In-Context Learning with Iterative Demonstration Selection

Anonymous ACL submission

Abstract

Spurred by advancements in scale, large language models (LLMs) have demonstrated strong few-shot learning ability via in-context learning (ICL). However, the performance of ICL has been shown to be highly sensitive to the selection of few-shot demonstrations. Selecting the most suitable examples as context remains an ongoing challenge and an open problem. Existing literature has highlighted the importance of selecting examples that are diverse or semantically similar to the test sample while ignoring the fact that the optimal selection dimension, *i.e.,* diversity or similarity, is task-specific. Based on how the test sample is answered, we propose Iterative Demonstration Selection (IDS) to leverage the merits of both dimensions. Using zero-shot chain-of-thought reasoning (Zero-shot-CoT), IDS iteratively selects examples that are diverse but still strongly correlated with the test sample as ICL demonstrations. Specifically, IDS applies Zero-shot-CoT to the test sample before demonstration selection. The output reasoning path is then used to choose demonstrations that are prepended to the test sample for inference. The generated answer is followed by its corresponding reasoning path for extracting a new set of demonstrations in the next iteration. After several iterations, IDS adopts majority voting to obtain the final result. Through extensive experiments on tasks including reasoning, question answering, and topic classification, we demonstrate that IDS can consistently outperform existing ICL demonstration selection methods.

1 Introduction

With the recent advancements in scaling up model parameters, large language models (LLMs) showcase promising results on a variety of few-shot tasks through in-context learning (ICL), where the model is expected to directly generate the output of the test sample without updating parameters. This is achieved by conditioning on a manually designed prompt consisting of an optional task description and a few demonstration examples (Brown et al., 2020). Fig. 1 shows an example describing how LLMs perform ICL on the topic classification task. Given a few text-topic pairs as demonstrations, ICL combines them with the test sample as input, to the LLM for inference. The output, *i.e.,* 'Technology', is generated by the model autoregressively without any parameter updates.

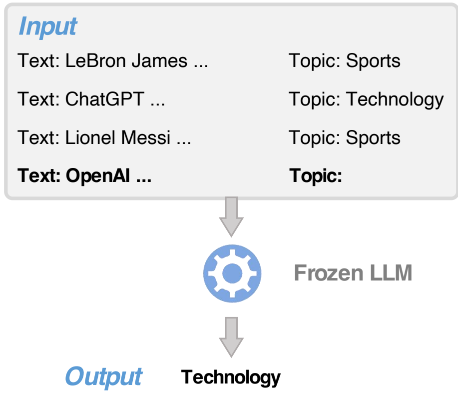

Figure 1: Illustration of in-context learning (ICL) on topic classification. A frozen large language model directly generates the topic 'Technology' for the test sample 'OpenAI ...' by taking the demonstrations and the test sample as input.

Despite the effectiveness, the performance of ICL has been shown to be highly sensitive to the selection of demonstration examples (Zhao et al., 2021). Different sets of demonstrations can yield performance ranging from nearly random to comparable with state-of-the-art models (Gao et al., 2021; Lu et al., 2022). To alleviate the above issue, researchers in ICL have proposed a number of methods to select a set of examples as few-shot demonstrations (Rubin et al., 2022; Liu et al., 2022; Li and Qiu, 2023; Wang et al., 2023b; Li et al., 2023a; Ma et al., 2023; An et al., 2023b). However, for LLMs for which parameters or detailed output distributions are not available (Sun et al., 2022), it is still a common practice to randomly select examples or select examples that are semantically similar to the test sample as demonstrations, *i.e.,* considering diversity or similarity. While several approaches investigate the combination of similarity and diversity when prompting with explanations, exploring compositional generalization, or choosing examples for annotation (Ye et al., 2023b; An et al., 2023a; Su et al., 2023), it is not yet clear how to determine and leverage the optimal dimension for different tasks in ICL and how the rationale for answering the query benefits the balance between these two dimensions.

Actually, the optimal dimension for selecting demonstration examples is task-specific. As we will show in §4, the diversity dimension is superior to the similarity dimension on CommonsenseQA while the similarity dimension outperforms the diversity dimension on AGNews and BoolQ. Thus, it is unreasonable to claim that one dimension is consistently better than the other across different tasks. To fully leverage the merits of both dimensions, we propose Iterative Demonstration Selection (IDS) for ICL (Fig. 2) by utilizing *how the test sample is answered*. IDS can iteratively select demonstration examples that are diverse but still have a strong correlation with the test sample through zero-shot chain-of-thought reasoning (Zero-shot-CoT) (Kojima et al., 2022). Specifically, Zero-shot-CoT, *e.g.,* "Let's think step by step.", is first applied to the test sample before selecting demonstrations to obtain a reasoning path. The training examples that are most semantically similar to the generated reasoning path are then selected as demonstrations. They are prepended to the test sample for inference. Note that IDS ensures that the generated answer is accompanied by the reasoning path through designed prompts. The new reasoning path is then used for extracting another set of demonstration examples by semantic similarity in the next iteration. After a few iterations, IDS adopts majority voting to obtain the final result. Empirical results on tasks spanning mathematical reasoning, commonsense reasoning, logical reasoning, question answering, and topic classification show that IDS can consistently outperform previous ICL demonstration selection baselines. In summary, our main contributions are:

- We consider both the diversity and similarity dimensions of ICL demonstration selection for LLMs. We identify that the optimal dimension for selecting demonstrations is task-specific and propose Iterative Demonstration Selection (IDS) based on how the test query is answered to fully leverage the merits of both dimensions.
- With extensive experiments and analysis, we demonstrate the effectiveness of IDS on a variety of tasks.

2 Related Work

This work mainly explores how to select few-shot in-context learning demonstrations for LLMs by leveraging Zero-shot-CoT. In light of this, we review four lines of research that form the basis of this work: few-shot learning, in-context learning basics, demonstration selection for in-context learning, and chain-of-thought reasoning.

2.1 Few-shot Learning

Few-shot learning aims to learn tasks with only a few labeled samples, which results in a big challenge, *i.e.,* over-fitting, for models as they typically require large amounts of data for training. Prior methods to address over-fitting mainly focused on augmenting the few-shot data (Gao et al., 2020; Qin and Joty, 2022), reducing the hypothesis space (Triantafillou et al., 2017; Hu et al., 2018), or optimizing the strategy for searching the best hypothesis (Ravi and Larochelle, 2017; Finn et al., 2017). More recently, LLMs have demonstrated strong few-shot learning ability through in-context learning without any parameter updates (Brown et al., 2020).

