
On the Noise Robustness of In-Context Learning for Text Generation

Hongfu Gao^{1,2*}, Feipeng Zhang², Wenyu Jiang^{1,3}, Jun Shu⁴, Feng Zheng⁵, Hongxin Wei^{1†}

¹Department of Statistics and Data Science, Southern University of Science and Technology

²School of Economics and Finance, Xi'an Jiaotong University

³National Key Laboratory for Novel Software Technology, Nanjing University

⁴School of Mathematics and Statistics, Xi'an Jiaotong University

⁵Department of Computer Science and Engineering, Southern University of Science and Technology

Abstract

Large language models (LLMs) have shown impressive performance on downstream tasks by in-context learning (ICL), which heavily relies on the quality of demonstrations selected from a large set of annotated examples. Recent works claim that in-context learning is robust to noisy demonstrations in text classification. In this work, we show that, on text generation tasks, noisy annotations significantly hurt the performance of in-context learning. To circumvent the issue, we propose a simple and effective approach called Local Perplexity Ranking (LPR), which replaces the “noisy” candidates with their nearest neighbors that are more likely to be clean. Our method is motivated by analyzing the perplexity deviation caused by noisy labels and decomposing perplexity into inherent perplexity and matching perplexity. Our key idea behind LPR is thus to decouple the matching perplexity by performing the ranking among the neighbors in semantic space. Our approach can prevent the selected demonstrations from including mismatched input-label pairs while preserving the effectiveness of the original selection methods. Extensive experiments demonstrate the effectiveness of LPR, improving the EM score by up to 18.75 on common benchmarks with noisy annotations. Our code is available at <https://github.com/ml-stat-Sustech/Local-Perplexity-Ranking>

1 Introduction

Large language models (LLMs) have shown remarkable performance on downstream tasks by *in-context learning* (ICL) with only a few task demonstrations [7, 10]. Without requiring explicit parameter updates, in-context learning consistently outperforms zero-shot inference on various tasks (e.g., classification and generation), making it a compelling alternative to supervised fine-tuning [13, 16]. In particular, the success of ICL heavily relies on the quality of demonstrations selected from a large set of annotated examples [21, 29, 51, 60]. For those candidates, input-label mappings solicited from humans [61, 73] or LLMs [58] can often be noisy, especially in complex tasks. This gives rise to the importance of *noise-robust ICL*, which aims to construct effective demonstrations in the presence of noisy and erroneous labels.

Previous works show that in-context learning on classification tasks is fairly robust to label noise in the in-context demonstrations [9, 12, 32, 37, 54, 55]. However, it is still mysterious how noisy labels affect the performance of ICL on text generation tasks. In this work, we present the first study on in-context learning with a *noisy* annotated dataset for generation. Surprisingly, we empirically find that label noise in the demonstrations significantly degrades ICL’s performance on generation tasks,

*Work done while working at SUSTech as a visiting scholar.

†Corresponding author (weihx@sustech.edu.cn)

which is different from previous results on classification. Moreover, increasing the number of selected demonstrations with a fixed noise rate or utilizing more effective selection methods (e.g., TopK [28] and DPP [62]) will intensify the negative effect of noisy labels. This motivates our method, which can universally improve the noise robustness of existing selection methods for in-context learning.

In this paper, we show that the issue of noisy annotations can be mitigated through the perplexity ranking of noisy candidates (i.e., input-label pairs) during selection. Our method, Local Perplexity Ranking (dubbed **LPR**), is motivated by our analysis of the perplexity deviation caused by noisy labels (i.e., incorrect answers). We find that wrong answers generally result in a higher perplexity of large language models compared to correct ones, in response to the same question. To explain this phenomenon, we decompose the perplexity into two components: inherent perplexity, which measures the task complexity of the question and the correct answer, and matching perplexity, which assesses the perplexity deviation caused by noisy outputs.

Therefore, our key idea behind Local Perplexity Ranking is to decouple the matching perplexity by performing the ranking among the neighbors in semantic space. This can be achieved by ranking candidates' perplexity alongside their nearest neighborhoods, which usually have similar levels of inherent perplexity. In particular, we replace each low-rank candidate selected by existing methods (e.g., random, TopK, and DPP) with its nearest neighbor that is highly ranked. In effect, our LPR strategy can prevent the selected demonstrations from containing mismatched input-label pairs while preserving the effectiveness of the original selection methods. In this way, we ensure the correctness and relevancy of demonstrations, thereby improving the noise-tolerant ability of in-context learning.

To verify the effectiveness of our method, we conduct extensive evaluations on six text generation datasets, including NQ [22], WebQ [5], SQuAD [46], SCIQ [56], GeoQuery [39] and NL2Bash [27] datasets. The results demonstrate that local perplexity ranking can largely improve the noise-robustness of all existing selection methods under irrelevant and relevant noises. For example, on SCIQ with 60% irrelevant label noise, LPR improves the exact match score of the TopK method from 29.31 to 48.06 – a significant direct improvement of **18.75**. Moreover, our method can be easily adopted in practice. The performance of LPR is insensitive to the hyperparameters, including the threshold γ and the number of local neighbors k . This approach can effectively generalize to various LLMs to improve their noise-robustness with in-context learning.

Our contributions are summarized as follows:

- We present the first study to show that annotation quality is crucial for in-context learning in text generation, where noisy annotations significantly hurt the performance. Increasing the set size of demonstrations cannot bridge the gap, as well as picking other selection methods.
- We propose Local Perplexity Ranking (LPR), a simple and effective method to enhance the noise robustness of in-context learning. The key idea is to decouple the matching perplexity by performing the ranking among the neighbors of each candidate in semantic space.
- We empirically show that LPR can improve the noise robustness of existing demonstration selection methods in ICL across various types of label noise. In addition to text generation, we also validate the effectiveness of our method in text classification tasks.

2 Preliminary

2.1 In-context learning for generation

We consider in-context learning (ICL) of large language models (LLMs) in generation tasks, where we aim to generate text outputs $\mathbf{y} = (y_1, \dots, y_{|\mathbf{y}|})$ (i.e., token sequences) conditioned on the inputs $\mathbf{x} = (x_1, \dots, x_{|\mathbf{x}|})$ and the context C_K . In particular, the context $C_K = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^K$ contains K task demonstrations (e.g., input-output pairs), selected from a large annotated dataset with N examples $\mathcal{D} = \{(\mathbf{x}_j, \mathbf{y}_j)\}_{j=1}^N$. Given a new test input text \mathbf{x}_{test} , we make the generation of output \mathbf{y}_{test} via large language models as

$$\mathbf{y}_{test} \sim \mathcal{P}_{LLM}(\mathbf{y}_{test} \mid \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^K, \mathbf{x}_{test}), \quad (1)$$

where \sim refers to decoding strategies (e.g. greedy decoding and nuclear sampling [17, 62]). Generation with the ICL procedure is especially attractive as it does not require the parameter updating of large language models, which is often expensive and impractical.

Table 1: An illustration of the effect of three different types of annotated dataset for in-context learning. The middle column is in-context demonstrations, and the last column is the Llama2-7B [49] model prediction. The model tends to learn the label of the demonstration.

Test Input	Support:	All forms of life are built of at least one cell. A cell is the basic unit of the structure and function of living things.	Prediction
	Question:	What are the smallest structural and functional units of all living organisms?	
Setting	In-Context Demonstration		Prediction
Clean	Support: Cells are organized into tissues, tissues are organized into organs.	Question: What is considered the smallest unit of the organ?	Cells
Irrelevant	Output: Cells	Support: Cells are organized into tissues, tissues are organized into organs.	Earth
Relevant	Question: What is considered the smallest unit of the organ?	Output: Earth	tissues
	Support: Cells are organized into tissues, tissues are organized into organs.	Output: tissues	

Existing studies show that the selection strategy of demonstration plays a crucial role in the ICL performance [25, 31, 43, 44, 47]. A naive method is to randomly sample the demonstrations from annotated examples without repetition [36]. To introduce the relevancy, TopK [28] proposes to select the closest examples to the test input in the embedding space

$$C_K = \mathcal{R}_K(\mathbf{x}_{test}) = \text{TopK}_{\mathbf{x}}(s(\mathbf{x}_{test}, \mathbf{x})),$$

where \mathcal{R} is a retriever, $s(\mathbf{x}_{test}, \mathbf{x})$ denotes the cosine similarity score between \mathbf{x}_{test} and examples \mathbf{x} from the annotated dataset. We use TopK to denote the top K examples ranked by the score.

These selection strategies focus on the inputs of demonstrations, assuming that all examples are labeled correctly in the large dataset [28, 36, 62]. However, collecting a large-scale dataset with perfectly correct labels is challenging and expensive, especially for generation tasks [2, 64]. In practice, researchers often use crowdsourcing [61, 73] or large language models (LLMs) [58] such as GPT-4 [38] to create input-output pairs for new tasks, which inevitably leads to some mistakes in the annotations. This motivates us to analyze the issue of label quality in ICL for generation tasks.

2.2 Setting of noisy annotations

Given a large-scale dataset with noisy annotations $\tilde{\mathcal{D}} = \{(\mathbf{x}_j, \tilde{\mathbf{y}}_j)\}_{j=1}^N$, the selected demonstration might contain mismatched input-output pairs $(\mathbf{x}, \tilde{\mathbf{y}})$, i.e., the output $\tilde{\mathbf{y}}$ might be not a correct answer to the input \mathbf{x} . Conditioned on the noisy demonstrations, the generation of output via ICL is made as

$$\mathbf{y}_{test} \sim \mathcal{P}_{LLM}(\mathbf{y}_{test} \mid \{(\mathbf{x}_i, \tilde{\mathbf{y}}_i)\}_{i=1}^K, \mathbf{x}_{test}). \quad (2)$$

In the real world, noisy annotations may arise from unintentional mistakes or limited knowledge, resulting in various types of noise in the demonstrations. In this work, we define two categories of noisy annotations based on the input-output relevance, as follows:

Irrelevant noise assumes that the generation of noisy annotations is conditionally independent of inputs. For example, crowdsource workers may make mistakes accidentally, introducing random words or sentences in annotations. This can be simulated by reconstructing the output with random words from a subset that does not contain tokens presented in the original input-output pairs.

Relevant noise is a more realistic setting where the corrupted output is relevant to the inputs despite its incorrectness. This type of corruption may occur due to the limited knowledge of annotators and LLMs. We simulate the relevant noise by generating related yet incorrect outputs using ChatGPT-4.

In Table 1, we present an ICL example of question answering (QA) tasks to illustrate the difference between the two noisy settings. In this example, the clean annotation for the test input is “Cells”. For noisy annotations, the irrelevant noise is randomly sampled as “Earth”, while the relevant noise “tissues” exists in the support of in-context demonstration. We proceed by analyzing the empirical effects of noisy annotations in generation tasks.

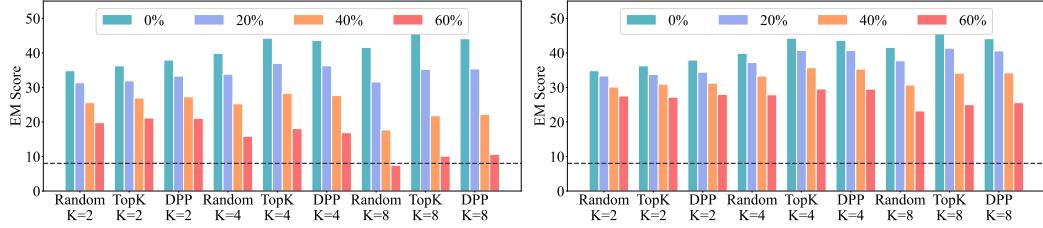


Figure 1: Average ICL performance with noisy annotations in various generation tasks across different demonstration settings. Both the two types of noises significantly deteriorate the performance of in-context learning on text generation tasks. The black line denotes zero-shot performance.

3 Empirical study of noisy ICL in text generation

In this section, we investigate the impact of noisy annotations on the performance of in-context learning for text generation. In particular, we conduct experiments on three types of generation tasks, including: question answering (NQ [22], WebQ [5]), reading comprehension (SQuAD [46], SCIQ [56]), code generation (GeoQuery [39], NL2Bash [27]). To simulate the noise, we generate noisy annotations with a pre-defined probability (e.g., 20%, 40%, 60%) in the annotated datasets. We use the output of an input from a different generation task as irrelevant noise, and adopt ChatGPT-4 to generate relevant yet false outputs as relevant noise. Furthermore, we compare the performance of noisy ICL with demonstrations across various set sizes (e.g., 2, 4, 8) and selection methods, including Random [36], TopK [28] and DPP [62]. Following previous work [16, 28, 62], we report the average Exact Match (EM) score with Llama2-7B [49].

ICL is not robust to noisy annotations in text generation. Figure 1 presents the empirical results of ICL methods with noisy annotations. The results show that both the two types of noises significantly deteriorate the performance of in-context learning on text generation tasks, which is different from the observations of ICL on classification tasks [9, 12, 32, 37, 54, 55]. In particular, a higher noise rate in annotated datasets leads to poorer performance of in-context learning. Moreover, irrelevant noises have a more negative influence than relevant noises, which may benefit the inference in the way of *task recognition* [40].

