More Samples or More Prompt Inputs? Exploring Effective Few-Shot In-Context Learning for LLMs with In-Context Sampling

Anonymous ACL submission

Abstract

While most existing works on LLM prompting techniques focus only on how to select a better set of data samples inside one single prompt input (In-Context Learning or **ICL**), why can not we design and leverage multiple prompt inputs together to further improve the LLM performance? In this work, we propose In-Context Sampling (**ICS**), a low-resource LLM prompting technique to produce confident predictions by optimizing the construction of multiple ICL prompt inputs. Extensive experiments with two open-source LLMs (FlanT5-XL and Mistral-7B) on four NLI datasets (e-SNLI, Multi-NLI, ANLI, and Contract-NLI) illustrate that ICS can consistently enhance LLM's prediction performance. An in-depth evaluation with three proposed data similarity-based ICS strategies suggests that these strategies can further elevate LLM's performance, which sheds light on a new yet promising future research direction.

1 Introduction

Large Language Models (LLMs) with billions of parameters, such as FLAN-T5 (Chung et al., 2022), LLaMA (Touvron et al., 2023b,d), and Mistral (Jiang et al., 2023), have demonstrated exceptional natural language interpretation capability in terms of understanding versatile prompt inputs[1]. In comparison with much smaller language models like BERT (Devlin et al., 2018) and GPT (Radford et al., 2018), such LLMs can understand not only more complex and detailed task narratives but also a few task examples with annotations within the prompt inputs, namely few-shot In-Context Learning (ICL) (Brown et al., 2020; Shin et al., 2022).

As a prominent prompting strategy to exploit LLMs' task-solving capabilities especially for unseen tasks, ICL inserts a few data examples as well as their corresponding annotations into the prompt input. The data examples, along with their annotations, serve as demonstrations[2] for the targeting task and are expected to facilitate LLMs' better understanding of the task narrative, the expected outputs, and potentially the underlying rationales needed for solving the task. Several recent works

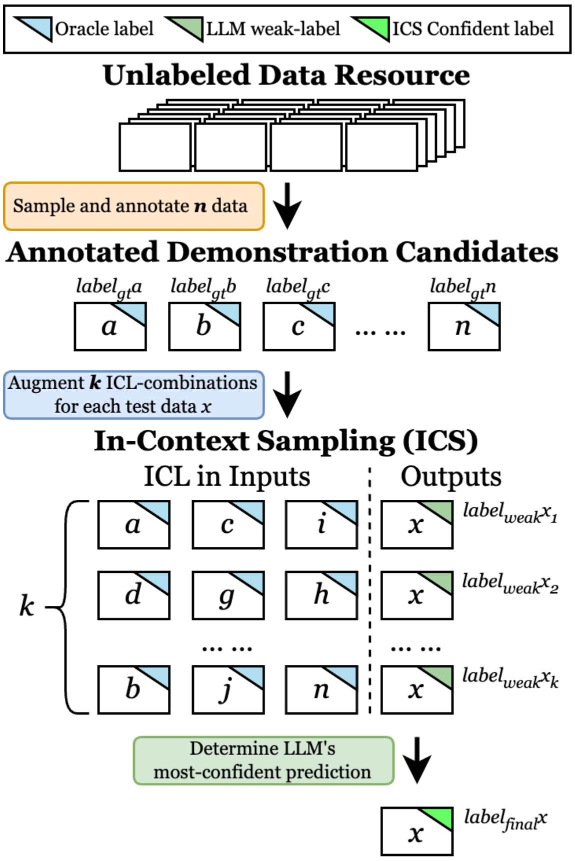


Figure 1: Our proposed ICS paradigm comprises three steps: 1) **sample** representative ICL demonstration candidates, 2) **augment** different ICL prompt inputs from the sampled candidates and acquire LLM's prediction for each input correspondingly, and 3) **vote** and determine LLM's most confident prediction.

[1] We use "prompt input" to refer to the composition of prompt structures, including the task narrative instructions, plus in-context examples, and the targeting data for inference.

[2] We use "examples" and "demonstrations" interchangeably to refer to the few-shot data examples in ICL within the prompt inputs.

investigate the influence of different ICL setups, including the number, ordering, and combinations of demonstrations (Wang et al., 2022; Lu et al., 2022; Yoo et al., 2022). However, there is no common ground for the best ICL strategy yet.

