the accuracy, likelihood, and information entropy of different strategies. The experimental results are shown in Figure 3. With respect to the gradient of accuracy as corruption increases, similarity-based selection exhibits a steeper decline than diversity-based selection. Moreover, the likelihood assigned to the correct query label by similarity-based selection is much lower than that of diversity-based selection. The information entropy exhibits a pronounced convex shape, with a point t such that it increases over $[0, t]$ and decreases over $[t, 1]$, undergoing a gradient reversal at $\eta = t$. The point t is located to the right of $1 - \frac{1}{|\mathcal{Y}|}$. Based on experiments combining likelihood and information entropy, on SST2 and AGNews we observe that the likelihood of correct labels under similarity-based selection is far below $\frac{1}{|\mathcal{Y}|}$, and the gradient of information entropy is inverted, indicating that the model is fitting to the corrupted labels. SST2 is a binary classification dataset. The lower the likelihood of the correct label, the more confident the model is in its incorrect predictions. This indicates that the mapping between incorrect inputs and labels in the in-context demonstrations is learned by ICL, overriding the prior knowledge acquired during pretraining. It reveals that label corruption leads ICL to learn incorrect relationships between input and label.

7 Hybrid Experiments

From the above experiments, Similarity helps ICL learn the representations of examples, while diversity can partially mitigate the interference caused by label unreliability. An effective demonstration selection method should incorporate both. We design a hybrid selection method, VoteK-TopK, in which half of the demonstrations are drawn using VoteK sampling and the other half using TopK; the same applies to the KMeans-TopK method. The results in Table 4 show that the hybrid method VoteK-TopK achieves higher average performance than the baseline methods across two inference models, with particularly notable improvements on SST2 and SST5 compared with TopK. Moreover, the KMeans-TopK method also shows improvements compared with other baselines. Such a hybrid algorithm leverages diversity to alleviate the imperfect supervision caused by label unreliability in similar demonstrations. Therefore, in future ICL research, studying how to trade off between similarity and diversity is crucial.

Method	GPT-j-6b					LlaMA3-8b				
	SST2	SST5	AGNews	Subj	Avg.	SST2	SST5	AGNews	Subj	Avg.
Random	92.3(0.9)	47.9(2.1)	78.3(2.8)	81.9(1.9)	75.1	96.3(0.4)	50.6(3.5)	86.2(1.3)	94.3(1.2)	81.9
TopK	94.7(2.0)	51.4(3.7)	87.3(0.4)	92.7(0.7)	81.5	96.1(1.1)	53.3(3.6)	90.3(1.1)	96.0(0.4)	83.9
VoteK	94.0(1.0)	47.5(2.9)	78.8(3.1)	81.6(7.1)	75.5	97.0(1.2)	51.9(5.6)	86.1(3.9)	95.4(2.7)	82.6
VoteK-TopK	**96.3(1.1)**	**52.1(4.9)**	**87.7(1.2)**	92.1(1.0)	**82.1**	**97.3(1.0)**	54.9(1.8)	90.2(0.6)	96.9(0.9)	**84.8**
KMeans	95.2(1.6)	51.9(3.8)	79.9(6.3)	81.1(10.0)	77.0	96.7(0.9)	**55.4(3.5)**	87.8(3.1)	96.5(1.7)	84.1
KMeans-TopK	94.8(1.9)	52.1(2.4)	86.8(0.7)	**93.3(0.4)**	81.8	96.8(1.0)	53.5(1.9)	**90.5(1.0)**	**97.5(1.1)**	84.6

Table 4: With 16 demonstrations, the Hybrid method outperforms baseline methods across classification datasets. The highest accuracy is highlighted in **bold** and the second is underlined.