The impact of demonstration selection. To provide a deep understanding of noisy annotations, we analyze the performance of noisy ICL across different demonstration settings, including the set size (i.e., K) and selection methods. Results in Figure 1 show that, under the noisy settings, selecting a larger set of demonstrations does not enhance — and may even worsen — the performance of text generation. For example, the ICL performances with $K = 8$ are basically lower than those with $K = 2$, which is inconsistent with the clean setting. In addition, the advantages of those powerful selection methods (i.e., TopK and DPP) are neutralized in the presence of noisy annotations.

Through the empirical analysis, we find that noisy annotations significantly hurt the performance of ICL in text generation tasks. More importantly, increasing the set size of demonstrations cannot bridge the gap, as well as picking an existing selection method, like DPP. This motivates us to design *noise-robust* methods, which can universally improve the noise robustness of in-context learning.

4 Methodology

In this section, we first analyze the perplexity deviation caused by noisy annotations and introduce the disentanglement of perplexity to explain the phenomenon. In light of this, we propose a novel method – local perplexity ranking – to improve the noise robustness of in-context learning for text generation. Our method can be easily incorporated into existing methods of demonstration selection.

4.1 Perplexity deviation of noisy annotations

For language models, perplexity measures the degree of uncertainty in generating new tokens. In particular, a low perplexity indicates that the model makes the prediction with high confidence.

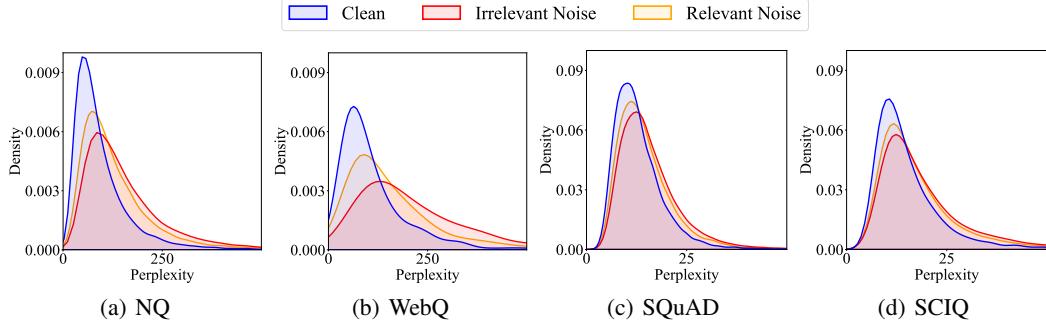


Figure 2: The distribution of perplexity of Llama2-7B [49] on clean and noisy annotations. Examples with noisy annotations indeed obtain higher perplexity than those with clean annotations.

Therefore, perplexity is commonly used to evaluate the language quality of generated content, e.g., detecting attack prompts [3], out-of-distribution instances [4, 57], hard-to-learn instances [13], and corrupted instances [64]. In light of this, we conjecture that mismatched input-output pairs may result in higher perplexity of LLMs due to their low co-occurrence rate. For instance, in the example presented in Table 1, the term “earth” rarely co-occurs with “cells” and “organ”, so LLMs are more likely to exhibit high perplexity in the input-output pair.

Empirical study To validate this assumption, we compare the perplexity of clean and noisy annotations in text generation tasks. Specifically, we concatenate each tokenized input-output pair (\mathbf{x}, \mathbf{y}) , and obtain the corresponding tokenized sequence $\mathbf{z} = (z_1, \dots, z_{|\mathbf{z}|}) = (x_1, \dots, x_{|\mathbf{x}|}, y_1, \dots, y_{|\mathbf{y}|})$, where $|\mathbf{z}| = |\mathbf{x}| + |\mathbf{y}|$. Now, the perplexity of \mathbf{z} is calculated as:

$$\text{Perplexity}(\mathbf{z}) = \exp\left\{-\frac{1}{|\mathbf{z}|} \sum_{i=1}^{|\mathbf{z}|} \log p_{\theta}(z_i | z_{<i})\right\}, \quad (3)$$

where $\log p_{\theta}(z_i | z_{<i})$ is the log-likelihood of the i -th token conditioned on the preceding tokens $z_{<i}$, from the given language model parameterized by θ .

In Figure 2, we present the perplexity distribution of Llama2-7B [49] on clean and noisy annotations of four datasets. The results illustrate that examples with noisy annotations indeed obtain higher perplexity than those with clean annotations, which confirms our assumption. In particular, relevant noises achieve slightly lower perplexity than irrelevant noises since relevant outputs are close to the inputs despite their erroneous information. However, the deviation of the perplexity distribution caused by noisy annotations is marginal, making it suboptimal to differentiate noisy annotations from clean ones. In the following, we explain this phenomenon with the disentanglement of perplexity.

Disentanglement of perplexity Given an input-output pair, the perplexity of large language models (LLMs) stems not only from how well the output matches the input, but also from the inherent complexity of the task. For example, a mathematical question with a correct answer can exhibit a higher perplexity than a question of daily life with an incorrect answer. Informally, we decompose the overall Perplexity into two components³, as shown below:

$$\text{Perplexity} = \text{Inherent Perplexity} + \text{Matching Perplexity}$$

Here, the inherent perplexity measures how the model is familiar with the task (i.e., the input and the correct output). The matching perplexity quantifies the perplexity deviation caused by noisy outputs, so it can be zero with correct outputs. A higher matching perplexity indicates that the output is more likely to be incorrect for the input. However, directly computing the matching perplexity is non-trivial as clean outputs are unknown. To circumvent the issue, we aim to design an effective method to decouple the matching perplexity from the overall perplexity.

³This disentanglement is conceptual rather than mathematical.

4.2 Local Perplexity Ranking

Intuition Motivated by the previous analysis, we propose *local perplexity ranking* (LPR), a general strategy that can improve the noise robustness of in-context learning. Our key idea is to decouple the matching perplexity by performing the ranking among the neighbors in semantic space. Here, our approach is built on two natural assumptions that are naturally satisfied in the real world:

1. The clean annotations are the majority in the annotated dataset.
2. Examples that are semantically similar share the same level of inherent perplexity.

In the literature, Assumption 2 is also supported by previous findings that paragraphs whose representations are close to each other share the same intrinsic task [14, 28, 73]. With the two assumptions, we can approximate the inherent perplexity of a candidate through its neighbors, where most examples are correctly annotated. In other words, the candidate is more likely to be wrongly annotated if its perplexity is relatively higher than its neighbors, and vice versa. With this in mind, we present the details of our approach in the following.

Finding the local neighbors Given a test input, we first sample a candidate set \tilde{C} with a pre-defined selection strategy, such as Random [36], TopK [28] or DPP [62]. For each candidate \mathbf{z}^* , we adopt k -Nearest-Neighbors (k -NN) to find its local neighbors that are close to the candidate in token space. Formally, the k local neighbors are obtained as: $N_k(\mathbf{z}^*) = \{\mathbf{z}_{\pi(1)}, \mathbf{z}_{\pi(2)}, \dots, \mathbf{z}_{\pi(k)}\}$, where $\pi(i)$ is the index of the example with the i -th smallest distance to the candidate. In particular, we use the cosine similarity score to measure the distance between the candidate \mathbf{z}^* and other examples \mathbf{z} :

$$\cos(\mathbf{z}_i, \mathbf{z}^*) = \frac{\mathbf{z}_i^\top \mathbf{z}^*}{\|\mathbf{z}_i\|_2 \|\mathbf{z}^*\|_2}.$$

Ranking the perplexity As discussed above, the local neighbors share the same level of inherent perplexity, which enables the comparison of their matching perplexity. For each candidate \mathbf{z}^* , we propose to rank the perplexity of examples in the cluster of local neighbors $\mathbf{z}^* \cup N_k(\mathbf{z}^*)$. Formally, we first sort all examples in the cluster in increasing order by the perplexity and obtain the original indices for the sorted scores as:

$$\mathcal{I} = \text{argsort}\{\text{Perplexity}(\mathbf{z}_n)\}_{n=1}^{k+1}, \quad \mathbf{z}_n \in (\mathbf{z}^* \cup N_k(\mathbf{z}^*)), \quad (4)$$

where $\text{Perplexity}(\cdot)$ is the overall perplexity defined in Equation 3. In this way, the high-ranking examples are more likely to be correctly annotated than the low-ranking example in the sorted list \mathcal{I} .

Substituting the noisy candidates To build the final demonstration set, we propose to replace the noisy candidates with their nearest neighbors that are more likely to be clean. In particular, we can determine whether a candidate should be replaced by:

$$g(\mathbf{z}_n) = \mathbb{1}\left(\frac{\text{Loc}(\mathbf{z}_n, \mathcal{I})}{k+1} \geq \gamma\right), \quad (5)$$

where γ is the pre-defined threshold (e.g., 50%), $\mathbb{1}(\cdot)$ is the indicator function and $\text{Loc}(\mathbf{z}_n, \mathcal{I})$ return the index of \mathbf{z}_n in the sorted list \mathcal{I} . It is worth noting that the proposed method is not sensitive to the value of the hyperparameter γ , as shown in Subsection 5.1. Then, for those candidates with $g(\mathbf{z}_n)$, we pick the substitutes from their neighbors by:

$$\min\{i \in N^k | g(\mathbf{z}_{\pi(i)}) = 0\},$$

where $\pi(i)$ is the index of the example with the i -th smallest distance to the candidate. After the replacement, we establish the final demonstration set for in-context learning. Noticeably, our method offers several compelling advantages:

- **Algorithm-agnostic:** LPR can be easily incorporated into existing demonstration selection methods, consistently improving the robustness against noisy annotations.
- **Easy to use:** LPR does not require heavy hyperparameter tuning, as it is insensitive to the threshold value (see Figure 3). LPR does not introduce much computational cost due to the efficient computation of perplexity (see Table 4).

Table 2: Main results on various datasets. The bold indicates the improvement by integrating LPR.