Additionally, despite LLMs' superb natural language interpretation and generation capability, real-world tasks requiring extensive domain expertise remain challenging for LLMs (e.g., children's education and mental issue detection (Chen et al., 2023a; Xu et al., 2023)), and thus, how to exploit LLMs' ability with ICL for solving these tasks is an under-explored topic but holds great promise.

We hypothesize that different ICL demonstrations provide LLMs with distinct knowledge about the task, leading to disparate understanding and predictions for the same data. Consequently, a research question emerges: **Can we augment multiple ICL prompt inputs efficiently to facilitate more accurate and confident LLM predictions?**

To address this question, we propose **In-Context Sampling (ICS)**, a low-resource methodology inspired by the *query-by-committee* strategy (Seung et al., 1992; Liere and Tadepalli, 1997) and the *few-shot In-Context Learning* approach. ICS follows a three-step pipeline as shown in Figure 1:

1. **Sample** demonstration candidates;
2. **Augment** ICL prompt inputs and predictions;
3. **Vote** the most confident label.

We also propose three data similarity-based ICS strategies inspired by established data sampling strategies for Active Learning (Settles, 2009). We believe ICS can be a more reliable prompting paradigm than the traditional ICL, better squeezing LLM's task-solving capabilities and seamlessly supporting "plug-and-play" customizations.

Our evaluation of the ICS paradigm comprises bi-fold. First, we benchmark the effectiveness of a baseline random ICS strategy with the traditional ICL approach with **two** open-source LLMs [3] (FLAN-T5-XL (Chung et al., 2022) and Mistral-7B (Jiang et al., 2023)) over **four** natural language inference (NLI) (Bowman et al., 2015) datasets. The four datasets include three general-domain NLI datasets of increasing difficulty (namely e-SNLI (Camburu et al., 2018), Multi-NLI (Williams et al., 2017), and ANLI (Nie et al., 2019)), and Contract-NLI (Koreeda and Manning, 2021a), a domain-specific NLI dataset for the real-world contract review task. We also investigate how different sample sizes and the number of ICL prompt inputs affect performance enhancement. Results indicate that ICS can consistently improve prediction accuracy and robustness despite LLMs demonstrating different levels of ICL capabilities.

We further investigate the additional advantages provided by three proposed ICS strategies through simulations with the best-performing setting from the previous experiment, compared with the random ICS and traditional ICL approaches on the aforementioned four datasets. Despite being conceptually straightforward, all three types of data-based strategies can effectively and consistently improve LLM performance, leading to a broader research scope to exploit ICS in the future.

2 Related Work

2.1 Large Language Models

Large Language Models (LLMs) (Brown et al., 2020; Touvron et al., 2023a,c; OpenAI, 2023) show impressive capability in understanding free-form instructions and generating high-quality content in a variety of tasks (Wei et al., 2021; Sanh et al., 2021; Chung et al., 2022). For instance, Wei et al. (2021) proposed FLAN-T5, a model trained to follow natural language instruction on over 60 NLP tasks. Ouyang et al. (2022) proposed a pipeline to instruction-finetune LLM with Reinforcement Learning from Human Feedback. In addition, various prompting methods such as Chain-of-Thoughts (Wei et al., 2023; Chung et al., 2022) and In-Context Learning (ICL) (Brown et al., 2020) have been developed to exploit LLMs' potential, where the former technique asks models to generate a sequence of rationales, and the latter methodology allows LLMs to learn from few-shot examples in the input context. Our ICS paradigm extends the traditional ICL approach to improve the performance and confidentiality of LLM predictions.