2.2 In-context Learning

Brown et al. (2020) first showed that a frozen GPT-3 model can achieve impressive results on a variety of few-shot NLP tasks through conditioning on manually designed prompts consisting of task descriptions and several demonstration examples. Since then many efforts have been made on in-context learning (ICL) (Dong et al., 2022). Chen et al. (2022); Min et al. (2022a); Wei et al. (2023a) demonstrated that the ICL ability of language models can be further improved through self-supervised or supervised training. Some analytical studies attempted to understand what factors affect ICL performance (Zhao et al., 2021; Shin et al., 2022; Wei et al., 2022a; Min et al., 2022b; Yoo et al., 2022; Wei et al., 2023b) and why ICL works (Xie et al., 2022; Olsson et al., 2022; Li et al., 2023b; Pan et al., 2023; Dai et al., 2023). Other ongoing research on ICL has also explored (*i*) demonstration designing, including demonstration selection (Liu et al., 2022; Rubin et al., 2022; Wang et al., 2023b), demonstration ordering (Lu et al., 2022), and demonstration formatting (Wei et al., 2022b; Wang et al., 2022c; Zhou et al., 2023; Zhang et al., 2023a), (*ii*) applications of ICL (Ding et al., 2022; Meade et al., 2023; Zheng et al., 2023), and (*iii*) ICL beyond text (Wang et al., 2023c; Huang et al., 2023; Zhu et al., 2023; Wang et al., 2023a).

2.3 Demonstration Selection for In-context Learning

The performance of ICL has been shown to be highly sensitive to the selection of demonstration examples (Zhao et al., 2021). Existing methods to solve this problem can be mainly divided into two categories. First, *unsupervised* methods rely on pre-defined metrics. Liu et al. (2022) proposed to select the closest neighbors as demonstrations. In contrast, Levy et al. (2022) selected diverse demonstrations to improve in-context compositional generalization. More recent studies have explored leveraging the output distributions or predictive uncertainty of language models to select few-shot demonstrations (Wu et al., 2022; Nguyen and Wong, 2023; Li and Qiu, 2023; Ma et al., 2023; Ling et al., 2024; Xu and Zhang, 2024) or self-generating demonstrations (Chen et al., 2023). Second, *supervised* methods involve model training. Rubin et al. (2022); Ye et al. (2023a); Li et al. (2023a); Luo et al. (2023); Wang et al. (2024) proposed to learn to retrieve demonstration examples. Wang et al. (2023b) posited LMs as implicit topic models to facilitate demonstration selection. In addition, some studies (Zhang et al., 2022; Scarlatos and Lan, 2023) attempted to select demonstrations based on reinforcement learning. However, it is still a common practice to randomly select examples or select examples that are semantically similar to the test sample as demonstrations for LLMs for which parameters or detailed output distributions are not available (Sun et al., 2022). Several methods investigated the combination of diversity and similarity in different scenarios, *e.g.*, prompting with explanations (Ye et al., 2023b), choosing examples for annotation (Su et al., 2023) and exploring compositional generalization (An et al., 2023a). Nevertheless, it remains unclear to us how to determine and leverage the optimal dimension for different tasks in ICL and how the reason for answering the test sample benefits the balance between the two dimensions, which motivates us to propose our simple but effective approach (IDS).

2.4 Chain-of-Thought Reasoning

Chain-of-thought (CoT) reasoning induces LLMs to produce intermediate reasoning steps before generating the final answer (Wei et al., 2022b). Depending on whether there are manually designed demonstrations, current CoT reasoning methods mainly include Manual-CoT and Zero-shot-CoT. In Manual-CoT, human-labeled reasoning paths are used to perform CoT reasoning (Wei et al., 2022b; Zhou et al., 2022; Wang et al., 2022b; Li et al., 2022; Wang et al., 2022a). In contrast, LLMs leverage self-generated rationales for reasoning in Zero-shot-CoT (Kojima et al., 2022; Zelikman et al., 2022; Zhang et al., 2023a; Diao et al., 2023). The ongoing research on CoT reasoning has also explored (*i*) multimodal reasoning (Zhang et al., 2023b; Wu et al., 2023), (*ii*) distilling knowledge from LLMs (Ho et al., 2022; Fu et al., 2023), and (*iii*) iterative optimization (Shinn et al., 2023; Madaan et al., 2023; Paul et al., 2023).

3 Problem Formulation

Given the test set $\mathcal{D}_{\text{test}}$ and the training set $\mathcal{D}_{\text{train}}$, the goal of ICL demonstration selection is to find an optimal subset $\mathcal{S} = \{(x_1, y_1), ..., (x_k, y_k)\}$ (k-shot) of $\mathcal{D}_{\text{train}}$ as demonstration examples for each test sample $(\hat{x}_i, \hat{y}_i)$ to maximize the overall task performance on $\mathcal{D}_{\text{test}}$. More formally, the optimal selection method $\tilde{h}$ is defined as:

$$\tilde{h} = \arg\max_{h \in \mathcal{H}} \sum_{i=1}^{|\mathcal{D}_{\text{test}}|} \delta_{\text{LLM}([h(\mathcal{D}_{\text{train}}, \hat{x}_i, \hat{y}_i), \hat{x}_i]), \hat{y}_i} \quad (1)$$

where $\mathcal{H}$ is the hypothesis space for searching demonstration examples, $h(\mathcal{D}_{\text{train}}, \hat{x}_i, \hat{y}_i)$ refers to demonstrations selected for $(\hat{x}_i, \hat{y}_i)$ using h, $[,]$ stands for concatenation, and $\delta_{a,b}$ is the Kronecker delta function: $\delta_{a,b} = 1$ if a equals b, otherwise $\delta_{a,b} = 0$. In this work, we aim to find the optimal method $\tilde{h}$ by leveraging Zero-shot-CoT.

4 What Makes Good In-Context Demonstrations?

As demonstrated in previous work (Zhao et al., 2021), the overall task performance is highly sensitive to the selection method h. Different sets of demonstration examples can yield significantly different performance. For example, Zhang et al.

	CommonsenseQA	BoolQ	AGNews
Similar-ICL-Consistency (**Similarity**)	76.0	**85.0**	**90.0**
Random-ICL-Voting (**Diversity**)	**79.0**	84.0	88.0

Table 1: Results of different methods on CommonsenseQA, BoolQ and AGNews. The optimal dimension for selecting ICL demonstrations is task-specific.

(2022) show that the minimum and maximum ICL performance due to random sampling differs by $> 30\%$ on 4 classification tasks, which emphasizes the importance of selecting good demonstrations for LLMs.