8 Discussions and Related Works

Sampling bias is introduced by similarity. The selection of demonstrations can induce a bias toward specific predictions that does not necessarily reflect the model's understanding of the task (Zhao et al., 2021; Fei et al., 2023). When only similarity-based sampling is used, the selected data tend to be concentrated in a local region and capture only local characteristics, leading LLMs to learn a narrow representation and thus introducing sampling bias. Under this condition, LLMs become highly sensitive to label unreliability in the demonstrations, which can be viewed as an amplification effect of label unreliability caused by sampling bias.

Large models are many-shot learners. A recent perspective suggests that ICL can be viewed as a form of many-shot learning to achieve effective learning (Agarwal et al., 2024). In a many-shot setting, some work proposes selecting a few similar demonstrations along with a large number of random examples, an approach that can effectively improve accuracy (Golchin et al., 2024). The motivation behind this design is to help inference models identify similar patterns for learning while leveraging diverse demonstrations to mitigate the impact of label unreliability, thereby enhancing both ICL optimization and generalization capabilities.

We rethink the relationship between inputs and labels. A highly cited study has shown that ICL is relatively insensitive to the unreliability of example labels, suggesting that label unreliability is not always a critical factor (Min et al., 2022). However, this study relies on random sampling strategies, where the inherent diversity of randomly selected demonstrations provides a degree of robustness against label unreliability, explaining the observed insensitivity. Further studies have found that larger models (Wei et al., 2023) or a greater number of in-context demonstrations (Kossen et al., 2024) are more sensitive to label unreliability, as models may rely heavily on the context examples, allowing unreliable labels to override their prior knowledge. Combining the mathematical formulation of label propagation with unreliable labels in Equation 6, the context can dominate the knowledge obtained from pre-training.

$$\left\| \sum_{i=1}^{k} s_i T y_i^* \right\|_F \gg \|s_0 y_0\|_F \tag{12}$$

As scale increases, LLMs are better able to capture the intrinsic relationship between the demonstrations and the query, which implies larger s_i values with $s_i \gg s_0$, potentially leading to overfitting to unreliable labels. Similarly, increasing the number of examples ($\sum_{i=1}^{k} s_i \gg s_0$) can exacerbate this effect. In our study, as shown in Figure 3, similarity-based selection alone may cause demonstrations to override prior knowledge due to label unreliability. This highlights that the relationship between inputs and labels is critically important.

9 Conclusion

Our work shows that similarity-based selection effectively captures relevant patterns, improving accuracy when labels are reliable, but it also makes the model more sensitive to label unreliability. In contrast, diversity-based selection spreads demonstrations across a wider range, mitigating the effects of unreliable labels and enhancing robustness. Combining both strategies balances learning relevant patterns with resisting imperfect supervision, leading to more stable ICL performance. We hope that future research can identify a solid strategy to trade off between similarity and diversity, adapted to specific scenarios.

Limitations

While our study highlights the trade-off between similarity and diversity in mitigating the impact of label unreliability, several limitations remain. Our analysis primarily considers synthetic or flipped label unreliability, which may not fully capture the complexity of real-world noisy or ambiguous annotations. The proposed insights are mainly validated on text classification and NLI tasks; their generalization to other modalities or more complex reasoning tasks remains to be explored. Finally, while balancing similarity and diversity improves robustness to label unreliability, determining the optimal trade-off in practice may require task-specific tuning and remains an open challenge.

References

Rishabh Agarwal, Avi Singh, Lei M Zhang, , and 1 others. 2024. Many-shot in-context learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:76930–76966.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. 2024. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in neural information processing systems*, 36.

Haoyang Chen, Richong Zhang, and Junfan Chen. 2025. Rethinking label consistency of in-context learning: An implicit transductive label propagation perspective.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, and 1 others. 2024. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 1107–1128.