Dataset	Method	Clean			Irrelevant Noise			Relevant Noise		
		0%	20%	40%	60%	20%	40%	60%	20%	40%
NQ	Random	14.51 \pm 0.51	10.97 \pm 0.29	7.37 \pm 0.45	4.23 \pm 0.46	12.00 \pm 0.65	9.67 \pm 0.45	6.40 \pm 1.02		
	+Ours	15.05\pm0.10	13.31\pm0.25	11.51\pm0.51	8.87\pm0.74	13.74\pm0.12	13.28\pm0.33	9.43\pm0.52		
	TopK	20.25 \pm 0.10	13.95 \pm 1.14	9.97 \pm 1.13	5.90 \pm 1.08	16.21 \pm 0.22	12.22 \pm 0.22	8.50 \pm 0.28		
	+Ours	19.19 \pm 0.19	17.15\pm0.50	13.54\pm0.41	9.64\pm0.25	17.25\pm0.69	14.82\pm0.51	11.98\pm0.60		
	DPP	20.35 \pm 0.76	14.69 \pm 0.94	9.87 \pm 0.49	5.97 \pm 0.48	15.47 \pm 1.00	11.28 \pm 0.42	7.89 \pm 0.25		
	+Ours	19.68 \pm 0.33	16.59\pm0.45	13.31\pm0.57	11.18\pm0.50	16.79\pm0.47	14.91\pm0.18	11.94\pm0.91		
WebQ	Random	20.37 \pm 0.64	15.18 \pm 1.06	10.39 \pm 0.83	4.83 \pm 0.17	18.29 \pm 0.43	15.92 \pm 0.68	13.50 \pm 0.17		
	+Ours	21.94\pm0.64	20.32\pm0.92	16.33\pm0.58	12.54\pm0.29	21.51\pm0.33	19.33\pm0.41	16.69\pm1.11		
	TopK	30.16 \pm 0.58	22.52 \pm 0.64	14.52 \pm 0.78	8.00 \pm 1.12	27.19 \pm 0.27	22.82 \pm 0.75	18.88 \pm 1.09		
	+Ours	29.24 \pm 0.34	26.55\pm0.24	21.67\pm1.28	14.54\pm1.02	28.49\pm0.43	25.44\pm0.68	21.28\pm0.12		
	DPP	29.40 \pm 0.39	22.11 \pm 0.81	13.72 \pm 0.27	7.33 \pm 0.68	26.18 \pm 1.04	21.53 \pm 0.61	16.80 \pm 0.17		
	+Ours	29.92\pm0.48	26.57\pm0.95	21.94\pm1.05	14.85\pm0.81	28.46\pm1.01	25.61\pm0.78	21.35\pm1.17		
SQuAD	Random	56.50 \pm 0.57	50.00 \pm 0.62	39.10 \pm 0.88	26.20 \pm 0.79	53.90 \pm 0.65	49.17 \pm 0.62	42.03 \pm 0.79		
	+Ours	57.73\pm0.79	56.87\pm0.47	48.50\pm0.86	43.00\pm0.86	57.70\pm1.31	53.93\pm0.33	47.93\pm0.48		
	TopK	56.97 \pm 0.69	51.83 \pm 1.03	42.83 \pm 1.68	29.10 \pm 2.92	54.77 \pm 0.69	49.37 \pm 1.37	41.37 \pm 2.09		
	+Ours	57.27\pm0.62	55.40\pm0.37	51.43\pm1.26	41.30\pm2.65	56.90\pm0.64	53.90\pm1.08	48.37\pm0.66		
	DPP	57.29 \pm 0.87	50.57 \pm 0.33	41.63 \pm 1.00	25.67 \pm 2.52	56.10 \pm 0.59	49.57 \pm 1.24	43.37 \pm 0.78		
	+Ours	58.10\pm0.29	56.73\pm0.61	52.53\pm0.33	42.93\pm0.88	57.50\pm0.54	55.90\pm0.18	50.77\pm0.39		
SCIQ	Random	68.15 \pm 0.28	59.19 \pm 1.57	44.19 \pm 2.89	28.21 \pm 2.96	64.59 \pm 1.42	58.39 \pm 0.16	49.54 \pm 0.80		
	+Ours	67.93 \pm 0.85	65.06\pm1.34	55.57\pm0.53	42.00\pm2.96	66.63\pm0.94	62.70\pm1.10	58.92\pm1.74		
	TopK	68.62 \pm 1.13	59.59 \pm 1.28	45.77 \pm 2.68	29.31 \pm 1.73	64.66 \pm 1.34	58.54 \pm 0.12	49.47 \pm 0.65		
	+Ours	70.06\pm0.32	66.67\pm0.81	57.44\pm1.04	48.06\pm1.53	67.76\pm0.50	63.96\pm1.71	56.32\pm2.18		
	DPP	67.29 \pm 0.35	57.69 \pm 1.83	45.34 \pm 1.56	28.50 \pm 1.78	64.88 \pm 0.43	58.91 \pm 0.64	50.00 \pm 0.85		
	+Ours	70.57\pm0.45	67.86\pm1.43	59.65\pm2.11	45.46\pm2.72	69.16\pm0.98	65.63\pm0.21	56.72\pm1.37		
GeoQuery	Random	27.97 \pm 0.99	23.18 \pm 0.62	17.44 \pm 1.56	14.10 \pm 0.74	26.48 \pm 0.17	26.13 \pm 0.05	26.25 \pm 0.40		
	+Ours	27.27 \pm 0.36	27.12\pm0.69	25.52\pm1.02	22.23\pm0.67	27.43\pm0.71	27.01\pm0.05	26.73\pm0.90		
	TopK	44.17 \pm 0.09	27.28 \pm 2.65	17.49 \pm 2.05	9.96 \pm 3.08	41.31 \pm 0.46	38.48 \pm 0.63	34.90 \pm 0.69		
	+Ours	43.32 \pm 0.05	42.25\pm1.00	33.80\pm1.43	24.39\pm1.08	42.59\pm0.37	39.40\pm0.37	37.74\pm1.23		
	DPP	45.81 \pm 0.71	31.79 \pm 5.93	21.54 \pm 3.36	10.61 \pm 0.15	42.97 \pm 1.96	39.91 \pm 0.42	33.34 \pm 0.53		
	+Ours	44.18 \pm 0.47	43.01\pm0.02	40.94\pm0.91	33.25\pm1.27	41.49 \pm 0.11	40.62\pm0.06	36.81\pm0.61		
NL2Bash	Random	27.91 \pm 0.37	25.37 \pm 0.21	15.77 \pm 0.91	8.95 \pm 0.65	27.20 \pm 1.06	28.09 \pm 0.51	26.27 \pm 0.56		
	+Ours	29.93\pm1.18	29.09\pm0.26	26.04\pm2.05	22.92\pm0.39	29.01\pm0.36	28.92\pm0.07	26.80\pm0.55		
	TopK	35.71 \pm 0.42	27.40 \pm 0.26	20.00 \pm 0.62	9.95 \pm 0.68	32.57 \pm 0.13	30.21 \pm 0.08	27.48 \pm 0.35		
	+Ours	33.92 \pm 0.70	32.51\pm1.59	30.50\pm1.02	23.47\pm1.52	31.33 \pm 0.04	31.39\pm1.70	29.49\pm0.06		
	DPP	37.77 \pm 0.02	31.52 \pm 0.12	23.23 \pm 0.34	11.16 \pm 2.14	32.74 \pm 0.29	32.56 \pm 0.61	26.72 \pm 1.58		
	+Ours	35.85 \pm 1.51	32.27\pm0.99	32.47\pm0.40	27.84\pm1.17	33.63\pm0.23	32.53 \pm 0.57	28.96\pm0.98		

5 Experiments

5.1 Experimental Setup

Datasets. We employ 6 generation datasets for the evaluations, including **Open-Domain Question-Answering**: NQ [22], WebQ [5]; **Reading Comprehension**: SQuAD [46] and SCIQ [56]; **Code Generation**: GeoQuery [39] and NL2Bash [27]. Due to limited space, these tasks' input/output, statistics, split and evaluation metrics are reported in Appendix A.2.

Models and ICL methods. For the main results, we use Llama-2-7B-Chat [49] as the LLM throughout our experiments. We also provide experiments on other models including Llama2-13B-Chat [49], Mistral-7B [19] and OPT-6.7B [66]. We use bert-base-uncased sentence encoder as the similarity tokenizer [11, 62]. We conduct experiments with existing demonstration selection methods, including **Random** [36], **TopK** [28] and **DPP** [62]. For hyperparameters, we set the number of neighbors $k = 4$ and the threshold $\gamma = 50\%$ by default. The details of our implementation is presented in Appendix A.2.

5.2 Main Results

Can LPR improve the noise-robustness of in-context learning? Table 2 presents the average in-context learning performance of the baselines and our method on six generation tasks, under various types of noisy annotations. A salient observation is that our method drastically improves the noise-robustness performance of the existing demonstration selection methods by employing LPR.

Table 3: Average test performance of the baselines and our method using varying large language models across various noise types. The results are shown as Naive/+Ours. The bold indicates the improved results by integrating LPR.

Method	Clean		Irrelevant Noise				Relevant Noise		
	0%	20%	40%	60%	20%	40%	60%		
Llama2-13B [49]	45.13/ 45.27	38.58/ 43.47	29.00/ 39.24	18.93/ 30.46	42.18/ 44.32	37.10/ 41.88	30.67/ 36.76		
Mistral-7B [19]	34.89/34.12	32.12/ 33.59	26.28/ 31.56	19.24/ 27.03	33.43/ 33.91	30.52/ 32.64	26.63/ 30.00		
OPT-6.7B [66]	23.46/ 24.03	17.26/ 21.31	11.32/ 17.29	7.68/ 12.91	20.16/ 22.40	17.58/ 20.22	14.95/ 17.52		

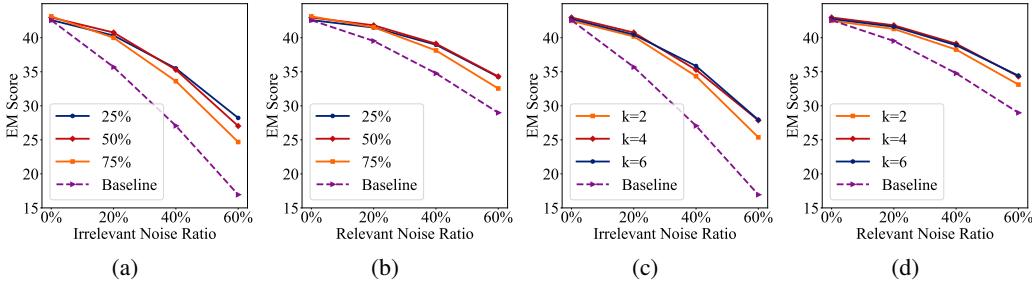


Figure 3: The average test performance with different thresholds τ and numbers of local neighbors k across various noise types. Figure (a) and (b) analyze how the hyperparameter τ affects the performance of LPR. Figure (c) and (d) illustrate the influence of the hyperparameter k .

For example, on the NQ with 60% irrelevant noise, our approach improves the EM score of the naive random selection method from 28.21 to 42.00 - a **13.79** of direct improvement. Moreover, we show that the LPR can boost performance for a wide range of existing demonstration selection methods such as TopK [28] and DPP [62]. For example, we observe that, on SCIQ with 60% irrelevant label noise, LPR improves the exact match score of the TopK method from 29.31 to 48.06 – a significant direct improvement of **18.75**. Our method also establish strong robustness against all types of noisy annotations. Appendix A.3 reports the results with various demonstration sizes.

How does the threshold γ affect the noise-robustness of LPR? In Figure 3 (a) and (b), we ablate how the parameter γ in our method (cf. Eq. 5) affects the noise-robust performance. The base indicates all candidate demonstrations are selected without our method. It's noteworthy that LPR shows robustness to the choice of threshold γ , even if we set $\gamma = 75\%$ also yield significant EM score improvements. We can also observe that as the threshold γ decrease, the noise-robust performance also improve, especially under 60% noise conditions. Due to space constraints, we only report the average results of multiple baselines on various generation tasks.

Does LPR work with the different number of k nearest neighbors? We evaluate how the number of nearest neighbors k in our method affects the LPR performance. Specifically, We vary the number of neighbors $k = \{2, 4, 6\}$. As is shown in Figure 3 (c) and (d), an increase in the number of nearest neighbors beyond 0 leads to an evident improvement in EM score, and the performance starts to reach a point of saturation with the further addition of neighbors. Concernedly, more perplexity of nearest neighbors needs to be calculated as k value increase, but the improvement is limited. For simplicity, we employ a moderate range of neighbors and use $k=4$ throughout our experiments.

Is LPR effective with different LLMs? To show our proposed method is model-agnostic, we conduct experiments on a diverse collection of model architectures and present the results in Table 3. From the results, we observe that our method consistently improves the ICL performance when using Llama2-13B [49], Mistral-7B [19] and OPT-6.7B [66]. For instance, with Mistral-7B, using our method boosts the ICL performance using the random selection method from 19.24 to 27.07, an average **7.83** of direct improvement on 6 datasets with irrelevant-60% noisy annotations.

6 Discussion

Global Perplexity Ranking vs. Local Perplexity Ranking. While our method has demonstrated strong promise in in-context learning, one may also ask: *can a similar effect be achieved by selecting*

Table 4: Average test performance comparison between global perplexity ranking and local perplexity ranking. The results are shown as *Global/Local*. Bold numbers are superior results.

Method	Clean		Irrelevant Noise				Relevant Noise		Time (h)
	0%	20%	40%	60%	20%	40%	60%	Time (h)	
Random	39.32/ 40.66	38.94/38.89	34.41 /32.98	27.82 /26.59	39.23/ 39.90	36.38/ 37.31	31.76/ 33.24	2.88/ 0.55	
TopK	40.57 / 43.94	39.94/ 41.44	35.85/ 36.02	31.79 /28.38	40.33/ 42.60	38.69/ 39.53	33.88/ 34.48	3.06/ 0.57	
DPP	42.33/ 44.32	40.18/ 41.94	36.20/ 36.86	30.91 /28.60	40.42/ 42.98	38.49/ 40.51	32.24/ 35.20	3.21/ 0.64	
Average	40.74/ 42.97	39.68/ 40.76	35.49 /35.28	30.17 /27.86	39.99/ 41.83	37.85/ 39.12	32.63/ 34.31	3.05/ 0.57	

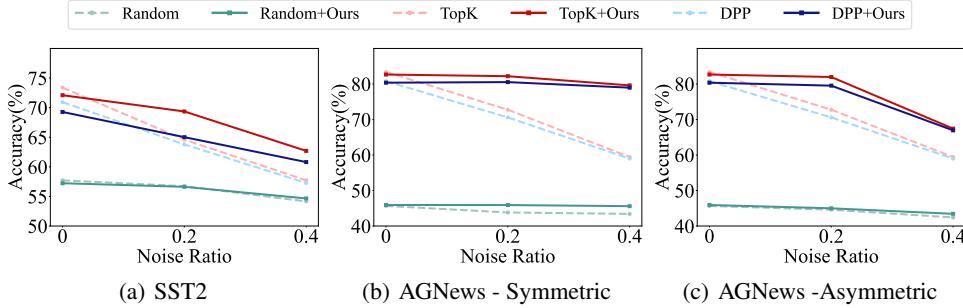


Figure 4: Average test accuracy on SST2 [48] and AGNews [67]. Different colors indicate the selection methods. The solid lines denote existing selection methods, and the dotted lines represent the method integrated by our method. We omit the noisy type on the binary classification – SST2.

demonstrations with the lowest perplexity in the whole dataset? In this ablation, we compare our method with a global perplexity ranking method that selects demonstrations with the lowest perplexity values of input-label pairs from a large candidate set (e.g., $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{100}$).