A natural question is: what makes good in-context demonstrations? For LLMs, it is still a common practice to select a subset $\mathcal{S}$ consisting of examples that are diverse or semantically similar to the test sample as demonstrations, *i.e.,* considering the diversity or similarity of $\mathcal{S}$. To investigate whether one dimension is consistently better than the other one across different tasks, we conduct some pilot experiments on CommonsenseQA (Talmor et al., 2019), BoolQ (Clark et al., 2019) and AGNews (Zhang et al., 2015). Specifically, we randomly sample 100 examples from the original test set for experiments and conduct 4-shot learning using GPT-3.5 (gpt-3.5-turbo).

Following Zhang et al. (2023a), we use Sentence-BERT (Reimers and Gurevych, 2019) to encode all samples. For each test sample, the Similar-ICL method selects the top-4 similar training data based on cosine similarity while the Random-ICL method randomly samples 4 training examples as few-shot demonstrations. Inspired by Wang et al. (2022b), we apply *self-consistency* with 3 decoding paths (temperature 0.7) to Similar-ICL (named **Similar-ICL-Consistency**) and run Random-ICL 3 times before majority voting (named **Random-ICL-Voting**) to improve the robustness.

The results of different methods on four datasets are reported in Table 1. We can observe that the diversity dimension outperforms the similarity dimension on CommonsenseQA while the similarity dimension is superior to the diversity dimension on BoolQ and AGNews. Therefore, the optimal dimension for selecting demonstration examples is task-specific. Thus, it is unreasonable to claim that one dimension is consistently better than the other one in ICL demonstration selection.

Intuitively, semantically similar examples can help the model correctly answer the test query as they might share similar input-output patterns with the test sample which could unleash GPT-3.5's power of text generation. To further understand why the similarity dimension underperforms the diversity dimension on CommonsenseQA, we present a case study in Table 2. We can see that the answer of the final demonstration example extracted by Similar-ICL-Consistency, *i.e.,* 'most buildings' is also in the options list of the test sample, which misleads the decision process of the model, leading to a wrong answer. In addition, the selected demonstrations might not include enough important information as high similarity also results in redundancy.

Considering the strengths and weaknesses of both dimensions, we aim to design a method that can select demonstration examples that are diverse (minimizing misleading information) but still strongly correlated with the test sample, which is introduced in the next section.

5 Iterative Demonstration Selection

Based on the observations and considerations in §4, we introduce Iterative Demonstration Selection (IDS) for ICL demonstration selection by leveraging *how the test sample is answered* (see Fig. 2 for an illustration). Intuitively, the demonstrations that are similar to the *reason* for answering a sample are strongly correlated with this sample. Therefore, we propose to incorporate zero-shot chain-of-thought reasoning (Zero-shot-CoT) into IDS to iteratively select demonstration examples that are diverse but still have a strong correlation with the test sample.

Specifically, for each test sample $\hat{x}_i$, IDS mainly consists of four steps:

1. We apply **Zero-shot-CoT**, *i.e.,* "Let's think step by step." to the test sample $\hat{x}_i$ before selecting demonstrations to obtain a reasoning path R.

2. The **reasoning path** R is then used to select top-k (k is the number of shot) most semantically similar training examples $\{(x_1, y_1), ..., (x_k, y_k)\}$ as few-shot demonstrations. We use Sentence-BERT (Reimers and Gurevych, 2019) to encode the reasoning path R and training examples to obtain the contextual representations and use cosine similarity to measure the similarity between representations.

3. The selected k training examples $\{(x_1, y_1), ..., (x_k, y_k)\}$ are then prepended to the test sample $\hat{x}_i$ for ICL. During inference, we ensure that the generated answer $\hat{\mathrm{A}}$ is accompanied by its corresponding reasoning path $\hat{\mathrm{R}}$

Similar-ICL-Consistency	**Random-ICL-Voting**
Which choice is the correct answer to the question? **Examples**: **Question**: If you have cleaned off dust here it may be difficult to do your homework where? Answer Choices: (A) desktop (B) closet (C) most buildings (D) surface of earth (E) stove **Answer**: A **Question**: Where is dust likely to be under? Answer Choices: (A) closet (B) ground (C) windowsill (D) attic (E) carpet **Answer**: E **Question**: Where would you find a dustbin that is being used? Answer Choices: (A) utility closet (B) ground (C) cupboard (D) broom closet (E) kitchen **Answer**: E **Question**: Dust accumulates where? Answer Choices: (A) ceiling (B) library (C) surface of earth (D) **most buildings** (E) desktop **Answer**: D The response should follow the format: Answer: {A, B, C, D or E} Here is the test data. **Question**: John wanted to clean all of the dust out of his place before settling down to watch his favorite shows. What might he hardest do dust? Answer Choices: (A) closet (B) under the bed (C) television (D) attic (E) **most buildings**	Which choice is the correct answer to the question? **Examples**: **Question**: She had a busy schedule, she had to run errands and pick up the kids the second she did what? Answer Choices: (A) make time for (B) take money (C) go outdoors (D) leave work (E) field **Answer**: D **Question**: What is the worst outcome of an injury? Answer Choices: (A) cause death (B) cause bleeding (C) falling down (D) become infected (E) claim insurance **Answer**: A **Question**: Mom said that Sarah should stay in bed until she was able to go to school again. What did mom say to Sarah when she tried to get up? Answer Choices: (A) you're sick (B) were sick (C) more rest (D) rest more (E) get back under the covers **Answer**: A **Question**: John got a raise, but he lost rank. Overall, it was a good what? Answer Choices: (A) demotion (B) push down (C) go off strike (D) lower (E) go off strike **Answer**: A The response should follow the format: Answer: {A, B, C, D or E} Here is the test data. **Question**: John wanted to clean all of the dust out of his place before settling down to watch his favorite shows. What might he hardest do dust? Answer Choices: (A) closet (B) under the bed (C) television (D) attic (E) most buildings
Answer: E ✗	**Answer**: D ✓

Table 2: Examples of Similar-ICL-Consistency (first decoding path) and Random-ICL-Voting (first run) for constructing demonstration examples. The upper part is the input to LLMs, including few-shot demonstrations, and the lower part is the predicted answer. Similar-ICL-Consistency gives the wrong answer 'most buildings' which is actually the output of the final demonstration example, indicating that the decision process of the model is misled by this similar sample.

through designed prompts, *e.g.,* "The response should follow the format: Topic: {world, sports, business or technology}\nReason: {reason}". Note that **Zero-shot-CoT** is also applied in this step to improve the quality of generated reasoning paths. After ICL, we go back to Step 2 for *iterations* using the *new* reasoning path $\hat{\mathrm{R}}$.