Y. Fei, Y. Hou, Z. Chen, and 1 others. 2023. Mitigating label biases for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

S. Golchin, Y. Chen, R. Han, and 1 others. 2024. Towards compute-optimal many-shot in-context learning. In *Proceedings of the Second Conference on Language Modeling*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

J. Kossen, Y. Gal, and T. Rainforth. 2024. In-context learning learns label relationships but is not conventional learning. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Itay Levy, Ben Bogin, and Jonathan Berant. 2023. Diverse demonstrations improve in-context compositional generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1401–1422. Association for Computational Linguistics.

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*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

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*.

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

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

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

Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2024. Revisiting demonstration selection strategies in in-context learning. *arXiv preprint arXiv:2401.12087*.

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

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

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, and 1 others. 2022. Selective annotation makes language models better few-shot learners. *arXiv preprint arXiv:2209.01975*.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. *arXiv preprint arXiv:2305.08377*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ben Wang and Aran Komatsuzaki. 2021. Gpt-j-6b: A 6 billion parameter autoregressive language model.

J. Wei, J. Wei, Y. Tay, and 1 others. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Noam Wies, Yoav Levine, and Amnon Shashua. 2023. The learnability of in-context learning. *Advances in Neural Information Processing Systems*, 36:36637–36651.

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

Zhenyu Wu, Yaoxiang Wang, Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Jingjing Xu, and Yu Qiao. 2023. Openicl: An open-source framework for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 489–498.

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*.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. In *The eleventh international conference on learning representations*.

Zi Zhao, Eric Wallace, Shihao Feng, and et al. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the International Conference on Machine Learning*, PMLR, pages 12697–12706.

A Appendix: Proof

Prove the equivalence of the following objective functions.

$$\max \left\| \sum_{i=1}^{k} \tilde{s}_i \, y_i^* \right\|_F^2 \iff \min \sum_{i=1}^{k} \tilde{s}_i \, \|y_i^* - \bar{y}^*\|_F^2 \tag{13}$$

$$\bar{y}^* = \sum_{i=1}^{k} \tilde{s}_i y_i^* \tag{14}$$

Equation 13 can be interpreted as maximizing the weighted expectation with weights $\tilde{s}_i$ on the left-hand side, and minimizing the weighted variance with weights $\tilde{s}_i$ on the right-hand side. Expand the right-hand side using the variance formula.

$$\mathrm{Var}(\mathbf{Y}) = \mathbb{E}[\mathbf{Y}^2] - (\mathbb{E}[\mathbf{Y}])^2 \tag{15}$$

The weighted variance of the samples is defined as:

$$\sum_{i=1}^{k} \tilde{s}_i \, \|y_i^* - \bar{y}^*\|_F^2 = \sum_{i=1}^{k} \tilde{s}_i \, \|y_i^*\|_F^2 - \left\| \sum_{i=1}^{k} \tilde{s}_i y_i^* \right\|_F^2 \tag{16}$$

The ground-truth label y_i^* is usually represented as a one-hot vector in classification tasks, so that $||y_i^*||_F^2 = 1$. In addition, the normalization factor satisfies $\sum_{i=1}^{k} \tilde{s}_i = 1$.

$$\sum_{i=1}^{k} \tilde{s}_i \left\| y_i^* - \bar{y}^* \right\|_F^2 = 1 - \left\| \sum_{i=1}^{k} \tilde{s}_i y_i^* \right\|_F^2 \quad (17)$$

Therefore, maximizing $\left\| \sum_{i=1}^{k} \tilde{s}_i y_i^* \right\|_F^2$ is equivalent to minimizing the weighted variance $\sum_i \tilde{s}_i \left\| y_i^* - \bar{y}^* \right\|_F^2$.

B Appendix: Dataset

The experiments included seven seven datasets, comprising five classification tasks and two natural language inference tasks, including SST-2, SST-5, AGNews, CR, Subj, MNLI and QNLI with specific details shown in Table 5.

Dataset	Task	#train	#test	#class
SST-2	Classification	6,920	1821	2
SST-5	Classification	8,544	2210	5
AGNews	Classification	120,000	7600	4
Subj	Classification	8000	2000	2
CR	Classification	3394	376	2
MNLI	Inference	392702	19643	3
QNLI	Inference	104743	5463	2

Table 5: The statistics of the datasets.