Table 4 presents the performance comparison between our method and the global perplexity ranking method. While both the two perplexity ranking methods improve the robustness of ICL against noisy annotations, the global approach obtains inferior performance compared to our proposed method in most cases, especially in the cases of clean and low noise rates. In efficiency, Table 4 also show that the local ranking approach requires only 20% of the time required by the global ranking. This is because our method only calculates the perplexity of the local neighbors for each candidate, instead of using a large candidate pool. Overall, we show that the global ranking method cannot outperform the local ranking while introducing much more computational loads.

Transfer to text classification tasks. Text classification is a common task of in-context learning, which may also suffer from a noisy annotation issue. To this end, we verify the effectiveness of the proposed method in text classification. Here, we consider two classification tasks (SST2 [48] and AGNews [67]) with popular label noise types: the symmetric noise and the asymmetric noise [8, 33]. We report the average accuracy with GPT-Neo-2.7B [6] on datasets with the two noise types. More detailed experimental settings are presented in Appendix A.2.

Figure 4 demonstrates that noise annotations barely hurt the performance of ICL when employing the random demonstration selection method [36]. However, the performance of ICL is significantly compromised when utilizing more effective selection methods like TopK [28] and DPP [62]. After integrating our method, both TopK and DPP methods are significantly improved in the inference performance, which indicates the noise robustness of our method in text classification.

Potential failure cases. Our approach is built on two assumptions that are naturally satisfied in the real world (See Section 4.2). In this section, we conduct experiments on four generation tasks, including NQ, WebQ, SCIQ, and SQuAD, to determine whether our proposed method remains effective when one of these two assumptions is dissatisfied. The detailed analysis is presented below.

Assumption 1 (Data): clean annotations are the majority in the annotated dataset. Given a dataset with extremely high noise ratios (e.g., 60%, 70%, 80%, 90%), the perplexity ranking of local neighbors may not reflect the correctness of the annotations, as most (even all) neighbors can be wrongly annotated. To explicitly show that, we conduct an experiment to validate the performance of LPR under extremely high noise ratios. The Table 5 below presents the average EM score of the baselines

Table 5: Average test performance of the baselines and our method for four generation tasks on four datasets with extremely high noise ratios (e.g., 60%, 70%, 80%, 90%). The results are shown as Naive/+Ours. The bold indicates the improved results by integrating LPR.

Method	Irrelevant Noise				Relevant Noise			
	60%	70%	80%	90%	60%	70%	80%	90%
Random	15.80/ 26.60	11.61/ 16.97	7.98/ 11.24	4.79/ 5.45	27.87/ 33.25	24.67/ 28.29	22.51/ 24.45	20.15/ 21.20
TopK	18.08/ 28.08	14.62/ 18.24	10.16/ 10.96	6.25/ 7.17	29.55/ 34.48	26.02/ 29.23	23.28/ 25.87	21.21/ 22.68
DPP	16.87/ 28.61	15.10/ 18.01	9.93/ 10.03	6.46/ 7.18	29.51/ 35.19	25.85/ 28.86	23.28/ 25.27	20.83/ 21.95

Table 6: Average test performance of the baselines and our method using varying large language models (e.g. OPT-1.3B, OPT-2.7B, OPT-6.7B [66]) across various noise types. The results are shown as Naive/+Ours. The bold indicates the improved results by integrating LPR.

Method	Clean		Irrelevant Noise				Relevant Noise	
	0%	20%	40%	60%	20%	40%	60%	60%
OPT-1.3B	13.06/ 13.22	10.48/ 10.96	8.66/ 9.63	5.95/ 6.41	12.21/ 12.58	11.33/ 11.53	10.42/ 10.81	
OPT-2.7B	15.30/ 15.70	12.68/ 13.23	10.53/ 11.45	7.01/ 9.02	14.15/ 14.73	13.21/ 14.33	11.86/ 12.85	
OPT-6.7B	23.46/ 24.03	17.26/ 21.31	11.32/ 17.29	7.68/ 12.91	20.16/ 22.40	17.58/ 20.22	14.95/ 17.52	

and our method. We use Llama2-7B [49] as the LLM throughout our experiments. The results show that the improvements of our approach decrease as the noise ratios increase. For example, when the irrelevant label noise ratio increases from 60% to 90%, the improvement of our method for the TopK method decreases from 10.26 to 0.92.

Assumption 2 (Model): examples that are semantically similar share the same level of inherent perplexity. The model affects the the performance of LPR through the concept of inherent perplexity. This assumption cannot hold if the model is not capable of precisely measuring the semantic distance between examples. In this case, the local neighbors may not share the same level of inherent perplexity so that we cannot compare the Matching Perplexity. To validate this, we conduct experiments with language models with various sizes, including OPT-1.3B, OPT-2.7B and OPT-6.7B [66]. The results in Table 6 reveal that the performance of LPR decreases as the parameter size of language models decreases. For instance, for 60% irrelevant noise, the improvement of our method decreases from 5.23 to 0.46 when the parameter size of the language model decreases from 6.7B to 1.3B.

7 Conclusion

In this paper, we introduce Local Perplexity Ranking (**LPR**), a general strategy that can universally enhance the noise robustness of in-context learning on generation tasks. To the best of our knowledge, this work is the first to analyze the noisy annotations in ICL for text generation. Our key idea is to decouple the matching perplexity by performing the ranking among the neighbors in semantic space. In particular, we replace each low-ranked candidate with its nearest neighbor that is highly ranked. Extensive experiments demonstrate that LPR can improve the noise robustness of existing demonstration selection methods in ICL across various noise types. Our approach is easy to use in practice, as it is insensitive to the hyperparameters and does not introduce heavy computational cost. **Limitations.** LPR is suboptimal in cases of high noise rates due to the assumption that clean annotations are the majority in the dataset. In addition, we do not provide a theoretical analysis to show how noisy annotations affect ICL, which will be an interesting direction for future research.

8 Acknowledgements

This research is supported by the Shenzhen Fundamental Research Program (Grant No. JCYJ20230807091809020). Feipeng Zhang is supported by the National Natural Science Foundation of China (Grant No. 72171192) and the Youth Innovation Team of Shaanxi Universities. Jun Shu is supported in part by the National Natural Science Foundation of China (Grant No. 12326606). Feng Zheng is supported in part by the National Natural Science Foundation of China (Grant No. 62122035). We gratefully acknowledge the support of the Center for Computational Science and Engineering at the Southern University of Science and Technology for our research.

References

- [1] Maha Agro and Hanan Aldarmaki. Handling realistic label noise in BERT text classification. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing*, pages 11–20, 2023.
- [2] Dmitriy Alexandrov, Anastasia Zakharova, and Nikolay Butakov. Does noise really matter? investigation into the influence of noisy labels on bert-based question answering system. *International Journal of Semantic Computing*, pages 1–20, 2024.
- [3] Gabriel Alon and Michael Kamfonas. Detecting language model attacks with perplexity. *arXiv preprint arXiv:2308.14132*, 2024.
- [4] Udit Arora, William Huang, and He He. Types of out-of-distribution texts and how to detect them. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10687–10701, 2021.
- [5] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, 2013.
- [6] Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow. *10.5281/zenodo.5297715*, 2021.
- [7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 1877–1901, 2020.
- [8] Mingcai Chen, Hao Cheng, Yuntao Du, Ming Xu, Wenyu Jiang, and Chongjun Wang. Two wrongs don't make a right: Combating confirmation bias in learning with label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14765–14773, 2023.
- [9] Chen Cheng, Xinzhi Yu, Haodong Wen, Jinsong Sun, Guanzhang Yue, Yihao Zhang, and Zeming Wei. Exploring the robustness of in-context learning with noisy labels. *arXiv preprint arXiv:2404.18191*, 2024.
- [10] Julian Coda-Forno, Marcel Binz, Zeynep Akata, Matt Botvinick, Jane Wang, and Eric Schulz. Meta-in-context learning in large language models. *Advances in Neural Information Processing Systems*, 36:65189–65201, 2023.
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.
- [12] Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. Mitigating label biases for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14014–14031, 2023.
- [13] Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. Demystifying prompts in language models via perplexity estimation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10136–10148, 2023.
- [14] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Pre-training to learn in context. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4849–4870, 2023.

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

[16] Shivanshu Gupta, Matt Gardner, and Sameer Singh. Coverage-based example selection for in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13924–13950, 2023.

[17] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

[18] Runhui Huang, Yanxin Long, Jianhua Han, Hang Xu, Xiwen Liang, Chunjing Xu, and Xiaodan Liang. Nlip: Noise-robust language-image pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 926–934, 2023.

[19] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[20] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, pages 22199–22213, 2022.

[21] Jannik Kossen, Yarin Gal, and Tom Rainforth. In-context learning learns label relationships but is not conventional learning. In *International Conference on Learning Representations*, 2024.

[22] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, pages 452–466, 2019.

[23] Itay Levy, Ben Bogin, and Jonathan Berant. 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, 2023.

[24] Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*, 2024.

[25] Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. Unified demonstration retriever for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4644–4668, 2023.

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

[27] Xi Victoria Lin, Chenglong Wang, Luke Zettlemoyer, and Michael D. Ernst. NL2Bash: A corpus and semantic parser for natural language interface to the linux operating system. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 2018.

[28] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, 2022.

[29] Quanyu Long, Yin Wu, Wenya Wang, and Sinno Jialin Pan. Decomposing label space, format and discrimination: Rethinking how llms respond and solve tasks via in-context learning. *arXiv preprint arXiv:2404.07546*, 2024.

[30] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 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)*, pages 8086–8098, 2022.

[31] Man Luo, Xin Xu, Zhuyun Dai, Panupong Pasupat, Mehran Kazemi, Chitta Baral, Vaiva Imbrasaite, and Vincent Y Zhao. Dr.icl: Demonstration-retrieved in-context learning. *arXiv preprint arXiv:2305.14128*, 2023.

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

[33] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

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

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

[36] Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, 2022.

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

[38] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024.

[39] Vaishali Pal, Andrew Yates, Evangelos Kanoulas, and Maarten de Rijke. MultiTabQA: Generating tabular answers for multi-table question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6322–6334, 2023.

[40] Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. What in-context learning “learns” in-context: Disentangling task recognition and task learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8298–8319, 2023.

[41] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[42] Gaurav Patel, Jan Allebach, and Qiang Qiu. Seq-UPS: Sequential uncertainty-aware pseudo-label selection for semi-supervised text recognition. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6169–6179, 2023.

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

[44] Chengwei Qin, Aston Zhang, Anirudh Dagar, and Wenming Ye. In-context learning with iterative demonstration selection. *arXiv preprint arXiv:2310.09881*, 2023.

[45] Linlu Qiu, Peter Shaw, Panupong Pasupat, Tianze Shi, Jonathan Herzig, Emily Pitler, Fei Sha, and Kristina Toutanova. Evaluating the impact of model scale for compositional generalization in semantic parsing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9157–9179, 2022.

[46] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, 2016.

[47] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, 2022.

[48] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 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, 2013.

[49] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

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

[51] Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. In *Advances in Neural Information Processing Systems*, pages 15614–15638, 2023.

[52] Hongxin Wei, Huiping Zhuang, Renchunzi Xie, Lei Feng, Gang Niu, Bo An, and Yixuan Li. Mitigating memorization of noisy labels by clipping the model prediction. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[53] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, pages 24824–24837, 2022.

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

[55] Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.

[56] Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, 2017.

[57] Qianhui Wu, Huiqiang Jiang, Haonan Yin, Börje Karlsson, and Chin-Yew Lin. Multi-level knowledge distillation for out-of-distribution detection in text. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7317–7332, 2023.

[58] Tongshuang Wu, Haiyi Zhu, Maya Albayrak, Alexis Axon, Amanda Bertsch, Wenxing Deng, Ziqi Ding, Bill Guo, Sireesh Gururaja, Tzu-Sheng Kuo, Jenny T. Liang, Ryan Liu, Ihita Mandal, Jeremiah Milbauer, Xiaolin Ni, Namrata Padmanabhan, Subhashini Ramkumar, Alexis Sudjianto, Jordan Taylor, Ying-Jui Tseng, Patricia Vaidos, Zhijin Wu, Wei Wu, and Chenyang

Yang. LLMs as workers in human-computational algorithms? replicating crowdsourcing pipelines with llms. *arXiv preprint arXiv:2307.10168*, 2023.

[59] Zhenyu Wu, Yaoxiang Wang, Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Jingjing Xu, and Yu Qiao. 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, 2023.

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

[61] Yan Yan, Rómer Rosales, Glenn Fung, Ramanathan Subramanian, and Jennifer Dy. Learning from multiple annotators with varying expertise. *Machine Learning*, page 291–327, 2014.

[62] Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. Compositional exemplars for in-context learning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 39818—39833, 2023.

[63] Xichen Ye, Xiaoqiang Li, songmin dai, Tong Liu, Yan Sun, and Weiqin Tong. Active negative loss functions for learning with noisy labels. In *Advances in Neural Information Processing Systems*, pages 6917–6940, 2023.

[64] Hang Zhang, Yeyun Gong, Xingwei He, Dayiheng Liu, Daya Guo, Jiancheng Lv, and Jian Guo. Noisy pair corrector for dense retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11439–11451, 2023.