4. After q rounds of iterations between Step 2 and 3, we adopt **majority voting** on all $\hat{\mathrm{A}}$ to obtain the final result $\hat{\mathrm{A}}_{final}$.

Obviously, the selected demonstration examples are strongly correlated with the original test sample, *i.e.,* achieving similarity, as they are selected by the generated reasoning paths (see Appendix A.4 for quantitative analysis of reasoning paths). And they can be different during iterations to achieve diversity because the reasoning paths vary in different iterations. Note that there is *no* reasoning path in few-shot demonstrations (as shown in the green part in Fig. 2). The reasoning path only exists in the output of LLMs.

Algorithm 1 Selection process of IDS

Require: Training set $\mathcal{D}_{\text{train}}$, test set $\mathcal{D}_{\text{test}}$, LLM$_\theta$, number of demonstrations k, number of iterations q and answer set $\hat{\mathrm{A}}_{all} = \varnothing$

ENCODE all samples in $\mathcal{D}_{\text{train}}$ using Sentence-BERT ▷ Encode training set
for $\hat{x}_i$ in $\mathcal{D}_{\text{test}}$ **do**
  APPLY Zero-shot-CoT to $\hat{x}_i$ to obtain the reasoning path R ▷ Zero-shot-CoT
  for $j = 1, \ldots, q$ **do**
    ENCODE R using Sentence-BERT ▷ Encode reasoning path
    USE R to select top-k most similar examples $\mathcal{S} = \{(x_1, y_1), \ldots, (x_k, y_k)\}$ from $\mathcal{D}_{\text{train}}$ as demonstrations ▷ KNN selection
    $(\hat{\mathrm{A}}, \hat{\mathrm{R}}) = \mathrm{LLM}_\theta(\mathcal{S}, \hat{x}_i)$ ▷ ICL with Zero-shot-CoT
    $\mathrm{R} = \hat{\mathrm{R}}$, $\hat{\mathrm{A}}_{all} = \hat{\mathrm{A}}_{all} \cup \{\hat{\mathrm{A}}\}$ ▷ Update reasoning path and answer set
  end for
  ADOPT majority voting for $\hat{\mathrm{A}}_{all}$ to obtain the final result $\hat{\mathrm{A}}_{final}$ for the test sample $\hat{x}_i$ ▷ Majority voting
end for

In addition, we illustrate the whole selection

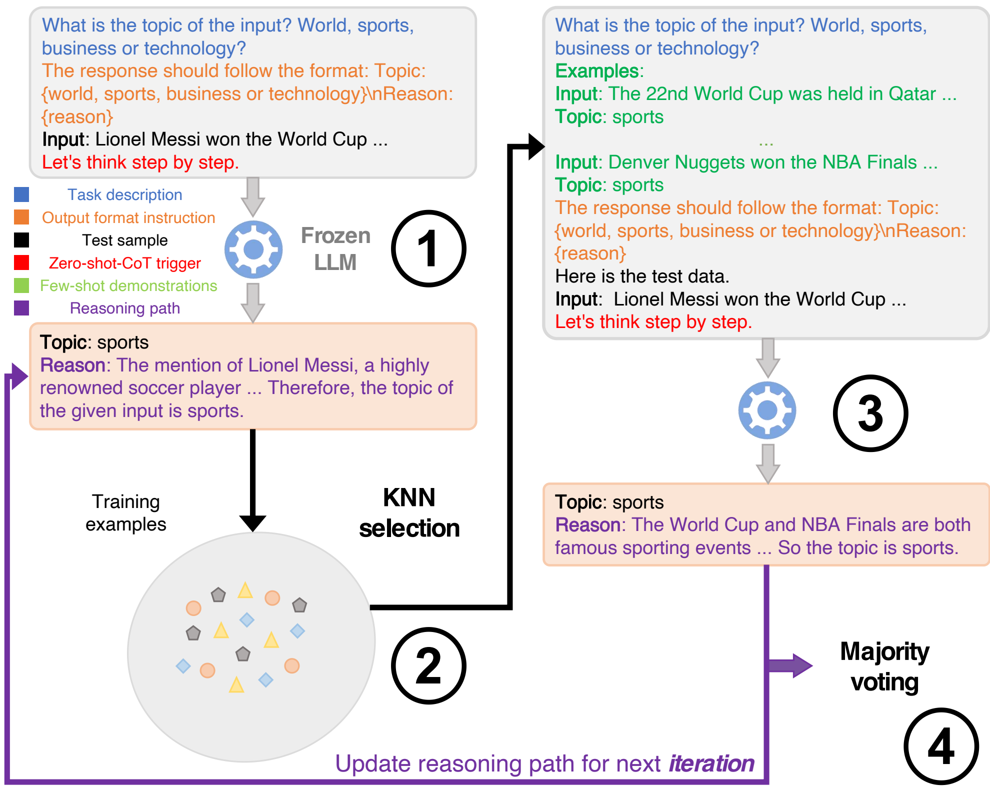

Figure 2: Illustration of our proposed Iterative Demonstration Selection (IDS). IDS first applies Zero-shot-CoT to the test sample to obtain a reasoning path, which is then used to select few-shot demonstrations from training examples through KNN. The selected demonstration examples are prepended to the test sample for ICL. To obtain the new reasoning path for extracting another set of demonstrations in the next iteration, an instruction for output format is inserted before the test sample. After several iterations, IDS uses majority voting to obtain the final result.

process in Alg. 1 and show the instructions and input formats of different types of tasks for ICL in Appendix A.1.

6 Experiments

In this section, we first describe the tasks and datasets, and then introduce methods compared in our work. Finally, we present the experimental results.

6.1 Experimental Setup

Tasks and Datasets We mainly investigate 6 different datasets covering 5 representative task categories: mathematical reasoning (GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021)), commonsense reasoning (CommonsenseQA (Talmor et al., 2019)), logical reasoning (LogiQA (Liu et al., 2020)), question answering (BoolQ (Clark et al., 2019)) and topic classification (AGNews (Zhang et al., 2015)). For each dataset, we randomly sample at most 10000 examples from the original training set as $\mathcal{D}_{\text{train}}$ and at most 2000 test examples as $\mathcal{D}_{\text{test}}$ for evaluating the performance of selected demonstrations. The detailed information of different datasets is shown in Appendix A.2. To reduce the randomness, we run every experiment five times with different random seeds (resulting in different training and test samples if not using the whole set) and report the average results. Without specification, we use $k = 4$ number of demonstrations following Wang et al. (2023b) and set the number of iterations q to 3.