C Appendix: Ablation Study

To demonstrate the generality of our mathematical derivations, we also conduct partial experiments using LlaMA3-8b (Grattafiori et al., 2024), GPT-j-6b (Wang and Komatsuzaki, 2021), LlaMA2-7b (Touvron et al., 2023), Mistral-7b (Jiang et al., 2023) and Gemma-7b (Team et al., 2024) as inference models, all-roberta-large-v1, all-MiniLM-L12-v2, all-distilroberta-v1 and all-mpnet-base-v2 as embedding models (Reimers, 2019).

By replacing three other inference models, LlaMA2-7b, Mistral-7b and Gemma-7b, and three other embedding models, all-MiniLM-L12-v2, all-distilroberta-v1 and all-mpnet-base-v2, as shown in Figure 4 and Figure 5 for ablation studies, it can be observed that with random labels, the accuracies of VoteK and KMeans are higher than that of TopK, indicating that selecting similar demonstrations causes ICL to fit label unreliability, which is a general phenomenon. The results with ground-truth labels are shown in Table 4, while those with uncertain labels are presented in Table 5 and those with corrupted labels are presented in Table 6.

D Appendix:Prompt Template

Here, we present the prompt template used for PPL inference. X_i is the input and Y_i is the label of the i^{th} in-context demonstrations. X is the input of the test data. Y is the category for classification. For all categories, LLMs calculate the lowest perplexity with different labels. "Input" and "Label" in the prompt word template are substituted with specific words in Table 6.

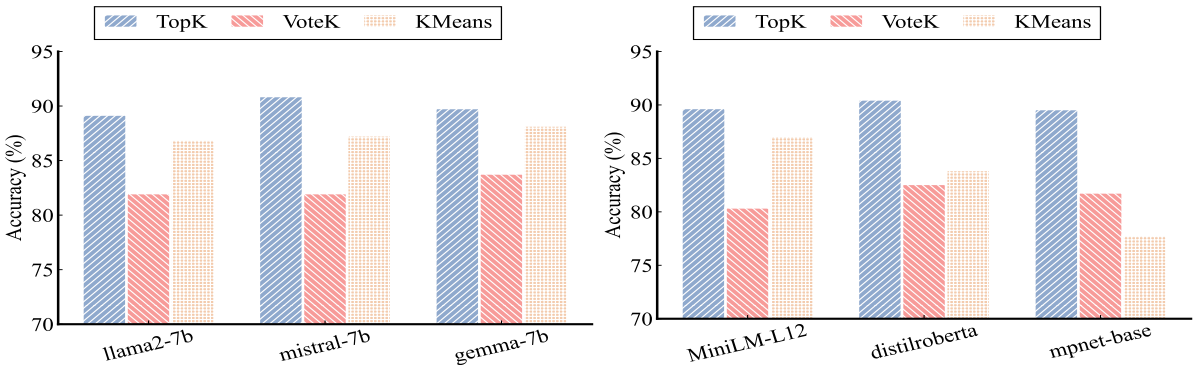
Dataset	Prompt Template
SST-2	Review: "X" Sentiment: positive Review: "X" Sentiment: negative
SST-5	Review: "X" Sentiment: terrible Review: "X" Sentiment: bad Review: "X" Sentiment: okay Review: "X" Sentiment: good Review: "X" Sentiment: great
CR	Review: "X" Sentiment: positive Review: "X" Sentiment: negative
Subj	Input: "X" Type: objective Input: "X" Type: subjective
AGNews	"X" It is about world. "X" It is about sports. "X" It is about business. "X" It is about science and technology.
MNLI	<C> Can we know <X>? Yes. <C> Can we know <X>? Maybe. <C> Can we know <X>? No.
QNLI	<C> Can we know <X>? Yes. <C> Can we know <X>? No.