2.2 In-Context Learning Optimization

Optimizing ICL performance has garnered significant attention recently. Dong et al. (2023) summarized three categories for different ICL optimization approaches: fine-tuning with ICL, ICL sample selection, and analyzing order sensitivity. Fine-tuning with ICL generally requires a significant amount of computing resources and effort to tune model parameters, such that Wei et al. (2021) proposed an instruction tuning method that improves

[3] We also experimented with Llama2 (Touvron et al., 2023d) and discussed its limited performance in Appendix E

both zero-shot and few-shot In-Context Learning performance. Sample selection in ICL has been demonstrated to have a considerable impact on model performance (Zhang et al., 2022b; Rubin et al., 2022; Li et al., 2023). Zhang et al. (2022b) initiated a reinforcement learning technique to select more advantageous samples for in-context demonstration. Rubin et al. (2022) proposed a two-staged method with an unsupervised retriever followed by a supervised model. Some work focused on reducing LLM's ICL order sensitivity issue. Lu et al. (2022) proposed multiple sample sorting methods, while Liu et al. (2022) introduced a method for arranging examples based on their semantic similarity. A few other works attempted to exploit the benefits of the ICL pipeline to improve model performance, better alignment, and minimize reliance on external demonstrations (Yu et al., 2023; Lin et al., 2023; Kim et al., 2022).

2.3 Sampling Strategies

The data sampling strategy is a key element of many low-resource learning paradigms that attempt to select the most representative examples, such as Active Learning (AL) (Settles, 2009). Following established works, the data sampling strategies have been mainly categorized into three categories: *model-based*, *data-based*, and *hybrid* (Settles, 2009; Olsson, 2009; Fu et al., 2013; Schröder and Niekler, 2020; Ren et al., 2021; Zhang et al., 2022c; Schröder et al., 2022).

Model-based strategies aim to find the data with the most model uncertainty (Wang et al., 2017; Zeng et al., 2019). For instance, Margatina et al. (2021) and Zhang et al. (2022a) explored using the divergence of a model's prediction as a measurement of model uncertainty. Data-based strategies, on the other hand, aim to find the most diverse or representative data in the data space (Erdmann et al., 2019; Prabhu et al., 2019; Karamcheti et al., 2021). Such that Deng et al. (2018); Sinha et al. (2019) leveraged adversarial learning to select the most representative data. In contrast to model-based strategies, data-based strategies are generally model-agnostic and demand fewer computational resources but necessitate the analysis of unlabeled samples. Hybrid or ensemble Sampling Strategies integrate various strategy types in unison (Krogh and Vedelsby, 1994; Tang et al., 2002; Melville and Mooney, 2004; Donmez et al., 2007; Zhu et al., 2008; Ambati et al., 2011). For instance, Qian et al. (2020) proposed a combined approach of a diversity-based and an uncertainty-based tactic to benefit from both strategies.

3 ICS Prompting Paradigm

Given a natural language task instruction I and a datum to predict $x \in \mathcal{D}$, LLMs can take the In-Context Learning (ICL) input format, denoted as:

$$\{I + (x_1^{icl}, y_1^{icl}) + ... + (x_m^{icl}, y_m^{icl}) + x\} \qquad (1)$$

where (x_m^{icl}, y_m^{icl}) denotes an oracle-annotated in-context demonstration. We believe in-context demonstrations can provide LLMs with two types of knowledge: 1) **explicit** insights to interpret the task instruction I and expected outputs through $(y_1^{icl}, ..., y_m^{icl})$ and 2) **implicit** guidance for how to solve the task via demonstrations $(x_m^{icl} \rightarrow y_m^{icl})$. We hypothesize that **different sets of ICL demonstrations provide LLMs with disparate implicit knowledge about the task**; thus, LLMs may alter their predictions for the same data x given different ICL prompt inputs, but the predictions will eventually converge to a most confident result.

Our hypothesis stands on the shoulder of the *query-by-committee* (Seung et al., 1992; Liere and Tadepalli, 1997) strategy that has been around for a long time. The original concept is to ask a committee of models to vote on whether the unlabeled data needs to be annotated, where the voting models focus on competing hypotheses. However, most existing works focused on measuring the disagreements among committee models (Engelson and Dagan, 1996; McCallum et al., 1998) and creating different committees with probabilistic and non-probabilistic models (Dagan and Engelson, 1995; Freund and Schapire, 1997; Abe and Mamitsuka, 1998; Melville and Mooney, 2004; Tomanek and Hahn, 2009; Sarawagi and Bhamidipaty, 2002).

In this work, we present **In-Context Sampling (ICS)**, a low-resource paradigm for LLMs through effectively augmenting ICL prompt inputs, as shown in Figure 1. We view the ICS strategy as exploring efficient approaches to create committee ICL prompt inputs and query LLMs for the most confident prediction. ICS consists of three steps:

1. **Sample** demonstration candidates and acquiring oracle annotations,
2. **Augment** prompt inputs and label predictions with different ICL combinations, and
3. **Vote** the most confident label as the final prediction from augmented labels.