[65] Shaokun Zhang, Xiaobo Xia, Zhaoqing Wang, Ling-Hao Chen, Jiale Liu, Qingyun Wu, and Tongliang Liu. IDEAL: Influence-driven selective annotations empower in-context learners in large language models. In *International Conference on Learning Representations*, 2024.

[66] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[67] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, 28, 2015.

[68] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. What makes good examples for visual in-context learning? *Advances in Neural Information Processing Systems*, 36, 2024.

[69] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. In *International Conference on Learning Representations*, 2023.

[70] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, pages 46595–46623, 2023.

[71] Yu Zhou, Hongtao Xie, Shancheng Fang, and Yongdong Zhang. Semi-supervised text detection with accurate pseudo-labels. *IEEE Signal Processing Letters*, pages 1272–1276, 2022.

[72] Dawei Zhu, Michael A. Hedderich, Fangzhou Zhai, David Adelani, and Dietrich Klakow. Is BERT robust to label noise? a study on learning with noisy labels in text classification. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 62–67, 2022.

[73] Zhaowei Zhu, Zihao Dong, and Yang Liu. Detecting corrupted labels without training a model to predict. In *Proceedings of the 39th International Conference on Machine Learning*, pages 27412–27427, 2022.

A Appendix

A.1 Related Work

In-context learning In-context learning (ICL) has become a new paradigm for natural language processing (NLP), where LLMs make predictions only based on contexts augmented with a few demonstrations [7, 34–36]. The popularity of ICL also raises growing concerns regarding its instability: given different selected demonstrations, ICL’s performance can vary from near state-of-the-art to random [9, 12, 15, 26, 32, 37, 55, 54]. Existing studies show that ICL’s performance is highly sensitive to order [30, 60], template [37] and labels [50] of selected demonstrations. For example, on the one hand, some previous studies show that flip classification of demonstration can significantly hurt ICL performance on classification tasks [51, 60]. On the other hand, many researches show that ICL is fairly robust to noisy demonstrations [32, 37, 55, 54]. However, the existing studies only focus on classification tasks and the research of generation tasks is limited. We expand the previous finding from text classification tasks to generation tasks and find that demonstrations selected from noisy annotations significantly hurt the ICL performance of generation tasks.

In practice, researchers often use crowdsourcing [61, 73] or large language models (LLMs) [58] such as GPT-4 [38] to create input-output pairs for new tasks, which inevitably leads to some mistakes in the annotations. However, the existing demonstration selection methods for generation tasks such as TopK [28] or DPP [62] only consider the input of demonstrations and assume the demonstrations are selected from a completely clean dataset such as [16, 28, 62, 65]. In comparison, we aim to propose a training-free demonstration selection method for generation tasks that can consistently and significantly improve the robustness of the existing methods under noisy annotations.

Learning with noisy labels Label noise is common in many real-world datasets, especially generation tasks [2, 64]. The existing approaches to learning with noisy labels can be classified into two types: (1) training noise-robust models with noisy training datasets: designing noise-robust loss function [1, 63, 52, 72] or designing noise-robust model architectures [2, 18, 64] to mitigate label noise. However, this method is not suitable for ICL, which usually hypothesizes that users are unable to apply fine-tuning techniques [68]. (2) detecting noisy labels and reducing their impacts: comparing model predictions with noisy labels [42, 71] or checking the noisy label consensuses of nearby features [73]. Different from the above literature that focuses on classification tasks, we mainly consider a training-free solution to improve noise-robust ICL for generation tasks.

A.2 Experimental Setting

Datasets We conduct experiments on 6 generation tasks, and examples of each dataset are shown in Tables 12 and 13. For open-domain question-answering tasks, we choose the Natural Questions (NQ) dataset [22] and WebQuestions (WebQ) [5]. For reading comprehension tasks, we choose two reading comprehension datasets: Stanford Question Answering (SQuAD) Dataset [46] and Science Questions (SCIQ) dataset [56]. For code generation tasks, we choose Generating Tabular Answers for Multi-Table Question Answering (GeoQuery) Dataset [39] and Natural Language Interface to the Linux Operating System (NL2Bash) dataset [27]. Following previous studies [16, 25, 62], we report Exact Match (EM) for NQ, WebQ, SQuAD and SCIQ, BLEU for NL2Bash and GeoQuery. We collect these dataset from Huggingface. The train sets of these datasets are regarded as examples datasets and the test sets are used to evaluate the performance of ICL. We randomly subsample 20,000 examples from the train set to generate noisy annotations and select demonstrations. We provide a few examples of noisy annotations of each dataset in Tables 14, 15 and 16.

Baselines Our model LPR is essentially a data-centric retriever for in-context demonstration selection. We consider both learning-free and other learning-based retrievers as baselines:

1. **Random** randomly selects demonstrations from a example set without repetition [36].
2. **TopK** retrieve demonstration that are semantically-similar to a test query sample [28].
3. **DPP** uses the original BERT embedding as above without fine-tuning, and adopts MAP inference for subset retrieval [62].

Experiment details We run our experiments on NVIDIA L40 GPU. We adopt a large portion of the code from the OpenICL repository [59, 60]. The whole experiment around one week on 8 GPUs and each experiment around one hour on a single GPU.

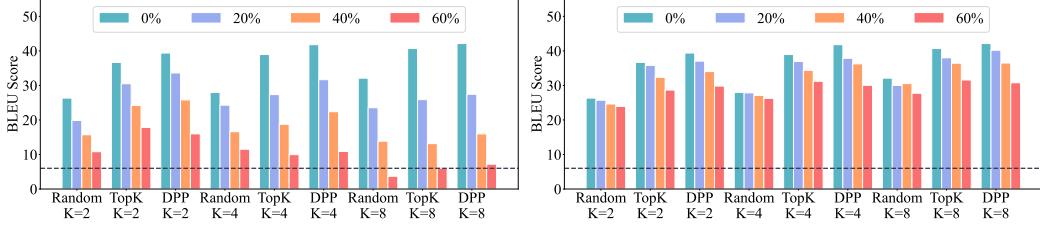


Figure 5: Average results of ICL with noisy annotations in various generation tasks across different demonstration settings. Both the two types of noises significantly deteriorate the performance of in-context learning on code generation tasks. The black line denotes zero-shot performance.

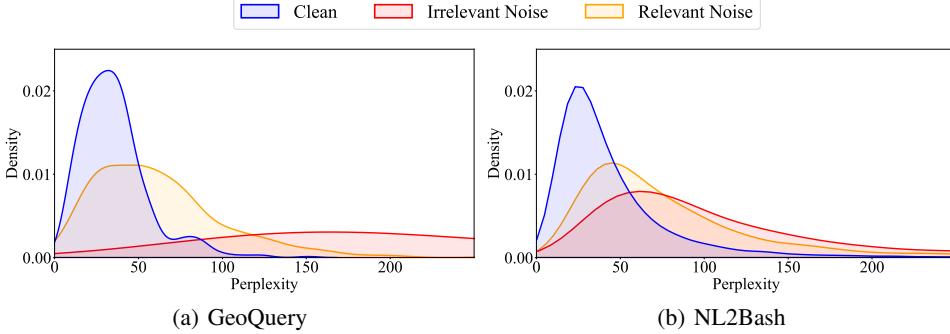


Figure 6: The distribution of perplexity of Llama2-7B [49] on clean and noisy annotations. Examples with noisy annotations indeed obtain higher perplexity than those with clean annotations.

Transfer to classification tasks Inspired by the idea implemented in above assumption, we assume that examples that are semantically similar share the similar task, indicating they should belong to same classification. We don't need to calculate the perplexity of input-output pair and only identify whether the classification of candidate demonstration is same with its local neighbors or not. Similar to generation tasks, we replace the noisy candidates with their nearest neighbors that are more likely to be clean. We investigate whether our local-based method can transfer across to classification tasks.

A.3 More empirical results

Empirical study of noisy ICL in text generation In this section, we provide the detailed results of GeoQuery and NL2Bash. Following existing studies [25, 62], we adopt BLEU score [41] to evaluate ICL performance on code generation tasks. Figure 5 shows that both the two types of noises significantly deteriorate the performance of in-context learning on code generation tasks. This phenomenon motivates us to further investigate the noise-robustness of in-context learning.

Perplexity deviation of noisy annotations In Figure 6, we present the perplexity distribution of Llama2-7B [49] on clean and noisy annotations of GeoQuery and NL2Bash datasets. As a complement, we observe that examples selected from noisy annotations set indeed obtain higher perplexity than those collected from clean annotations, which confirms the deviation can also transfer to code generation tasks.

Analysis of small noise ratios. In this section, we conduct experiments on datasets with smaller noise ratios (e.g. 5%, 10%, 15%). Figure 7 (a) and (b) present the average EM score on four generation tasks, including NQ, WebQ, SQuAD, and SQuAD. Figure 7 (a) and (b) show that our method can benefit the ICL performance from a small noise rate (e.g. 5%)

Open Benchmark Evaluation. Long-form and open-domain QA tasks such as MT-bench [70] and Arena-Hard [24] serve as valuable additions to the current standardized LLM benchmarks. In this section, we conduct experiments on these complex and open tasks to confirm the effectiveness of our method. The results on MT-bench [70] and Arena-Hard [24] are shown in the Figure 7 (c) and (d), which presents the average answer grading (0-10) [70] of baselines and our method. Figure 7 shows

Table 7: Average test performance of Zero-Shot, In-context learning, Chain-of-Thought (COT) and our proposed method across various noise types. The results are shown as Naive/+Ours. The bold indicates the improved results by integrating LPR.

Method	Clean		Irrelevant Noise			Relevant Noise		
	0%	20%	40%	60%	20%	40%	60%	
Zero-Shot	7.46							
Zero-Shot-COT	10.06							
Random-ICL/+Ours	27.94/ 28.60	24.28/ 28.11	16.61/ 25.78	11.53/ 22.58	26.84/ 28.27	27.11/ 28.95	26.26/ 26.76	
TopK-ICL/+Ours	39.94/ 38.62	27.34/ 36.38	18.75/ 32.15	9.96/ 23.93	38.94/ 36.92	34.35/ 36.39	31.19/ 33.62	
Manual-COT/+Ours	31.91/31.80	26.57/ 30.62	17.95/ 26.64	15.30/ 23.61	30.57/ 32.06	29.01/ 31.02	27.13/ 30.54	
Auto-COT/+Ours	45.69/45.44	30.51/ 40.10	20.51/ 34.94	10.86/ 27.32	41.38/ 42.78	35.91/ 40.73	27.90/ 37.10	

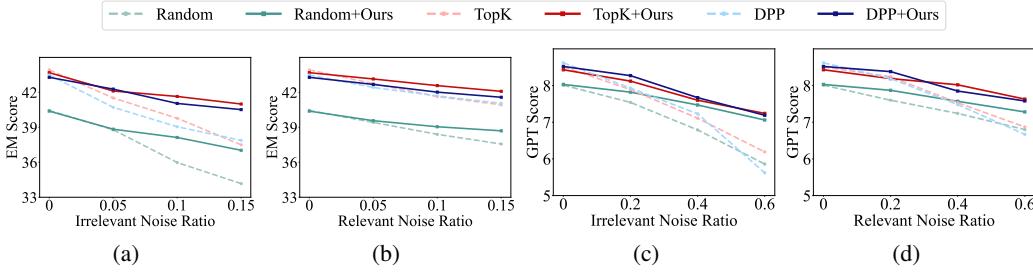


Figure 7: (a) and (b) demonstrate average EM scores of the baselines and our method for four generation tasks on four datasets with smaller noise ratios. (c) and (d) report average GPT-4 score [70] of the baselines and our method for two long-form and open-domain QA datasets.

that our approach significantly improves the efficacy of existing selection methods on long-form question-answering tasks.

Evaluation on the not-demonstration-selection-based baselines. Here, we add Zero-Shot baseline, as well as some CoT-related baselines, including Zero-Shot-CoT [20] and Manual-CoT [53], Auto-CoT [69]. Specifically, Manual-CoT [53] and Auto-CoT [69] require to select demonstrations from an annotated examples set. The table below presents the BLEU score of the baselines and our method on the two code generation tasks: Geoquery [39] and NL2Bash [27]. We use Llama2-7B [49] as the LLM throughout our experiments. The results in Table 7 show that our method can outperform Zero-Shot and Zero-Shot-CoT [20], and improve the noise robustness of Manual-CoT [53] and Auto-CoT [69].

Transfer to similarity score In LPR, we select reference demonstrations for candidate examples using cosine similarity. While cosine similarity captures some aspects of semantic similarity, it is limited to a single embedding [13]. Another measure of similarity necessitates an accurate characterization of the word levels. One way might be to use larger syntactic substructures of the input as terms with BM25, which is a sparse information retrieval algorithm belonging to a class of TF-IDF measures that view the test input and the candidates as bags of terms and measures relevance as a weighted recall of these terms:

$$\text{BM25}(\mathbf{x}_i, \mathbf{x}^*) = \sum_i^n W_i R(q_i, \mathbf{x}^*)$$

where q_i is each token of \mathbf{x}_i , $R(q_i, \mathbf{x}^*)$ and W_i are the term frequency and inverse document frequency statistics that measure the coverage of a particular term and the relative importance of terms. In this section, we replace the cosine similarity score with the BM25 similarity score to verify the effectiveness of our proposed method.