Table 6: Details of Prompt template on various datasets.

E Appendix: Comparative Experiments

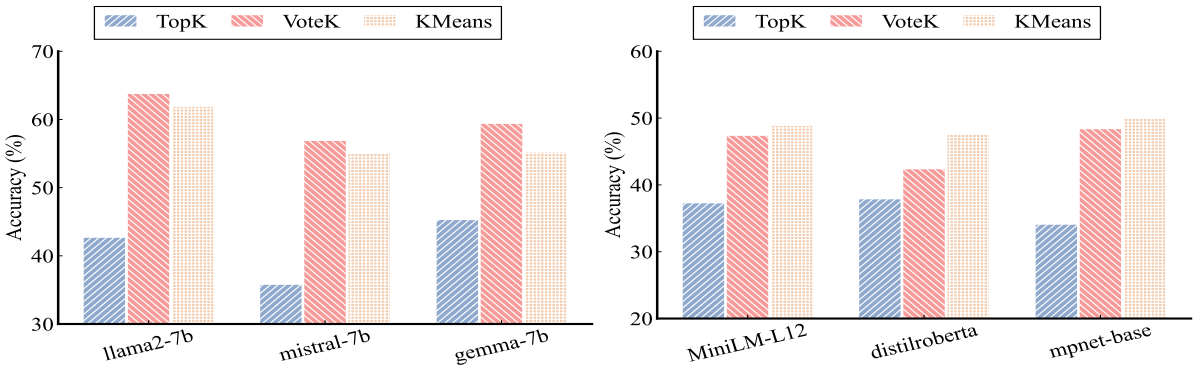
Methed	Similarity	Diversity
Random	-	-
TopK	✓	-
BM25	✓	-
Votek	-	✓
KMeans	-	✓

Table 7: Different selection strategies consider different demonstration characteristics.



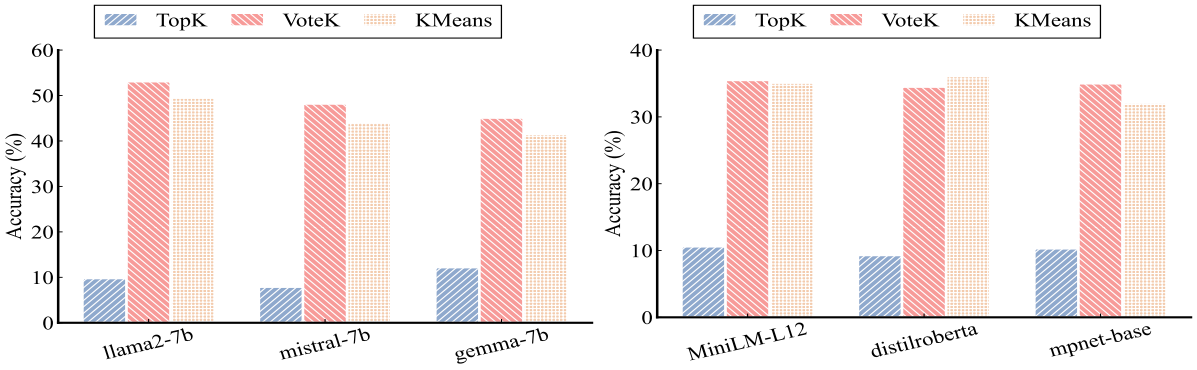
(a) Different inference models (b) Different embedding models

Figure 4: (a) Comparison with ground-truth labels on AGNews using various inference models and all-roberta-large-v1 embeddings. (b) Same comparison using various embedding models and LLaMA3-8B inference.



(a) Different inference models (b) Different embedding models

Figure 5: (a) Comparison with uncertain labels on AGNews using various inference models and all-roberta-large-v1 embeddings. (b) Same comparison using various embedding models and LLaMA3-8B inference.