Before diving deep into the details of each step in ICS, we want to emphasize that our prototyped ICS strategies in this work are model-agnostic. We will demonstrate the consistent effectiveness of a random baseline ICS strategy over the traditional ICL approach across four datasets and two LLMs in Section 4.1. More importantly, our ICS supports **"plug-and-play"** customizations by switching to different sampling, augmenting, and voting strategies with minimum effort. In addition to justifying the effectiveness of our proposed ICS pipeline and investigating the influence of different factors on performance improvement and robustness, we propose three types of model-agnostic ICS strategies and demonstrate their further improvements over the random ICS pipeline in Section 4.2. The following sections illustrate each ICS step in detail as well as our proposed three data similarity-based ICS strategies: diversity, similarity, and hybrid. We also leave a broad research area to explore strategy variations in future work.

Algorithm 1 Proposed Data-based ICS Strategies

```
function ICS_STRATEGY($D, n, strategy$)   ▷ $D$ : array of data content; $n$ : sample size; $strategy$ : strategy type
    $A \leftarrow (s(D_i, D))_{i\in[1,|D|]}$   ▷ Average score
    $S \leftarrow argsort(A)$   ▷ Descending order
    if $strategy$ = "diversity" then
        $t = \left\lfloor \frac{|D|}{n} \right\rfloor$   ▷ Step
        Return $(S_i)_{\substack{i||t \\ 1\leq i\leq|D|}}$
    else if $strategy$ = "similarity" then
        Return $(S_i)_{i\in[1,n)}$
    else if $strategy$ = "hybrid" then
        $t = \left\lfloor \frac{|D|}{(n/2)} \right\rfloor$
        $R_{div} = (S_i)_{\substack{i||t \\ 1\leq i\leq|D|}}$
        $S' = S \ominus R_{div}$   ▷ Array subtract.
        $R_{sim} = (S'_i)_{i\in[1,n/2)}$
        Return $R_{div} \oplus R_{sim}$   ▷ Array concat.
    end if
end function
```

3.1 Demonstration Candidate Sampling

How to effectively select unlabeled examples to benefit model performance shares the same spirit as the Active Learning (AL) data sampling strategy (Settles, 2009), where an AL strategy iteratively samples few examples for annotation and fine-tuning the model. The AL strategies are often categorized into three types, as illustrated above in Section 2: data diversity-based, model probability-based, and hybrid strategies. Existing work stated that the effectiveness of model-based strategies might differ from model to model (Yao et al., 2023), which could introduce irreverent factors when we benchmark our ICS versus the traditional ICL approach. In this work, we implement three different data similarity-based, model-agnostic strategies for ICS and evaluate their effectiveness in Section 4.2, in addition to the baseline **Random** strategy where we demonstrate the effectiveness compared with traditional ICL approach in Section 4.1. The mathematical notations of our proposed strategies are illustrated in Algorithm 1.

Diversity This strategy adheres to established cluster-based strategies (i.e., core-set) (Sener and Savarese, 2017; Yao et al., 2023), aiming to identify examples **representative of all unlabeled data while maximizing the diversity among these selected instances**. The concept of ensuring data diversity derives from the established *density-weighted* sampling strategies (Settles and Craven, 2008; Shen et al., 2004). They assume the instances that can provide the most helpfulness should be the ones that are representative of the input space (He et al., 2023). In other words, the diversity among selected data should be maximized. Specifically, our strategy calculates the cosine similarity for each data x_i, encoded with sentence-transformer (Reimers and Gurevych, 2019), with the following formula, where embed represents sentence-transformer embedding:

$$s(x_i, \mathcal{D}) = \cos\left(\text{embed}(x_i), \frac{1}{|\mathcal{D}|}\sum_{j=1}^{|\mathcal{D}|}\text{embed}(x_j)\right) \quad (2)$$

Subsequently, we rank the data by similarity score and retrieve n examples with the same interval, ensuring the sampling diversity. for instance, to sample 4 demonstrations from 10 ranked unlabeled data, we choose the 1^{st}, 4^{th}, 7^{th}, and 10^{th} data.