Methods Compared We mainly use GPT-3.5 (gpt-3.5-turbo) as the LLM and compare our IDS with the following methods in the experiments for selecting ICL demonstrations:

- **Top-k-Consistency** (Liu et al., 2022) selects the *top-k* semantically similar examples from the training set $\mathcal{D}_{\text{train}}$ as demonstrations for each test sample and applies *self-consistency* (Wang et al., 2022b) with q decoding paths (temperature 0.7) to match the number of iterations. Following Zhang et al. (2023a), all samples are encoded by Sentence-BERT (Reimers and Gurevych, 2019) to obtain contextual representations for calculating the cosine similarity.

Method	BoolQ	GSM8K	MATH	CommonsenseQA	LogiQA	AGNews	Average
Vote-k	$86.7_{\pm 0.7}$	$76.5_{\pm 0.5}$	$35.7_{\pm 0.2}$	$75.2_{\pm 0.3}$	$45.4_{\pm 0.3}$	$88.1_{\pm 1.2}$	$67.9_{\pm 0.2}$
MMR	$86.4_{\pm 0.8}$	$75.5_{\pm 0.7}$	$34.8_{\pm 0.3}$	$74.9_{\pm 0.2}$	$44.7_{\pm 0.3}$	$87.6_{\pm 1.1}$	$67.3_{\pm 0.3}$
G-fair-Prompting	$84.8_{\pm 0.7}$	$76.9_{\pm 0.6}$	$34.6_{\pm 0.3}$	$75.5_{\pm 0.3}$	$43.8_{\pm 0.4}$	$88.9_{\pm 1.0}$	$67.4_{\pm 0.2}$
Skill-KNN	$85.9_{\pm 0.5}$	$76.5_{\pm 0.3}$	$35.1_{\pm 0.2}$	$75.2_{\pm 0.2}$	$44.6_{\pm 0.2}$	$88.7_{\pm 0.9}$	$67.7_{\pm 0.1}$
Top-k-Consistency	$87.1_{\pm 0.2}$	$76.1_{\pm 0.5}$	$35.6_{\pm 0.3}$	$74.5_{\pm 0.2}$	$45.7_{\pm 0.4}$	$89.3_{\pm 0.8}$	$68.1_{\pm 0.1}$
Random-Voting	$87.3_{\pm 0.6}$	$75.6_{\pm 0.4}$	$35.4_{\pm 0.1}$	$77.0_{\pm 0.2}$	$45.1_{\pm 0.3}$	$87.0_{\pm 1.6}$	$67.9_{\pm 0.2}$
Cluster-Voting	$86.4_{\pm 0.7}$	$76.8_{\pm 0.3}$	$34.9_{\pm 0.4}$	$76.5_{\pm 0.3}$	$44.1_{\pm 0.3}$	$86.8_{\pm 1.2}$	$67.6_{\pm 0.3}$
IDS	$\mathbf{87.8}_{\pm 0.8}$	$\mathbf{78.5}_{\pm 0.4}$	$\mathbf{37.5}_{\pm 0.2}$	$\mathbf{78.1}_{\pm 0.1}$	$\mathbf{46.9}_{\pm 0.2}$	$\mathbf{89.8}_{\pm 0.8}$	$\mathbf{69.8}_{\pm 0.1}$

Table 3: Accuracy (%) of different methods on 6 datasets. **Bold** indicates the best result. IDS is consistently better than all previous baselines.

	Top-k-Consistency	IDS	Random-Voting
Average Similarity Score	0.68	0.48	0.32

Table 4: Average similarity scores between test examples and the corresponding selected demonstrations of three methods (Top-k-Consistency, IDS and Random-Voting).

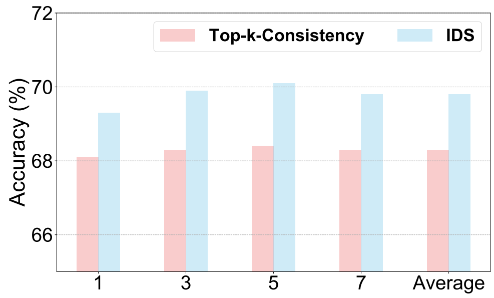

Figure 3: Accuracy (%) of Top-k-Consistency and IDS with different numbers of reasoning paths or iterations.

- **Random-Voting** randomly selects k examples from $\mathcal{D}_{\text{train}}$ as few-shot demonstrations for every test sample and runs experiments q times before majority voting.

- **Cluster-Voting** partitions $\mathcal{D}_{\text{train}}$ into k clusters and selects a representative example from each cluster to form demonstrations. Following Zhang et al. (2023a), we choose the sample closest to the centroid in each cluster as the representative example. Same as Random-Voting, after running experiments q times, Cluster-Voting adopts majority voting to obtain the final result.

Besides, we also compare IDS with several latest ICL demonstration selection approaches: **Vote-k** (Su et al., 2023), **MMR** (Ye et al., 2023b), **G-fair-Prompting** (Ma et al., 2023) and **Skill-KNN** (An et al., 2023b) (see Appendix A.3 for more details of baselines). Similar to Top-k-Consistency, we apply self-consistency to these baselines to match the number of iterations q. Note that we find that simultaneously generating answers and reasoning paths can improve the ICL performance in general even if the target task is not a reasoning task in the conventional sense, *e.g.,* topic classification. Therefore, we apply the same prompt, *e.g.,* "The response should follow the format: Topic: {world, sports, business or technology}\nReason: {reason}", and *Zero-shot-CoT* to baseline methods.

6.2 Main Results

Table 3 shows the average performance scores of different methods on all investigated datasets. From the results, we can observe that

- Our proposed IDS consistently outperforms previous baselines on all datasets with a negligible increase in API request cost (Zero-shot-CoT in the first step), which demonstrates that our method can indeed effectively and efficiently select better ICL demonstration examples by incorporating the reason for answering the test query.
- On average, IDS yields about 1.7% performance boost compared with the best baseline Top-k-Consistency as it can fully leverage the merits of both selection dimensions (diversity and similarity). While the performance gain on a few simple benchmarks looks somewhat small (because the baseline results are already pretty high, *e.g.,* the baseline performance of BoolQ and AGNews is above 85%), IDS performs much better than baselines on more complex tasks. For example, IDS can bring an average relative improvement of about 4% on mathematical reasoning tasks compared with Top-k-Consistency.