Our results in Table 8 show that the improvement still holds when BM25 is used as the cluster, confirming the superiority and robustness of our method compared with the naive demonstration selection methods. The above result also demonstrates that examples that are semantically similar in both token space and word space share the same level of inherent perplexity.

Reordering Some studies demonstrate that in-context learning is highly sensitive for demonstrations' ordering when using random demonstrations [25, 30, 60]. Specifically, the same randomly sampled demonstrations with different orders can lead to the performance between random guesses and near state-of-the-art. In LPR, we reorder exemplars based on their similarities to the test input in ascending

Table 8: Average results with BM25 as similarity score. The result of The bold indicates the improvement by integrating LPR.

Dataset	Method	Clean		Irrelevant Noise			Relevant Noise		
		0%	20%	40%	60%	20%	40%	60%	
NQ	Random	14.51 \pm 0.51	10.97 \pm 0.29	7.37 \pm 0.45	4.23 \pm 0.46	12.00 \pm 0.65	9.67 \pm 0.45	6.40 \pm 1.02	
	+Ours	15.15\pm0.20	13.45\pm0.85	10.98\pm0.47	7.51\pm0.43	14.14\pm0.48	12.08\pm0.53	10.12\pm0.53	
	TopK	20.25 \pm 0.10	13.95 \pm 1.14	9.97 \pm 1.13	5.90 \pm 1.08	16.21 \pm 0.22	12.22 \pm 0.22	8.50 \pm 0.28	
	+Ours	19.35 \pm 0.21	17.51\pm0.08	14.44\pm0.41	10.25\pm0.45	17.39\pm0.45	14.38\pm0.90	11.71\pm0.71	
WebQ	DPP	20.35 \pm 0.76	14.69 \pm 0.94	9.87 \pm 0.49	5.97 \pm 0.48	15.47 \pm 1.00	11.28 \pm 0.42	7.89 \pm 0.25	
	+Ours	19.45 \pm 0.97	17.29\pm1.19	14.08\pm0.98	10.45\pm0.68	17.07\pm0.84	15.02\pm0.59	12.14\pm0.75	
	Random	20.37 \pm 0.64	15.18 \pm 1.06	10.39 \pm 0.83	4.83 \pm 0.17	18.29 \pm 0.43	15.92 \pm 0.68	13.50 \pm 0.17	
	+Ours	21.18\pm0.14	19.83\pm0.71	16.40\pm0.28	10.89\pm0.24	20.38\pm0.71	18.54\pm0.48	15.92\pm0.48	
SQuAD	TopK	30.16 \pm 0.58	22.52 \pm 0.64	14.52 \pm 0.78	8.00 \pm 1.12	27.19 \pm 0.27	22.82 \pm 0.75	18.88 \pm 1.09	
	+Ours	28.82 \pm 0.72	26.51\pm0.39	22.03\pm1.26	14.74\pm0.25	27.56\pm0.20	25.08\pm0.36	21.58\pm0.21	
	DPP	29.40 \pm 0.39	22.11 \pm 0.81	13.72 \pm 0.27	7.33 \pm 0.68	26.18 \pm 1.04	21.53 \pm 0.61	16.80 \pm 0.17	
	+Ours	29.15 \pm 0.21	26.30\pm0.93	20.93\pm1.42	13.72\pm0.57	27.83\pm0.33	25.08\pm0.93	20.57\pm1.27	
SCIQ	Random	56.50 \pm 0.57	50.00 \pm 0.62	39.10 \pm 0.88	26.20 \pm 0.79	53.90 \pm 0.65	49.17 \pm 0.62	42.03 \pm 0.79	
	+Ours	56.47 \pm 0.25	54.73\pm1.10	51.53\pm1.59	43.03\pm1.51	54.77\pm0.76	52.83\pm0.97	49.70\pm0.08	
	TopK	56.97 \pm 0.69	51.83 \pm 1.03	42.83 \pm 1.68	29.10 \pm 2.92	54.77 \pm 0.69	49.37 \pm 1.37	41.37 \pm 2.09	
	+Ours	56.83 \pm 0.19	55.60\pm1.45	50.33\pm0.62	40.83\pm2.82	55.70\pm0.99	53.07\pm0.65	48.17\pm1.92	
GeoQuery	DPP	57.29 \pm 0.87	50.57 \pm 0.33	41.63 \pm 1.00	25.67 \pm 2.52	56.10 \pm 0.59	49.57 \pm 1.24	43.37 \pm 0.78	
	+Ours	57.20 \pm 1.00	56.50\pm0.83	52.70\pm0.86	44.73\pm1.19	56.43\pm1.13	53.47\pm0.81	50.57\pm1.19	
	Random	68.15 \pm 0.28	59.19 \pm 1.57	44.19 \pm 2.89	28.21 \pm 2.96	64.59 \pm 1.42	58.39 \pm 0.16	49.54 \pm 0.80	
	+Ours	69.25\pm0.86	64.14\pm1.47	54.37\pm1.88	37.64\pm0.58	66.49\pm1.14	62.24\pm0.86	54.19\pm0.82	
NL2Bash	TopK	68.62 \pm 1.13	59.59 \pm 1.28	45.77 \pm 2.68	29.31 \pm 1.73	64.66 \pm 1.34	58.54 \pm 0.12	49.47 \pm 0.65	
	+Ours	70.11\pm0.36	63.79\pm2.87	57.58\pm1.52	38.90\pm2.93	66.55\pm2.74	60.23\pm5.44	51.95\pm4.59	
	DPP	67.29 \pm 0.35	57.69 \pm 1.83	45.34 \pm 1.56	28.50 \pm 1.78	64.88 \pm 0.43	58.91 \pm 0.64	50.00 \pm 0.85	
	+Ours	69.78\pm1.00	64.94\pm1.42	55.34\pm2.12	41.21\pm1.52	67.64\pm0.86	63.85\pm2.05	56.43\pm2.50	
GeoQuery	Random	27.97 \pm 0.99	23.18 \pm 0.62	17.44 \pm 1.56	14.10 \pm 0.74	26.48 \pm 0.17	26.13 \pm 0.05	26.25 \pm 0.40	
	+Ours	29.99\pm0.50	29.35\pm0.26	25.69\pm0.91	25.11\pm0.64	29.77\pm0.35	28.09\pm0.50	26.80\pm0.55	
	TopK	44.17 \pm 0.09	27.28 \pm 2.65	17.49 \pm 2.05	9.96 \pm 3.08	41.31 \pm 0.46	38.48 \pm 0.63	34.90 \pm 0.69	
	+Ours	43.06 \pm 0.60	41.61\pm1.00	41.19\pm1.42	32.76\pm0.45	40.99\pm38.71	38.87\pm0.49	36.26\pm0.03	
NL2Bash	DPP	45.81 \pm 0.71	31.79 \pm 5.93	21.54 \pm 3.36	10.61 \pm 0.15	42.97 \pm 1.96	39.91 \pm 0.42	33.34 \pm 0.53	
	+Ours	43.92 \pm 3.44	41.32\pm3.55	38.37\pm4.19	26.78\pm3.32	41.70\pm1.22	39.79\pm2.13	35.34\pm2.10	
	Random	27.91 \pm 0.37	25.37 \pm 0.21	15.77 \pm 0.91	8.95 \pm 0.65	27.20 \pm 1.06	28.09 \pm 0.51	26.27 \pm 0.56	
	+Ours	29.15\pm0.21	26.30\pm0.93	20.93\pm1.42	13.72\pm0.57	28.83\pm0.33	28.08\pm0.93	27.57\pm1.27	
NL2Bash	TopK	35.71 \pm 0.42	27.40 \pm 0.26	20.00 \pm 0.62	9.95 \pm 0.68	32.57 \pm 0.13	30.21 \pm 0.08	27.48 \pm 0.35	
	+Ours	32.42 \pm 0.26	29.85\pm2.99	30.10\pm2.11	23.67\pm1.02	31.18\pm38.71	31.03\pm3.80	28.84\pm2.48	
	DPP	37.77 \pm 0.02	31.52 \pm 0.12	23.23 \pm 0.34	11.16 \pm 2.14	32.74 \pm 0.29	32.56 \pm 0.61	26.72 \pm 1.58	
	+Ours	36.69 \pm 3.30	32.63\pm3.32	29.10\pm4.10	23.56\pm2.65	33.18\pm2.51	32.19 \pm 3.46	28.65\pm1.80	

order after the process of our method, in accordance with common practices [23, 28, 45, 47, 62]. Here we compare our method with reordering and without reordering to explore the effect of reordering on example-specific demonstrations retrieved by our method.

Across our experiments, Table 9 shows our method without the reordering process still improves the existing demonstration selection methods across various types of noise. The above results indicate that local perplexity ranking rather than the reordering process is crucial for the success of noise-robust ICL. Additionally, we believe high-quality demonstrations are less sensitive to the ordering and stabilize in-context learning, which is consistent with the previous work [16].

Various demonstration sizes To verify the effectiveness of our proposed method, we also present the ICL performance with our method across various set sizes K . Concretely, Tables 10 and 11 report the number of demonstrations to be 2 and 8 and show our method effectively mitigates the issue of noisy annotation in various demonstration selection methods across various demonstration sizes.

Table 9: Average results without reordering process. The result of The bold indicates the improvement by integrating LPR.