Similarity The similarity strategy shares the same procedure as the diversity strategy of calculating the averaged similarity score for each unlabeled data. Nevertheless, the similarity strategy aims to find examples that are **of the highest similarity to the whole unlabeled training data space** so that the sampled data will most likely be similar to the actual testing data. The underlying concept of this strategy is analogous to a family of *density-weighted* sampling strategies that look for the ones that appear most in the unlabeled data space or are most similar to unlabeled data (Fujii et al., 1999; Xu et al., 2003; Haffari and Sarkar, 2009). We follow the same mathematical procedure 2 above to calculate and rank the unlabeled data by the averaged similarity score. Then, differing from the diversity strategy, we retrieve n highest-ranked examples from the ranked list.

Hybrid Similar to the aforementioned line of ensemble strategies that incorporate different strategies altogether in Section 2, our hybrid strategy expects to benefit from both above-mentioned strategies, which aims to locate examples that are either representative of the sampling space or of the highest similarity to the whole space. Subsequently, this hybrid strategy comprises two steps: first, sample $n/2$ examples following the diversity strategy, then sample $n/2$ examples following the similarity strategy from the remaining list.

3.2 ICL Prompt Inputs Augmentation

As described in Section 3 and shown in Figure 1 above, ICS augments label predictions for the same data by constructing multiple disparate ICL combinations from the demonstration candidates sampled in the previous step. Many recent works (Chen et al., 2023b; Levy et al., 2023; Zhang et al., 2022b; Rubin et al., 2022; Nguyen and Wong, 2023; Lu et al., 2022; Liu et al., 2022) attempted different ICL constructions by altering the demonstrations' numbers, orderings, prompts, or sampling strategies. Nevertheless, there's no commonly recognized best strategy yet, and we believe models will learn disparate implicit guidance for solving the task via different demonstrations. In this work, we utilize four Natural Language Inference (NLI) datasets of varying difficulties and fix **three** as the number of demonstrations per prompt input, consistent with the number of NLI categories.

Still, the computation could be massive if we permutate every combination of the candidates. for example, 50 demonstration candidates can result in $19,600$ 3-demonstration ICL combinations. We believe, however, that ICS does not need every ICL combination to find the model's most confident label. Analogous to the *query-by-committee* concept, where a few representative committee models vote for the best prediction, we plan to investigate a reasonable amount of "committees" (i.e., prompt inputs) that balance between establishing robust and reliable predictions and minimizing costs (i.e., computational resources, time, annotation efforts.

The task of augmenting ICL prompt inputs can be naturally viewed as a variation of the candidate sampling task for the previous step, where the underlying concept for both steps attempts to sample a few examples that could be potentially helpful to LLMs through different approaches. Despite that, the optimal strategy for candidate sampling may not be optimal for augmenting prompt inputs in terms of effectiveness and helpfulness. In this work, we benchmark ICS over traditional ICL with a random strategy for augmenting prompt inputs in Section 4.1. Analogous to the sampling step, we implement and evaluate three similarity-based, model-agnostic strategies proposed in Section 4.2 to iteratively select demonstrations for each prompt input. Specifically, for each data to be predicted, we iteratively sample three demonstrations from the candidate list with a certain strategy for k times, constructing k different prompt inputs and, thus, acquiring k predicted labels. At every iteration, we remove previously selected examples from the candidate list to avoid using each demonstration multiple times. For ICS strategy evaluation, we leverage the best-performing parameters from the benchmark experiment, where n=100 and k=10.

3.3 Confident Prediction Voting

Once we acquire a set of predicted labels from the abovementioned ICS steps for each datum to be predicted, we can apply different voting algorithms to find LLM's most confident prediction. A straightforward design could be a majority vote algorithm to select the prediction with the most appearances among all the predictions for the current data, which is analogous to finding the mode value mathematically: $y^{final} = \text{mode}(y_1^{ics}, ..., y_k^{ics})$, where y_k^{ics} denotes the prediction for each augmented prompt input of data x. In this work, we leverage the majority vote algorithm in our prototyped ICS pipelines. We can further consider the model's different prediction confidences for a more complex algorithm design. Additionally, we can envision ICS to **provide model-confident unsupervised labels** to iteratively fine-tune LLM in resource-deficient scenarios where expert annotations are difficult and expensive to access.