To delve deeper into how different dimensions are leveraged in selected demonstrations, we report the average similarity scores between test samples and the corresponding demonstrations of different methods in Table 4. Specifically, we randomly select 500 test examples for each dataset and use Sentence-BERT to obtain contextual representations for calculating similarity scores. We can see

	GPT-3.5	GPT-4
Top-k-Consistency	68.3	73.9
IDS	**69.9**	**75.4**

Table 5: Accuracy (%) of Top-k-Consistency and IDS with different LLMs (GPT-3.5 and GPT-4). For GPT-4, we randomly sample 200 test examples per dataset due to the high cost.

that the average similarity score of IDS is between that of Top-k-Consistency and Random-Voting, indicating that it can indeed strike a balance between two selection dimensions (see Appendix A.5 for more analysis on the diversity of the selected demonstration examples).

6.3 Analysis

Different Numbers of Iterations Our experiments and analysis so far use $q = 3$ iterations. To verify whether the performance gain of IDS is consistent across different numbers of iterations, we conduct controlled experiments with $q = \{1, 5, 7\}$. The average results of the 6 datasets with a randomly selected seed are reported in Fig. 3. IDS consistently outperforms the best baseline Top-k-Consistency with different q (even $q = 1$, *i.e.,* without voting), emphasizing the importance of rationales in selecting demonstration examples. Interestingly, the performance of ICL does not always improve with the number of iterations, which might be because increased iterations can also lead to unnecessary noise; we provide an in-depth analysis in Appendix A.6.

Robustness to Model Types To demonstrate the robustness of IDS to model types, we conduct controlled experiments with GPT-4. Specifically, we randomly select one seed and sample 200 test examples per dataset for experiments due to the expensive cost. From the average results shown in Table 5, we can observe that IDS still achieves better performance than Top-k-Consistency when using GPT-4 as the LLM, showing its robustness to different LLMs.

Generalization to Open-source LLMs To better verify the generalization ability of IDS, we use vLLM (Kwon et al., 2023) to serve Llama-2-chat models (Touvron et al., 2023) for experiments and compare IDS with Top-k-Consistency on two datasets: BoolQ and GSM8K. We randomly sample 500 test examples for experiments and report the results in Table 6, which demonstrates that IDS can successfully generalize to open-source LLMs of different sizes.

	BoolQ			GSM8K		
	7B	13B	70B	7B	13B	70B
Top-k-Consistency	77.1	81.3	84.2	14.6	24.8	49.6
IDS	**78.5**	**82.2**	**85.4**	**16.6**	**27.1**	**51.4**

Table 6: Accuracy (%) of different methods with Llama-2-chat models.

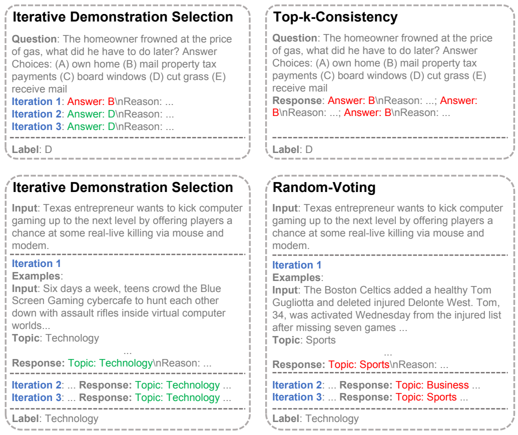

Figure 4: Several case studies of model responses. We color correct outputs in green, and wrong outputs in red.

Case Study To further understand the advantage of IDS, we show several cases in Fig. 4. As shown in the upper part of the figure, IDS can iteratively select more diverse demonstration examples than Top-k-Consistency which may be able to correct errors from previous iterations. Compared with Random-Voting, IDS can find examples that share more similar input-output patterns with the test sample to induce the LLM to generate correct answers (the lower part of the figure).

In addition, we show the results with different numbers of demonstrations, the robustness of IDS to different embedding models and Zero-shot-CoT triggers, and the results on two additional datasets in Appendix A.7 ~ A.10, respectively.

7 Conclusion

In this work, we have introduced Iterative Demonstration Selection (IDS) that can iteratively select examples that are diverse but still strongly correlate with the test sample as demonstrations to improve the performance of in-context learning (ICL) by leveraging the rationale for answering the test sample. Extensive experimental results and analysis show that IDS can consistently outperform previous ICL demonstration selection baselines.

Limitations

This work has several limitations. First, due to the inference cost of ChatGPT, we do not conduct experiments on the entire test set. Besides, we include 6 datasets covering 5 different task types in this work. A further improvement could be to explore more diverse types of tasks.

References

Shengnan An, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Jian-Guang Lou, and Dongmei Zhang. 2023a. How do in-context examples affect compositional generalization? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11027–11052, Toronto, Canada. Association for Computational Linguistics.

Shengnan An, Bo Zhou, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Weizhu Chen, and Jian-Guang Lou. 2023b. Skill-based few-shot selection for in-context learning. *arXiv preprint arXiv:2305.14210*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Mingda Chen, Jingfei Du, Ramakanth Pasunuru, Todor Mihaylov, Srini Iyer, Veselin Stoyanov, and Zornitsa Kozareva. 2022. Improving in-context few-shot learning via self-supervised training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3558–3573, Seattle, United States. Association for Computational Linguistics.

Wei-Lin Chen, Cheng-Kuang Wu, Yun-Nung Chen, and Hsin-Hsi Chen. 2023. Self-ICL: Zero-shot in-context learning with self-generated demonstrations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15651–15662, Singapore. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang. 2023. Active prompting with chain-of-thought for large language models. *arXiv preprint arXiv:2302.12246*.

Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq Joty, and Boyang Li. 2022. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. *arXiv preprint arXiv:2301.12726*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Tianyu Gao, Xu Han, Ruobing Xie, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. 2020. Neural snowball for few-shot relation learning. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7772–7779. AAAI Press.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*.

Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Few-shot charge prediction with discriminative legal attributes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 487–498, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. 2023. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS 2022)*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Itay Levy, Ben Bogin, and Jonathan Berant. 2022. Diverse demonstrations improve in-context compositional generalization. *arXiv preprint arXiv:2212.06800*.

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

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

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2022. On the advance of making language models better reasoners. *arXiv preprint arXiv:2206.02336*.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. 2023b. Transformers as algorithms: Generalization and stability in in-context learning.