Dataset	Method	Clean		Irrelevant Noise			Relevant Noise		
		0%	20%	40%	60%	20%	40%	60%	
NQ	Random	14.51 \pm 0.51	10.97 \pm 0.29	7.37 \pm 0.45	4.23 \pm 0.46	12.00 \pm 0.65	9.67 \pm 0.45	6.40 \pm 1.02	
	+Ours	15.35\pm0.83	14.58\pm0.33	12.38\pm0.09	9.24\pm1.24	14.28\pm0.46	12.95\pm0.91	9.93\pm0.94	
	TopK	20.25 \pm 0.10	13.95 \pm 1.14	9.97 \pm 1.13	5.90 \pm 1.08	16.21 \pm 0.22	12.22 \pm 0.22	8.50 \pm 0.28	
	+Ours	19.65 \pm 0.24	16.88\pm0.40	13.21\pm0.38	9.47\pm0.38	17.42\pm0.36	14.58\pm0.26	11.61\pm0.59	
	DPP	20.35 \pm 0.76	14.69 \pm 0.94	9.87 \pm 0.49	5.97 \pm 0.48	15.47 \pm 1.00	11.28 \pm 0.42	7.89 \pm 0.25	
	+Ours	18.57 \pm 0.24	17.45\pm0.37	14.48\pm0.85	11.44\pm0.29	17.75\pm0.29	15.45\pm0.70	12.18\pm0.87	
WebQ	Random	20.37 \pm 0.64	15.18 \pm 1.06	10.39 \pm 0.83	4.83 \pm 0.17	18.29 \pm 0.43	15.92 \pm 0.68	13.50 \pm 0.17	
	+Ours	22.08\pm0.31	20.38\pm0.74	16.91\pm0.61	12.16\pm0.21	21.64\pm0.71	19.01\pm0.78	17.06\pm1.35	
	TopK	30.16 \pm 0.58	22.52 \pm 0.64	14.52 \pm 0.78	8.00 \pm 1.12	27.19 \pm 0.27	22.82 \pm 0.75	18.88 \pm 1.09	
	+Ours	29.69 \pm 0.22	26.96\pm0.66	22.12\pm1.08	15.98\pm0.60	29.07\pm0.04	27.26\pm0.40	22.33\pm1.13	
	DPP	29.40 \pm 0.39	22.11 \pm 0.81	13.72 \pm 0.27	7.33 \pm 0.68	26.18 \pm 1.04	21.53 \pm 0.61	16.80 \pm 0.17	
	+Ours	29.15 \pm 0.21	26.30\pm0.93	20.93\pm1.42	13.72\pm0.57	27.83\pm0.33	25.08\pm0.93	20.57\pm1.27	
SQuAD	Random	56.50 \pm 0.57	50.00 \pm 0.62	39.10 \pm 0.88	26.20 \pm 0.79	53.90 \pm 0.65	49.17 \pm 0.62	42.03 \pm 0.79	
	+Ours	55.93 \pm 0.75	54.23\pm1.11	51.67\pm0.39	41.37\pm0.66	55.67\pm0.52	53.13\pm0.63	49.07\pm0.74	
	TopK	56.97 \pm 0.69	51.83 \pm 1.03	42.83 \pm 1.68	29.10 \pm 2.92	54.77 \pm 0.69	49.37 \pm 1.37	41.37 \pm 2.09	
	+Ours	57.83\pm0.97	54.87\pm0.83	50.97\pm0.70	39.00\pm3.12	56.40\pm0.37	52.77\pm0.83	47.63\pm0.94	
	DPP	57.29 \pm 0.87	50.57 \pm 0.33	41.63 \pm 1.00	25.67 \pm 2.52	56.10 \pm 0.59	49.57 \pm 1.24	43.37 \pm 0.78	
	+Ours	57.47\pm0.25	57.53\pm0.97	52.03\pm0.39	44.00\pm1.10	57.27\pm0.40	55.00\pm0.22	50.27\pm1.51	
SCIQ	Random	68.15 \pm 0.28	59.19 \pm 1.57	44.19 \pm 2.89	28.21 \pm 2.96	64.59 \pm 1.42	58.39 \pm 0.16	49.54 \pm 0.80	
	+Ours	68.56\pm1.17	64.88\pm1.22	54.94\pm1.00	40.63\pm2.62	66.67\pm1.34	62.41\pm0.24	54.03\pm1.69	
	TopK	68.62 \pm 1.13	59.59 \pm 1.28	45.77 \pm 2.68	29.31 \pm 1.73	64.66 \pm 1.34	58.54 \pm 0.12	49.47 \pm 0.65	
	+Ours	70.00\pm0.25	66.26\pm0.35	56.32\pm1.90	41.03\pm1.89	68.19\pm0.13	63.27\pm0.75	55.17\pm2.12	
	DPP	67.29 \pm 0.35	57.69 \pm 1.83	45.34 \pm 1.56	28.50 \pm 1.78	64.88 \pm 0.43	58.91 \pm 0.64	50.00 \pm 0.85	
	+Ours	70.00\pm0.61	66.26\pm1.27	56.03\pm2.04	43.44\pm2.54	68.70\pm0.21	63.22\pm1.84	54.77\pm2.47	
GeoQuery	Random	27.97 \pm 0.99	23.18 \pm 0.62	17.44 \pm 1.56	14.10 \pm 0.74	26.48 \pm 0.17	26.13 \pm 0.05	26.25 \pm 0.40	
	+Ours	28.58\pm0.59	28.60\pm0.03	28.89\pm1.77	22.61\pm0.53	27.80\pm0.27	28.45\pm0.32	26.86\pm0.76	
	TopK	44.17 \pm 0.09	27.28 \pm 2.65	17.49 \pm 2.05	9.96 \pm 3.08	41.31 \pm 0.46	38.48 \pm 0.63	34.90 \pm 0.69	
	+Ours	45.63\pm0.11	43.62\pm0.70	35.05\pm2.86	26.03\pm4.93	42.74\pm0.45	39.81\pm0.80	35.75\pm0.03	
	DPP	45.81 \pm 0.71	31.79 \pm 5.93	21.54 \pm 3.36	10.61 \pm 0.15	42.97 \pm 1.96	39.91 \pm 0.42	33.34 \pm 0.53	
	+Ours	44.73 \pm 0.56	45.10\pm0.50	40.26\pm1.06	32.54\pm1.25	41.64 \pm 0.64	40.78\pm0.89	35.09\pm0.79	
NL2Bash	Random	27.91 \pm 0.37	25.37 \pm 0.21	15.77 \pm 0.91	8.95 \pm 0.65	27.20 \pm 1.06	28.09 \pm 0.51	26.27 \pm 0.56	
	+Ours	25.54 \pm 2.19	25.02 \pm 1.25	23.05\pm2.22	21.28\pm2.12	27.63\pm0.58	24.21 \pm 0.66	24.09 \pm 0.38	
	TopK	35.71 \pm 0.42	27.40 \pm 0.26	20.00 \pm 0.62	9.95 \pm 0.68	32.57 \pm 0.13	30.21 \pm 0.08	27.48 \pm 0.35	
	+Ours	32.91 \pm 0.21	31.33\pm0.50	29.83\pm0.31	22.20\pm0.95	31.39 \pm 0.74	31.14\pm0.46	29.09\pm1.77	
	DPP	37.77 \pm 0.02	31.52 \pm 0.12	23.23 \pm 0.34	11.16 \pm 2.14	32.74 \pm 0.29	32.56 \pm 0.61	26.72 \pm 1.58	
	+Ours	38.37\pm0.32	31.81\pm1.08	24.27\pm2.14	13.09\pm0.05	34.43\pm0.91	32.32 \pm 1.94	28.76\pm0.88	

Table 10: Average in-context learning performance with 2 demonstrations on 6 datasets across various types of noisy annotation (over 3 runs). The bold indicates the improved results by integrating LPR.

Dataset	Method	Clean		Irrelevant Noise			Relevant Noise		
		0%	20%	40%	60%	20%	40%	60%	
NQ	Random	11.70±0.49	9.27±0.75	6.70±0.81	4.57±0.34	11.04±0.27	8.83±0.54	5.80±0.45	
	+Ours	12.01±0.71	11.38±0.92	10.71±0.35	8.27±0.75	11.58±0.78	10.38±0.87	8.90±0.51	
	TopK	14.61±0.49	12.18±0.05	9.53±0.53	6.42±0.38	11.96±0.62	8.89±0.21	7.14±0.21	
	+Ours	14.28±0.31	12.84±0.31	11.44±0.54	9.31±1.13	12.98±0.31	11.84±0.46	9.77±0.66	
	DPP	15.48±0.26	12.03±0.12	9.03±0.31	6.34±0.49	12.21±0.21	9.17±0.54	6.87±0.36	
WebQ	+Ours	14.68±0.61	13.91±0.65	12.11±1.28	10.04±1.26	14.44±0.50	12.41±0.33	10.33±0.45	
	Random	16.06±0.74	12.95±0.20	9.89±1.38	6.65±0.74	14.32±0.34	13.30±1.45	11.74±0.76	
	+Ours	16.28±0.65	16.20±0.30	13.36±0.33	14.41±1.22	16.58±0.34	15.21±0.51	13.89±0.43	
	TopK	20.04±0.13	16.33±0.17	11.88±0.07	8.43±0.45	17.59±1.41	14.82±0.16	11.52±0.65	
	+Ours	20.50±0.52	19.44±0.59	17.34±0.95	14.41±1.22	20.51±0.45	18.34±0.84	15.87±0.67	
SQuAD	DPP	22.68±0.61	18.10±0.86	13.12±0.44	8.66±0.51	19.60±0.17	17.32±0.57	13.97±0.63	
	+Ours	22.20±0.74	20.43±0.40	17.96±0.78	13.91±0.76	21.67±0.35	20.79±0.83	16.89±0.49	
	Random	45.07±0.37	41.13±1.03	34.77±0.39	27.60±1.40	43.27±0.59	38.83±1.21	35.63±0.62	
	+Ours	44.57±0.48	42.43±1.03	42.60±1.02	37.40±0.64	46.10±1.98	42.70±0.65	40.03±0.40	
	TopK	45.13±0.76	40.57±0.94	36.00±0.51	29.17±1.33	42.73±1.03	41.73±0.59	35.57±1.26	
SCIQ	+Ours	45.20±0.83	44.37±0.65	41.17±0.58	35.53±1.71	44.07±0.45	42.47±1.14	39.67±1.36	
	DPP	46.23±1.58	41.67±1.72	35.43±1.76	27.90±0.75	43.33±0.57	40.20±0.50	37.23±0.60	
	+Ours	46.67±0.59	44.53±0.68	42.17±1.03	37.10±0.41	44.77±0.29	43.23±0.92	41.20±0.22	
	Random	66.48±0.34	62.01±0.96	51.03±2.33	40.36±1.94	64.65±0.92	59.48±0.75	56.89±0.25	
	+Ours	65.17±0.56	62.64±0.69	57.59±1.01	48.79±1.60	64.08±0.80	61.61±0.33	56.26±1.92	
GeoQuery	TopK	65.17±0.49	58.50±1.10	50.29±1.06	40.54±1.68	62.76±1.36	58.27±0.73	54.54±0.86	
	+Ours	67.04±0.50	64.60±0.63	57.81±1.69	49.88±1.79	65.63±0.80	61.26±0.80	55.46±3.17	
	DPP	67.33±0.74	61.37±0.98	51.49±1.06	41.26±1.75	62.53±1.06	58.16±0.70	53.79±0.51	
	+Ours	67.24±0.92	64.48±1.60	59.37±0.84	50.57±2.34	66.95±2.41	62.12±1.18	56.84±2.00	
	Random	24.11±1.06	18.22±0.87	12.35±0.35	7.36±0.51	24.55±0.42	21.55±0.46	20.40±0.41	
NL2Bash	+Ours	23.67±0.98	22.23±0.27	19.99±0.17	16.29±0.96	22.84±0.68	22.78±0.22	22.34±1.24	
	TopK	41.48±0.41	32.11±0.69	26.11±1.92	18.57±3.32	40.08±1.57	36.97±1.29	33.27±1.88	
	+Ours	41.1±0.43	40.82±0.70	41.51±0.76	37.08±0.55	38.72±0.92	36.61±1.10	35.16±0.56	
	TopK	43.63±0.79	35.91±3.46	25.77±1.34	14.66±0.25	39.11±1.53	35.88±0.98	32.01±2.06	
	DPP	41.97±0.05	40.67±1.25	41.01±0.26	36.42±0.12	38.13±0.33	35.11±0.50	33.66±0.05	
NL2Bash	Random	28.56±0.89	21.45±2.64	19.07±0.75	14.25±2.48	26.87±0.82	25.69±0.26	24.47±0.73	
	+Ours	26.35±0.20	24.37±0.32	25.44±1.13	20.73±0.29	26.22±0.25	26.54±0.28	26.10±1.75	
	TopK	31.83±0.10	28.85±0.68	22.3±2.01	17.08±3.47	31.51±0.82	27.73±0.36	24.04±0.91	
	+Ours	35.10±0.06	34.11±0.51	30.73±0.47	26.04±1.00	34.02±0.41	30.22±0.54	27.12±0.81	
	DPP	35.13±1.07	31.31±0.85	25.84±0.92	17.28±0.72	34.95±0.43	32.14±0.59	27.61±0.85	
NL2Bash	+Our	33.79±0.33	30.84±0.63	30.34±1.02	26.41±1.58	32.88±1.34	31.40±0.59	28.82±0.30	

Table 11: Average in-context learning performance with 8 demonstrations on 6 datasets across various types of noisy annotation (over 3 runs). The bold indicates the improved results by integrating LPR.