4 Evaluations

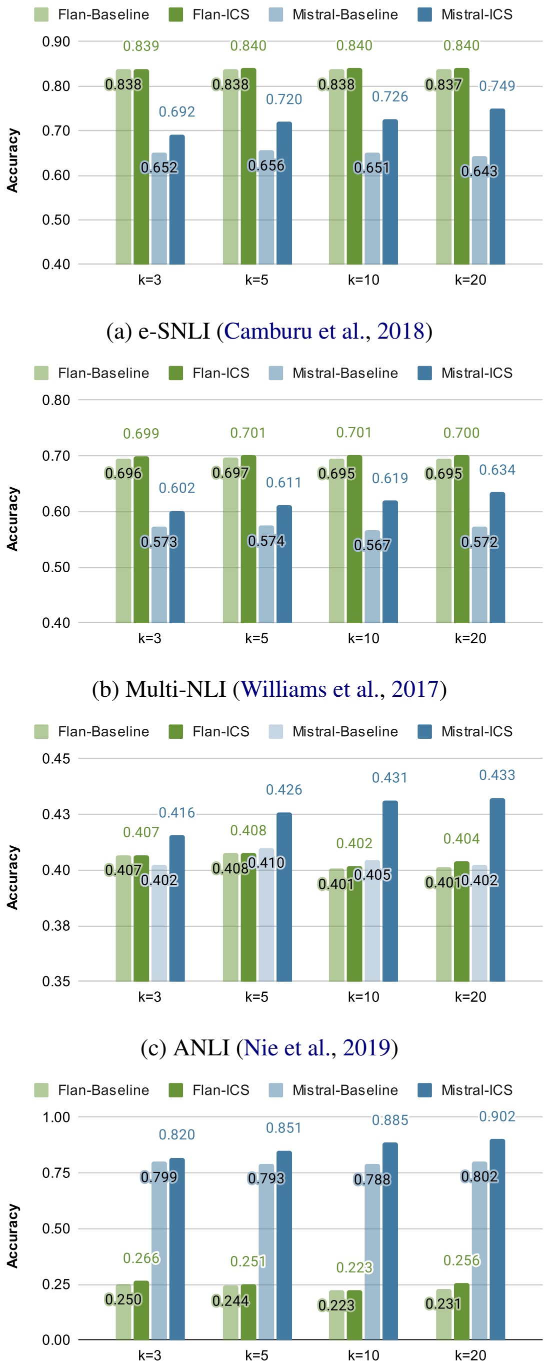
The evaluation of our proposed ICS paradigm comprises bi-fold. First, in Section 4.1, we execute a benchmark experiment between the random ICS strategy and traditional ICL approach on four datasets with two LLMs to demonstrate the paradigm effectiveness. Additionally, we attempt to identify a sample size and the amount of augmented ICL combinations that strike a balance across three perspectives: 1) encompass sufficient diversity to represent the underlying data adequately, 2) possess robustness toward confident predictions, and 3) minimize annotation costs. Subsequently, in Section 4.2, we pick the best-performing parameters from the first experiment to compare the additional advantages of the three proposed ICS strategies described above in Section 3.1.

4.1 Benchmark Evaluation: ICS vs. ICL

4.1.1 Setup

We conduct benchmark experiments to demonstrate the effectiveness of our ICS pipeline with a random sampling strategy for both sampling demonstration candidates and augmenting ICL prompt inputs, compared with the traditional ICL approach with the same amount of demonstrations in each prompt input. Specifically, we employ two open-source LLMs (FLAN-T5-XL (Chung et al., 2022) and Mistral-7B (Jiang et al., 2023)) and experiment on three generic NLI tasks of increasing difficulties: e-SNLI (Camburu et al., 2018), Multi-NLI (Williams et al., 2017), and ANLI (Nie et al., 2019), as well as Contract-NLI (Koreeda and Manning, 2021a), a domain-specific NLI task (dataset statistics in Appendix B). We originally considered Llama2 (Touvron et al., 2023d) but eventually excluded it because our preliminary experiment, discussed in Appendix E, shows that Llama2 tends to output the "neutral" category regardless of the inputs on ANLI.