Chen Ling, Xujiang Zhao, Wei Cheng, Yanchi Liu, Yiyou Sun, Xuchao Zhang, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, Guangji Bai, Liang Zhao, and Haifeng Chen. 2024. Uncertainty decomposition and quantification for in-context learning of large language models. In *2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

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

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

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

Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. 2023. Fairness-guided few-shot prompting for large language models. *arXiv preprint arXiv:2303.13217*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *Preprint*, arXiv:2303.17651.

Nicholas Meade, Spandana Gella, Devamanyu Hazarika, Prakhar Gupta, Di Jin, Siva Reddy, Yang Liu, and Dilek Hakkani-Tür. 2023. Using in-context learning to improve dialogue safety. *arXiv preprint arXiv:2302.00871*.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022a. 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, Seattle, United States. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022b. 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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tai Nguyen and Eric Wong. 2023. In-context example selection with influences. *arXiv preprint arXiv:2302.11042*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. 2022.

In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. 2023. 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, Toronto, Canada. Association for Computational Linguistics.

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2023. Refiner: Reasoning feedback on intermediate representations. *Preprint*, arXiv:2304.01904.

Chengwei Qin and Shafiq Joty. 2022. Continual few-shot relation learning via embedding space regularization and data augmentation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2776–2789, Dublin, Ireland. Association for Computational Linguistics.

Sachin Ravi and Hugo Larochelle. 2017. Optimization as a model for few-shot learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. 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, Seattle, United States. Association for Computational Linguistics.

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

Alexander Scarlatos and Andrew Lan. 2023. Reticl: Sequential retrieval of in-context examples with reinforcement learning. *arXiv preprint arXiv:2305.14502*.

Seongjin Shin, Sang-Woo Lee, Hwijeen Ahn, Sungdong Kim, HyoungSeok Kim, Boseop Kim, Kyunghyun Cho, Gichang Lee, Woomyoung Park, Jung-Woo Ha, and Nako Sung. 2022. On the effect of pretraining corpora on in-context learning by a large-scale language model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5168–5186, Seattle, United States. Association for Computational Linguistics.

Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Preprint*, arXiv:2303.11366.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023. Selective annotation makes language models better few-shot learners. In *The Eleventh International Conference on Learning Representations*.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. 2022. Black-box tuning for language-model-as-a-service. In *International Conference on Machine Learning*, pages 20841–20855. PMLR.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Eleni Triantafillou, Richard Zemel, and Raquel Urtasun. 2017. Few-shot learning through an information retrieval lens. *arXiv preprint arXiv:1707.02610*.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023a. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Liang Wang, Nan Yang, and Furu Wei. 2024. Learning to retrieve in-context examples for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1752–1767, St. Julian's, Malta. Association for Computational Linguistics.

Xinyi Wang, Wanrong Zhu, and William Yang Wang. 2023b. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. *arXiv preprint arXiv:2301.11916*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022a. Rationale-augmented ensembles in language models. *arXiv preprint arXiv:2207.00747*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022b. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022c. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Zhendong Wang, Yifan Jiang, Yadong Lu, Yelong Shen, Pengcheng He, Weizhu Chen, Zhangyang Wang, and Mingyuan Zhou. 2023c. In-context learning unlocked for diffusion models. *arXiv preprint arXiv:2305.01115*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. In *Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS 2022)*.

Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, et al. 2023a. Symbol tuning improves in-context learning in language models. *arXiv preprint arXiv:2305.08298*.

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

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2022. Self-adaptive in-context learning. *arXiv preprint arXiv:2212.10375*.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.

Shangqing Xu and Chao Zhang. 2024. Misconfidence-based demonstration selection for llm in-context learning. *arXiv preprint arXiv:2401.06301*.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023a. Compositional exemplars for in-context learning. *arXiv preprint arXiv:2302.05698*.

Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Veselin Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2023b. Complementary explanations for effective in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4469–4484, Toronto, Canada. Association for Computational Linguistics.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. 2022. Ground-truth labels matter: A deeper look into input-label demonstrations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2422–2437, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eric Zelikman, Yuhuai Wu, and Noah D Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *arXiv preprint arXiv:2203.14465*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28:649–657.

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

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023a. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023b. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? *arXiv preprint arXiv:2305.12740*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Appendix

A.1 Instructions and Input Formats of Different Tasks

We show the instructions and input formats of different types of tasks for in-context learning in Fig. 5.

A.2 Datasets Information

We show the detailed information of different datasets in Table 7.

A.3 Details of Baselines

In this work, we compare IDS with the following latest ICL demonstration selection approaches:

- **Vote-k** (Su et al., 2023) is an unsupervised, graph-based selective annotation method used for selecting and annotating diverse, representative examples. The annotated examples then serve as a pool for demonstration retrieval.
- **MMR** (Ye et al., 2023b) proposes a maximal marginal relevance-based approach for demonstration selection.
- **G-fair-Prompting** (Ma et al., 2023) leverages greedy search to select the example with the highest fairness score at each step.
- **Skill-KNN** (An et al., 2023b) generates skill-based descriptions for test queries and then uses these descriptions to select similar examples as demonstrations.

A.4 Measure of Reasoning Path Correlation

We report the average similarity score between test samples and the corresponding generated reasoning paths ($\text{score}_{\text{reason}}$), the average similarity score between test samples and randomly selected training examples ($\text{score}_{\text{random}}$), and the average similarity score between test samples and the most similar training examples ($\text{score}_{\text{similar}}$) in Table 8. For each dataset, we randomly select 500 test samples and use Sentence-BERT for similarity calculation. We can observe that $\text{score}_{\text{reason}}$ is slightly worse than $\text{score}_{\text{similar}}$ and much higher than $\text{score}_{\text{random}}$, indicating that the generated reasoning path is indeed strongly correlated with the test sample.

A.5 Analysis on Demonstration Diversity

In addition to the average similarity score between test samples and demonstrations, we further calculate the following metrics for IDS and Top-k-Consistency:

$$Q_S = \sum_{1\leq i<j\leq|S|} g(S_i, S_j)/C(|S|, 2) \quad (2)$$

where S is the set of the selected demonstration examples, and g is the function of measuring similarity. Q calculates the average pairwise similarity score of the demonstrations, which can be used to reflect whether they are diverse from each other. As can be seen from the results in Table 9, the average pairwise similarity score of IDS is much lower than that of Top-k-Consistency, verifying the diversity of demonstration examples selected by IDS.