(a) Different inference models (b) Different embedding models

Figure 6: (a) Comparison with corrupted labels on AGNews using various inference models and all-roberta-large-v1 embeddings. (b) Same comparison using various embedding models and LLaMA3-8B inference.

The performance of different selection strategies across various models and datasets, comparing ground-truth labels (GT), uncertain labels (U), and corrupted labels (C), is summarized in Table 8 with OpenICL (Wu et al., 2023) as the experimental framework. For demonstrations collected using the same strategy, we perform ICL with different labels and record the accuracy. Compared with ground-truth labels, the use of uncertain and corrupted labels results in a substantial performance degradation, demonstrating that label unreliability significantly affects the effectiveness of ICL.

Mod	Meth	Lbl	SST2	SST5	AGNews	Subj	CR	MNLI	QNLI	Avg.	Δ
GPT-j-6b	Random	GT	90.9(1.8)	44.7(2.6)	70.7(2.1)	73.1(1.3)	81.1(3.1)	41.2(2.0)	49.3(2.7)	64.4	-
		U	87.7(2.2)	39.1(2.9)	64.7(3.9)	56.5(3.6)	75.6(0.9)	42.0(2.6)	50.3(1.3)	59.4	↓5.0
		C	78.9(3.5)	40.3(2.0)	59.5(1.7)	43.9(3.4)	62.0(3.7)	39.7(2.5)	49.7(2.7)	53.4	↓11.0
	TopK	GT	94.2(1.3)	51.7(1.6)	86.7(1.2)	88.6(2.6)	89.3(0.8)	45.4(2.1)	53.0(2.5)	72.7	-
		U	82.7(2.3)	36.6(2.1)	54.8(1.9)	58.9(1.8)	69.4(1.7)	39.1(3.0)	49.6(2.1)	55.9	↓16.8
		C	37.6(2.9)	32.5(4.2)	17.5(2.7)	13.5(1.6)	27.3(1.5)	37.1(1.8)	48.3(1.3)	30.5	↓42.2
	BM25	GT	93.7(1.7)	47.8(3.2)	82.7(2.7)	84.3(2.4)	87.3(1.3)	42.8(3.7)	52.1(2.0)	70.1	-
		U	85.9(0.8)	39.3(3.7)	59.8(3.9)	54.7(1.9)	72.5(1.1)	37.8(4.4)	49.5(2.1)	57.1	↓13.0
		C	66.7(2.6)	35.6(3.8)	35.1(3.4)	24.7(3.3)	41.3(2.1)	37.5(3.0)	47.3(1.7)	41.2	↓28.9
	VoteK	GT	95.3(1.5)	50.6(4.7)	73.7(4.2)	72.7(7.0)	70.8(5.0)	42.1(5.3)	54.8(4.1)	65.7	-
		U	87.7(4.1)	41.6(3.4)	63.6(3.5)	60.4(4.7)	70.5(4.5)	39.9(2.2)	49.2(2.4)	59.0	↓6.7
		C	79.5(10.9)	37.5(3.7)	61.7(4.4)	48.2(3.9)	35.5(1.9)	41.1(3.5)	48.8(2.3)	50.3	↓15.4
	KMeans	GT	94.7(1.7)	42.9(5.6)	73.9(3.2)	80.3(9.3)	77.6(11.0)	40.4(6.2)	50.9(4.3)	65.8	-
		U	89.4(2.3)	38.5(3.2)	64.9(3.8)	59.9(2.3)	73.6(5.7)	42.0(4.7)	48.3(2.3)	59.5	↓6.3
		C	79.0(11.9)	36.5(3.8)	61.7(4.5)	35.7(7.9)	52.2(20.8)	41.0(1.8)	48.9(1.8)	50.7	↓15.1
LlaMA3-8b	Random	GT	95.7(0.9)	45.5(4.5)	84.0(2.1)	89.8(1.7)	89.3(0.9)	57.0(3.6)	55.0(2.0)	73.8	-