We intended to manipulate and investigate two controlled variables of ICS: **the size of sampled demonstration candidates** n, where $n \in \{50, 100, 250, 500\}$, and **the number of augmented prompt inputs** k **for each data to be predicted**, where $k \in \{3, 5, 10, 20\}$. We fix the number of demonstrations in each prompt input as three across all methodologies and experiments. The baseline is the vanilla ICL approach with randomly chosen three examples, denoted as $baseline$ in Figure 2 and ICL in tables from Appendix C. We consider 500 annotations a reasonable budget cap for various real-world, low-resource scenarios. Additionally, each setting is repeated and averaged over 10 trials to counter the sampling randomness. All the detailed experiment settings, including the task instruction narrative, are reported in Appendix A.



(a) e-SNLI (Camburu et al., 2018)

(b) Multi-NLI (Williams et al., 2017)

(c) ANLI (Nie et al., 2019)

(d) Contract-NLI (Koreeda and Manning, 2021a)

Figure 2: Benchmark experiment of FLAN-T5-XL and Mistral-7B on four datasets with 100 sampled demonstration candidates (n=100) for random ICS strategy compared with the baseline ICL approach.

4.1.2 Results

The complete evaluation results for every setting are reported in Appendix C as tables for the actual numerical prediction accuracy and in Appendix D as diagrams. We notice that the accuracy improve-

Sampling Strategy	Prompting Strategy	e-SNLI (Camburu et al., 2018)	Multi-NLI (Williams et al., 2017)	ANLI (Nie et al., 2019)	Contract-NLI (Koreeda and Manning, 2021a)
Diversity	Diversity	73.28 (↑ 8.54)	62.10 (↑ 5.20)	**42.78** (↑ 2.36)	87.66 (↑ 8.83)
Diversity	*Random*	73.68 (↑ 8.94)	62.27 (↑ 5.37)	42.77 (↑ 2.35)	89.42 (↑ 10.59)
Random	Diversity	73.47 (↑ 8.73)	61.21 (↑ 4.31)	42.33 (↑ 1.91)	87.53 (↑ 8.70)
Similarity	Similarity	73.63 (↑ 8.89)	61.79 (↑ 4.89)	42.47 (↑ 2.05)	90.44 (↑ 11.61)
Similarity	*Random*	**74.11** (↑ 9.37)	62.09 (↑ 5.19)	42.60 (↑ 2.18)	**90.48** (↑ 11.65)
Random	Similarity	73.74 (↑ 9.00)	62.17 (↑ 5.27)	42.63 (↑ 2.21)	88.88 (↑ 10.05)
Hybrid	Hybrid	73.86 (↑ 9.12)	**62.52** (↑ 5.62)	42.59 (↑ 2.17)	88.85 (↑ 10.02)
Hybrid	*Random*	73.96 (↑ 9.22)	62.41 (↑ 5.51)	42.56 (↑ 2.14)	89.73 (↑ 11.90)
Random	Hybrid	73.95 (↑ 9.21)	62.39 (↑ 5.49)	42.45 (↑ 2.03)	89.06 (↑ 10.23)
Random	*Random*	72.57 (↑ 7.83)	61.17 (↑ 4.27)	42.22 (↑ 1.80)	86.69 (↑ 7.86)
ICL (Baseline)		64.742	56.905	40.420	78.83

Table 1: Comparison of different ICS strategies versus the ICL baseline on four datasets with Mistral-7B (Jiang et al., 2023). We implement different strategy combinations and average each score over 40 trials. The change in prediction accuracy compared with the traditional ICL approach is reported in the parenthesis.

ment becomes insignificant once n goes beyond 100. This observation implies that a sample size over 100 can be considered diverse and representative enough for the NLI task we experimented with, and selecting more data would have only a marginal effect on representativeness. In Figure 2, we present the prediction accuracy of baseline ICL and our ICS strategy for every model and dataset when $n = 100$. We report the prediction accuracy as colored bars, where the green bars denote FLAN-T5-XL and the blue bars denote Mistral-7B.