Dataset	Method	Clean		Irrelevant Noise			Relevant Noise		
		0%	20%	40%	60%	20%	40%	60%	
NQ	Random	16.25 \pm 0.95	11.62 \pm 0.24	6.15 \pm 0.51	3.17 \pm 0.17	12.72 \pm 0.44	9.37 \pm 0.24	6.17 \pm 0.52	
	+Ours	16.55\pm0.47	14.08\pm0.61	11.88\pm0.69	8.64\pm0.70	14.58\pm0.42	13.31\pm0.64	9.74\pm0.74	
	TopK	21.09 \pm 0.42	14.91 \pm 1.26	8.57 \pm 0.40	5.47 \pm 0.21	17.98 \pm 0.34	12.71 \pm 1.02	8.87 \pm 0.85	
	+Ours	20.65 \pm 0.09	16.65\pm0.25	12.41\pm0.65	8.14\pm1.33	17.85\pm0.47	15.45\pm0.82	12.01\pm0.65	
WebQ	DPP	19.65 \pm 0.31	13.34 \pm 0.84	8.64 \pm 0.72	5.30 \pm 0.00	16.31 \pm 0.51	12.48 \pm 1.07	8.43 \pm 0.48	
	+Ours	18.68 \pm 0.29	16.62\pm0.45	13.71\pm0.57	9.54\pm1.14	17.49\pm0.41	15.82\pm0.17	12.22\pm1.21	
	Random	22.70 \pm 0.55	15.62 \pm 0.34	8.33 \pm 0.31	3.22 \pm 0.37	19.91 \pm 0.32	17.00 \pm 0.73	13.52 \pm 0.75	
	+Ours	22.93\pm0.47	20.62\pm0.60	16.18\pm0.89	9.71\pm0.23	22.36\pm0.27	20.52\pm0.28	17.85\pm1.05	
SQuAD	TopK	33.52 \pm 0.67	22.49 \pm 0.95	12.51 \pm 0.92	6.50 \pm 0.57	28.69 \pm 0.69	24.17 \pm 0.31	19.50 \pm 0.56	
	+Ours	31.64 \pm 0.10	26.91\pm0.37	19.52\pm1.29	12.71\pm1.50	29.29\pm0.48	26.32\pm1.39	21.86\pm0.60	
	DPP	31.49 \pm 0.27	22.66 \pm 0.95	12.51 \pm 0.48	5.27 \pm 1.43	27.64 \pm 0.64	22.90 \pm 0.52	17.82 \pm 0.00	
	+Ours	30.39 \pm 0.10	26.00\pm1.01	19.08\pm0.46	11.47\pm0.98	28.74\pm0.43	26.43\pm1.53	21.97\pm1.33	
SCIQ	Random	58.70 \pm 0.59	46.63 \pm 1.20	27.80 \pm 1.42	11.03 \pm 0.62	54.37 \pm 0.66	46.57 \pm 1.02	35.90 \pm 1.71	
	+Ours	57.73 \pm 0.79	56.87\pm0.47	48.50\pm0.86	33.00\pm1.31	57.70\pm0.65	53.93\pm0.33	47.57\pm0.90	
	TopK	58.97 \pm 0.42	49.80 \pm 1.44	34.87 \pm 1.68	15.53 \pm 2.23	56.63 \pm 0.80	49.60 \pm 0.78	36.37 \pm 1.16	
	+Ours	58.33 \pm 0.29	55.60\pm0.16	48.30\pm0.08	30.27\pm2.08	58.60\pm1.56	55.13\pm1.33	44.40\pm0.86	
GeoQuery	DPP	56.93 \pm 0.34	49.63 \pm 1.35	33.17 \pm 0.45	16.50 \pm 1.31	55.63 \pm 1.11	49.07 \pm 0.74	36.47 \pm 2.09	
	+Ours	57.67\pm0.82	56.30\pm0.14	50.53\pm1.14	34.03\pm2.43	57.83\pm0.25	53.87 \pm 1.25	44.87\pm2.05	
	Random	68.70 \pm 0.16	52.47 \pm 0.59	28.46 \pm 1.13	12.30 \pm 3.25	63.79 \pm 0.64	49.82 \pm 1.66	37.18 \pm 1.23	
	+Ours	69.54\pm0.33	62.58\pm1.22	43.79\pm2.46	26.49\pm1.83	66.55\pm1.70	58.62\pm0.65	42.13\pm3.07	
NL2Bash	TopK	68.91 \pm 0.22	53.73 \pm 0.22	31.25 \pm 1.35	12.87 \pm 2.46	62.13 \pm 0.77	50.05 \pm 1.05	35.32 \pm 1.45	
	+Ours	70.00\pm0.49	65.23\pm1.78	47.53\pm3.64	28.10\pm4.98	67.87\pm2.00	58.39\pm3.14	43.33\pm4.97	
	DPP	68.33 \pm 0.72	55.80 \pm 0.90	34.54 \pm 1.92	15.29 \pm 3.34	62.64 \pm 1.87	52.41 \pm 3.41	39.65 \pm 2.12	
	+Ours	68.39\pm0.43	65.14\pm1.50	48.67\pm1.62	29.13\pm2.44	68.04\pm1.59	58.22\pm2.01	46.09\pm1.70	
GeoQuery	Random	34.03 \pm 0.25	25.49 \pm 1.64	13.95 \pm 2.4	3.02 \pm 0.19	32.83 \pm 0.25	30.98 \pm 0.16	28.72 \pm 0.13	
	+Ours	32.48 \pm 0.46	33.36\pm1.19	31.07\pm1.66	23.76\pm0.92	32.42 \pm 0.14	31.03\pm2.29	29.88\pm0.84	
	TopK	45.18 \pm 0.48	22.07 \pm 4.26	10.12 \pm 0.19	3.61 \pm 1.39	42.63 \pm 0.56	40.43 \pm 0.28	34.06 \pm 0.76	
	+Ours	45.08 \pm 0.56	41.88\pm0.10	27.72\pm2.40	13.09\pm1.91	42.73\pm0.53	41.37\pm0.22	35.53\pm1.81	
NL2Bash	DPP	46.71 \pm 0.29	25.01 \pm 2.15	15.29 \pm 0.05	8.10 \pm 0.95	44.75 \pm 0.48	39.74 \pm 0.23	33.43 \pm 0.65	
	+Ours	45.89 \pm 0.27	46.09\pm0.57	35.01\pm0.95	23.43\pm2.93	44.91\pm1.02	40.70\pm0.59	34.66\pm0.60	
	Random	30.17 \pm 0.54	21.57 \pm 1.84	13.74 \pm 2.31	4.37 \pm 0.62	27.18 \pm 0.71	28.10 \pm 0.39	26.75 \pm 1.08	
	+Ours	29.30 \pm 0.97	28.38\pm0.19	27.51\pm2.27	18.32\pm1.50	26.58 \pm 0.37	28.45\pm0.09	27.26\pm0.80	
SQuAD	TopK	36.17 \pm 1.06	29.69 \pm 0.16	16.17 \pm 1.05	8.50 \pm 1.02	33.35 \pm 2.46	32.42 \pm 0.77	29.08 \pm 1.58	
	+Ours	35.16 \pm 0.03	33.32\pm0.20	27.73\pm0.97	17.82\pm4.71	33.14 \pm 0.72	32.75\pm0.20	29.69\pm0.39	
	DPP	37.55 \pm 0.56	29.87 \pm 3.04	16.65 \pm 0.36	6.30 \pm 1.20	34.61 \pm 0.46	32.24 \pm 0.69	28.11 \pm 0.64	
	+Our	36.93 \pm 0.89	34.65 \pm 0.23	30.45\pm1.17	25.36\pm1.00	34.57 \pm 1.14	33.31\pm1.26	31.90\pm0.74	

Table 12: All the datasets used in the experiments.

Task	Dataset	Train Set	Test Set
Open-Domain QA	NQ [22]	20,000	1,000
	WebQ [5]	1,261	1,213
Reading Comprehension	SQuAD [46]	20,000	1,000
	SCIQ [56]	6,059	581
Code Generation	GeoQuery [39]	530	253
	NL2Bash [27]	5,000	606

Table 13: Templates of tasks. Placeholders(e.g. <Question> and <Answer>) will be replaced by real questions or answers.

Dataset	Prompt	Example
NQ	Question: <Question> Answer: <Answer>	Question: The bundles of neurons in the cns are called? Answer: Nucleus
WebQ	Question: <Question> Answer: <Answer>	Question: Where are the libyan refugees going? Answer: Tunisia
SQuAD	Support: <Support> Question: <Question> Answer: <Answer>	Support: Among the philosophies that have influenced modern architects and their approach to building design are rationalism, empiricism, structuralism, poststructuralism. Question: Which philosophy followed structuralism? Answer: poststructuralism
SCIQ	Support: <Support> Question: <Question> Answer: <Answer>	Support: Gravity keeps the Moon orbiting Earth. Gravity keeps the planets orbiting the Sun. Question: What keeps the moon orbiting earth? Answer: gravity
GeoQuery	Question: <Question> Answer: <Answer>	Question: which state is Kalamazoo in Answer: SELECT city.statename FROM city WHERE city.cityname=kalamazoo
NL2Bash	Question: <Question> Answer: <Answer>	Question: Add "execute" to the permissions of all directories in the home directory tree Answer: find -type d -exec chmod +x {};

Table 14: An illustration of the effect of the label of demonstration, with three different types of input-label mapping of demonstration. The middle lines are demonstrations, and the last line is the model prediction. The model tends to learn the label of the demonstration.

NQ Test	Question: When did computer become widespread in homes and schools? Answer:	
Setting	In-Context Demonstration	Prediction
Clean	Question: When did the internet first become available to the public? Answer: 1980s	1980s
Irrelevant	Question: When did the internet first become available to the public? Answer: Crude Oil	Crude Oil
Relevant	Question: When did the internet first become available to the public? Answer: 2010s	2010s
WebQ Test	Question: When did computer become widespread in homes and schools? Answer:	
Setting	In-Context Demonstration	Prediction
Clean	Question: When did the internet first become available to the public? Answer: 1980s	1980s
Irrelevant	Question: When did the internet first become available to the public? Answer: Crude Oil	Crude Oil
Relevant	Question: When did the internet first become available to the public? Answer: 2010s	2010s

Table 15: An illustration of the effect of the label of demonstration, with three different types of input-label mapping of demonstration. The middle lines are demonstrations, and the last line is the model prediction. The model tends to learn the label of the demonstration.

		SQuAD Input	Setting	In-Context Demonstration	Prediction
Setting	Input	SQuAD Input	Support: The Super Bowl 50 Host Committee has vowed to be "the most giving Super Bowl ever", and will dedicate 25 percent of all money it raises for philanthropic causes in the Bay Area. The committee created the 50 fund as its philanthropic initiative and focuses on providing grants to aid with youth development, community investment and sustainable environments.		
			Question: What is the name of the fund that focuses on youth, community and sustainable environments?		
			Output:		
Setting	Input	SCIQ Input	Support: UNFPA works in partnership with governments , along with other United Nations agencies, communities, NGOs, foundations and the private sector, to raise awareness and mobilize the support needed to achieve its mission to promote the rights and health of women and young people.		
			Question: With what sort of agencies does UNFPA work?		25 percent
			Output: governments		
Setting	Input	SCIQ Input	Support: Cells are organized into tissues, tissues are organized into organs.		
			Question: What is considered the smallest unit of the organ?		Earth
			Output: Earth		
Setting	Input	SCIQ Input	Support: Cells are organized into tissues, tissues are organized into organs.		
			Question: What is considered the smallest unit of the organ?		50 fund
			Output: tissues		
Setting	Input	SCIQ Input	Support: All forms of life are built of at least one cell. A cell is the basic unit of the structure and function of living things.		
			Question: What are the smallest structural and functional units of all living organisms?		
			Output:		
Setting	Input	SCIQ Input	Support: Cells are organized into tissues, tissues are organized into organs.		
			Question: What is considered the smallest unit of the organ?		Cells
			Output: Cells		
Setting	Input	SCIQ Input	Support: Cells are organized into tissues, tissues are organized into organs.		
			Question: What is considered the smallest unit of the organ?		Earth
			Output: Earth		
Setting	Input	SCIQ Input	Support: Cells are organized into tissues, tissues are organized into organs.		
			Question: What is considered the smallest unit of the organ?		tissues
			Output: tissues		

Table 16: An illustration of the effect of the label of demonstration, with three different types of input-label mapping of demonstration. The middle lines are demonstrations, and the last line is the model prediction. The model tends to learn the label of the demonstration.

GeoQuery Test	Setting	Question: How high is the highest point of Alabama? Answer:	In-Context Demonstration
Clean		Question: How high is the highest point in Montana? Answer: <code>SELECT highlow.highest.elevation FROM highlow WHERE highlow.statename='Montana'</code>	
		Prediction: <code>SELECT highlow.highest.elevation FROM highlow WHERE highlow.statename='Alabama'</code>	
	Irrelevant	Question: How high is the highest point in Montana? Answer: <code>more than 900 million</code>	
Relevant		Prediction: <code>What are the highest point in Alabama</code>	
		Question: How high is the highest point in Montana? Answer: <code>SELECT city.cityname FROM city WHERE city.statename='Montana'</code>	
		Prediction: <code>SELECT city.cityname FROM city WHERE city.statename='Alabama'</code>	
NL2Bash Test		Question: List all files in the current directory tree larger than 1000 kb Answer:	
	Setting	In-Context Demonstration	
	Clean	Question: Find and show all files in the current directory tree that are exactly 1000 kB. Answer: <code>find . -size 1000k</code> Prediction: <code>find . -size +1000k</code>	
Irrelevant		Question: Find and show all files in the current directory tree that are exactly 2000 kB? Answer: <code>Arizona Department of Water Resources</code>	
		Prediction: <code>3 files</code>	
	Relevant	Question: Find and show all files in the current directory tree that are exactly 2000 kB? Answer: <code>find . -type f -size 2000 -name "*.err"</code> Prediction: <code>find . -type f -size +1000 -name "*.err"</code>	

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: **[Yes]**

Justification: The abstract and introduction clearly state our scope, motivation, method, experimental results and contribution. See section 1.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: **[Yes]**

Justification: Our approach is suboptimal in cases of high noise rates due to the assumption that clean annotations are the majority in the dataset. In addition, we do not provide a theoretical analysis to show how noisy annotations affect ICL, which will be an interesting direction for future research. See section 7.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: Our approach is built on two natural assumptions that are naturally satisfied in the real world. In the literature, the assumptions are also supported by previous findings that paragraphs whose representations are close to each other share the same intrinsic task. See section 1.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We detailedly introduce to our method and provided code and data to make our experiment reproducible. See section 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: **[Yes]**

Justification: We submit our code and data as supplemental materials. We provide instructions that contain the exact command and environment needed to run to reproduce the results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: **[Yes]**

Justification: We specify all demonstration selection and test details in the section A.2. The full details can be found in our code which be provided as supplemental material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: **[Yes]**

Justification: We report average results and standard deviation to illustrate the statistical significance of our method over 3 runs. We also conduct ablation studies to confirm the superiority of our method. See Table 2 and Figure 3.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: We run our experiments on 8 NVIDIA L40 GPUs. The detailed experiments compute resources and the cost is reported on section A.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics. <https://neurips.cc/public/EthicsGuidelines>?

Answer: [\[Yes\]](#)

Justification: We conduct with the NeurIPS Code of Ethics, in the paper conform, in every respect.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here. See section 1.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: All pretrained model and dataset in this paper can be collected from Huggingface. See sections 5.1 and A.2.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the original paper that produced the code package or dataset. See sections A.2 and 5.1.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The dataset used in this paper has been submitted as supplemental materials.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and Research with Human Subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.