A.6 Noise Caused by Increased Iterations

As observed from Fig. 3, the performance of ICL does not always improve with the number of iterations. We speculate that this is because too many iterations may also lead to unnecessary noise. As the number of iterations increases, the demonstrations selected in the latest iteration are more likely to have been chosen in previous iterations. Therefore, if these demonstrations result in wrong answers in previous iterations, these errors may be propagated to later iterations, *i.e.*, unnecessary noise caused by increased iterations. To better verify our hypothesis, we calculate (*i*) the proportion of demonstrations selected in iteration 5 or 7 that were also chosen in previous iterations (Prop_{pre}), and (*ii*) the proportion of demonstrations selected in iteration 5 or 7 that were chosen in previous iterations and resulted in wrong answers ($\text{Prop}_{\text{pre}}^{\text{wrong}}$).

Commonsense Reasoning

Which choice is the correct answer to the question?
Examples:
Question: If you poke yourself ... **Answer Choices**: (A) have fun ...
Answer: C
...
The response should follow the format: Answer: {A, B, C, D or E}\nReason: {reason}
Here is the test data.
Question: How can I store ... **Answer Choices**: ...
Let's think step by step.

Topic Classification

What is the topic of the input? World, sports, business or technology?
Examples:
Input: Cavs earn fourth straight win ...
Topic: Sports
...
The response should follow the format: Topic: {world, sports, business or technology}\nReason: {reason}
Here is the test data.
Input: Microsoft intros new mice, keyboards ...
Let's think step by step.

Question Answering

Please answer the question based on the context.
Examples:
Context: Sikma was voted as one of the ...
Question: is jack sikma in the hall of fame
Answer: Yes
...
The response should follow the format: Answer: {yes or no}\nReason: {reason}
Here is the test data.
Context: Blue is a playful female puppy ...
Question: is blue off of blue's clues a girl
Let's think step by step.

Logical Reasoning

Which choice is the correct answer to the question?
Examples:
Context: Li Lin is a civil servant, but not a college graduate.
Question: Which of the following is necessarily true? **Answer Choices**: (A) Not all university ...
Answer: B
...
The response should follow the format: Answer: {A, B, C or D}\nReason: {reason}
Here is the test data.
Context: The people in Harbin are all northerners, and some people in Harbin are not workers.
Question: If the above proposition is true, then which answer must be true? **Answer Choices**: ...
Let's think step by step.

Mathematical Reasoning

Please solve the following mathematical problem.
Examples:
Question: Eric, Ben, and Jack have some money. Eric has $10 less than Ben ...
Answer: The answer is 50
...
The response should follow the format: {reason} The answer is {your answer}
Here is the test data.
Question: Kim raises $320 more than Alexandra, who raises $430, and Maryam raises $400 more than Sarah, who raises $300. How much money did they all raise in total?

Figure 5: Instructions and input formats of five different categories of tasks (topic classification, question answering, commonsense reasoning, logical reasoning, and mathematical reasoning) for ICL. For Zero-shot-CoT in the first step of IDS, there is no demonstration example and the instruction "Here is the test data.".

	BoolQ	GSM8K	MATH	CommonsenseQA	LogiQA	AGNews
# **Training Samples**	9427 (full)	7473(full)	5000	9741 (full)	7376(full)	10000
# **Test Samples**	2000	1000	1000	1221 (full)	500	1000

Table 7: Deailed information of different datasets. # refers to 'the number of' and 'full' means the whole set. Note that different random seeds do not result in different samples if the whole set is used.

	$\text{score}_{\text{reason}}$	$\text{score}_{\text{random}}$	$\text{score}_{\text{similar}}$
Average Similarity Score	0.59	0.32	0.68

Table 8: Comparison between different average similarity scores.

	Top-k-Consistency	IDS
Average Pairwise Similarity	0.55	0.39

Table 9: Comparison of average pairwise similarity scores of demonstrations selected by different methods.

Iteration	5	7
Prop_{pre}	31.9%	**60.4%**
$\text{Prop}_{\text{pre}}^{\text{wrong}}$	13.1%	**38.7%**

Table 10: Comparison between different iterations.

	2	4	6	8
Top-k-Consistency	68.0	68.3	68.5	68.4
IDS	**69.4**	**69.9**	**69.9**	**69.7**

Table 11: Accuracy (%) of Top-k-Consistency and IDS with different numbers of demonstrations k.

We can see from Table 10 that the results of the 7th iteration are much higher than those of the 5th iteration, indicating the correctness of our claim.

A.7 Different Numbers of Demonstrations

While we use $k = 4$ demonstration examples for all experiments, we also evaluate the effectiveness of IDS with different k. We randomly choose one seed for experiments and report the average results of the 6 datasets in Table 11. We can see that IDS consistently outperforms Top-k-Consistency with different numbers of demonstrations. In addition, more demonstrations do **not** guarantee better ICL performance, which is consistent with the observation in Wang et al. (2023b).

A.8 Robustness to Embedding Models

Instead of using Sentence-BERT, we also explore adopting the OpenAI embedding model (text-embedding-ada-002) as the encoder. Specifically,

	BoolQ	CommonsenseQA	GSM8K
Top-k-Consistency	86.0	75.4	75.8
IDS	**87.2**	**78.0**	**77.6**

Table 12: Accuracy (%) of different methods with OpenAI embedding model (text-embedding-ada-002) on three datasets.

	Default	Trigger1	Trigger2
IDS	70.1	70.3	70.0

Table 13: Accuracy (%) of IDS with different Zero-shot-CoT triggers.

we conduct experiments on 3 datasets: BoolQ, CommonsenseQA and GSM8K. For each dataset, we randomly sample 500 test examples and compare IDS with the baseline Top-k-Consistency. The results reported in Table 12 demonstrate IDS's robustness to different embedding models.

A.9 Robustness to Zero-shot-CoT Triggers

To verify the robustness of IDS to Zero-shot-CoT triggers, we conduct controlled experiments with two new triggers: "Let's work this out in a step by step way to be sure we have the right answer." (Trigger1) and "Let's solve this problem step by step" (Trigger2). Specifically, we randomly sample 500 test examples per dataset for experiments and report the average results in Table 13, which demonstrates that IDS is indeed robust to different Zero-shot-CoT triggers.

A.10 Two Additional Datasets

To better demonstrate the generalization ability of IDS, we further conduct experiments on two additional datasets: MNLI (natural language inference) (Williams et al., 2018) and Emotion (emotion classification) (Saravia et al., 2018). The comparison between IDS and the baseline Top-k-Consistency is shown in Table 14, which verifies the strong generalizability of IDS.

	MNLI	Emotion
Top-k-Consistency	65.7	58.1
IDS	**67.4**	**60.3**

Table 14: Results on two additional datasets.