		U	85.2(2.8)	35.6(4.4)	49.1(1.4)	62.9(2.6)	72.2(1.5)	52.1(1.8)	54.7(3.1)	58.8	↓15.0
		C	54.9(1.1)	31.5(3.1)	36.0(2.4)	28.5(1.6)	48.2(4.1)	47.3(2.8)	50.9(4.5)	42.5	↓31.3
	TopK	GT	96.6(0.8)	51.8(2.6)	90.6(1.3)	95.5(1.0)	92.3(1.3)	57.8(2.5)	59.0(1.6)	77.7	-
		U	75.2(2.2)	31.7(3.2)	35.1(3.2)	65.8(5.0)	68.7(4.1)	48.3(2.7)	53.6(1.7)	54.1	↓23.6
		C	17.0(1.7)	24.7(2.1)	9.7(0.8)	8.8(1.7)	21.1(0.8)	44.2(2.6)	45.8(3.4)	24.5	↓53.2
	BM25	GT	95.7(0.8)	48.9(2.3)	88.6(1.2)	93.2(0.5)	92.1(1.7)	58.3(2.5)	56.9(1.8)	76.2	-
		U	81.7(2.6)	33.3(1.6)	41.5(2.9)	63.5(2.7)	70.7(0.8)	50.9(3.1)	51.4(3.5)	56.1	↓20.1
		C	41.3(2.9)	27.9(1.9)	19.1(1.6)	13.2(1.9)	29.7(2.4)	46.1(3.0)	46.2(2.5)	31.9	↓44.3
	VoteK	GT	96.9(1.5)	51.6(2.6)	82.5(6.6)	94.9(2.2)	86.0(3.8)	55.9(7.3)	53.7(5.8)	74.5	-
		U	80.7(4.4)	34.1(2.9)	52.1(6.1)	65.7(3.8)	71.3(3.1)	49.9(3.3)	51.5(0.4)	57.9	↓16.6
		C	30.1(15.4)	27.3(3.8)	39.3(11.9)	24.1(5.3)	35.5(1.9)	48.3(1.9)	48.9(2.2)	36.2	↓38.3
	KMeans	GT	96.7(0.8)	45.2(4.9)	86.1(4.8)	87.3(16.0)	90.2(3.1)	54.9(4.2)	57.6(6.6)	74.0	-
		U	78.8(4.4)	35.5(3.5)	48.9(2.4)	66.0(2.3)	76.1(2.9)	50.8(3.7)	51.7(1.6)	58.3	↓15.7
		C	22.9(12.8)	30.8(4.9)	33.5(3.2)	24.0(13.4)	45.9(16.9)	46.9(6.3)	47.1(2.5)	35.9	↓38.1

Table 8: Comparison of different methods on models with ground-truth labels (GT), uncertain labels (U), and corrupted labels (C), reporting accuracy and standard deviation, where Δ indicates the decrease relative to GT.

F Measurement Metrics

In the experiments shown in Figures 2 and 3, we measure three metrics—accuracy, likelihood, and information entropy. Here we provide likelihood and information entropy's mathematical definitions.

Likelihood represents the probability that the ICL assigns to the correct label $\mathbf{y}^*$.

$$\tilde{P}(\mathbf{y}^* \mid C, x) = \frac{P(\mathbf{y}^* \mid C, x)}{\sum_{y \in \mathcal{Y}} P(y \mid C, x)} \tag{18}$$

Here, we use the normalized probability and report the average value.

Information entropy reflects the degree of confidence that an LLM has in its own predictions. Its mathematical definition is given as follows.

$$H(y \mid C, x) = -\sum_{y \in \mathcal{Y}} \tilde{P}(y \mid C, x) \log \tilde{P}(y \mid C, x) \tag{19}$$

The higher the information entropy, the more uncertain the LLM is; the lower the entropy, the more confident the model is.