By comparing the accuracy differences in every diagram between the baseline ICL approach and our ICS strategy for each model, we can observe that ICS, even just benchmarked with a random sampling strategy, can **consistently improve both LLMs' prediction performance** in every (n, k) combination, justifying the validity of our proposed ICS paradigm. It is not difficult to notice that the accuracy improvement provided by the ICS strategy for FLAN-T5-XL is much less than that for Mistral-7B, where Mistral-7B illustrates more than 5% average improvement across all datasets with our ICS strategy. Additionally, we observe that FLAN-T5-XL results in extremely poor performance on Contract-NLI, implying that the model lacks domain knowledge to solve this task. We discuss the potential reasons for the disparate performance between models in Section 5.

4.2 ICS Strategy Evaluation

4.2.1 Setup

Given the observations from the previous benchmark experiment, the best-performing ICS setting in terms of the candidate sampling size and the size of augmented prompt inputs is when n=100 and k=10. In this ICS strategy evaluation experiment, we utilize this set of parameters and further investigate the effectiveness of different ICS strategies we introduced in Section 3.1 over the random ICS and baseline ICL strategies. We implement different ICS strategy combinations to conduct an in-depth analysis of the sampling strategies at each ICS step: sampling demonstration candidates and augmenting the prompt inputs. We determine Mistral-7B as the backbone because it performs higher effectiveness toward ICL and more robust performance on the domain-specific dataset from the benchmark experiment, compared with FLAN-T5-XL.

Because of the massive size of e-SNLI and Multi-NLI (540k and 390k in train splits, correspondingly) , we borrow the concept from Active Learning simulations (Yao et al., 2023) to efficiently evaluate the strategies with a reasonable amount of data and acquire the averaged score over multiple trials. Specifically, for each trial, we randomly sample $3,000$ and $1,000$ data from the train and test split correspondingly as the actual train and test data for the current trial. We then conduct each setting 40 trials to minimize the randomness provided by subsampling training and testing data and report the averaged prediction accuracy in Table 1.

4.2.2 Results

In addition to the prediction accuracy of different ICS strategy combinations, we also report the change in prediction accuracy compared with the baseline ICL approach in the parenthesis, where green denotes improvement. We can easily observe that all three ICS sampling strategies (diversity, similarity, and hybrid) can **consistently and significantly improve the prediction accuracy of Mistral-7B** compared with the baseline setting, with more than 9% improvement on e-SNLI and two-digits elevation on Contract-NLI. It is worth noticing that all the ICS settings with non-random strategies in at least one ICS step can outperform the benchmark ICS setting that utilizes the random strategy for both sampling and prompt augmentation. As we compare the effectiveness across different ICS strategies, we can observe that no single best strategy exists, even for the same NLI task. This observation is aligned with our motivation and the aforementioned existing works that different ICL demonstrations provide distinct knowledge about the task, and there's no single best ICL strategy yet. Specifically, the diversity strategy stands out on ANLI, whereas the hybrid strategy outperforms the other strategies on Multi-NLI, and the similarity strategy surpasses the others on e-SNLI as well as Contract-NLI.

Additionally, we observe that non-random strategies do not lead to consistent performance improvement for augmenting ICL prompt inputs by comparing them with the random strategy. For example, leveraging the random strategy for augmenting prompt inputs outperforms the similarity strategy on all four datasets, implying that **high similarity among the demonstrations within each prompt input is not preferred**. On the other hand, we can observe a significant performance improvement in leveraging non-random strategies demonstration candidate sampling compared to the random strategy, leading to the conclusion that all three strategies demonstrate more contributions during demonstration candidate sampling compared with augmenting ICL prompt inputs. We also hypothesize that more carefully curated strategies are needed to sample ICL combinations effectively, leaving a broader avenue for future research.

Furthermore, we notice **the improvement provided by ICS sampling strategies is inversely proportional to the difficulty of the tasks**. If we consider the model's baseline ICL performance from Section 4.1 as a faithful indicator of dataset difficulty, we can conclude the dataset ordering in ascending order of task difficulty will be e-SNLI, Multi-NLI, and ANLI, where the performance improvement provided by ICS strategies is the smallest on ANLI and the largest on e-SNLI.

Our evaluation of different ICS strategies illustrates promising results that fundamental similarity-based algorithms can effectively increase ICS enhancement, leading to broader future research avenues in exploiting the benefits of more carefully curated ICS strategies with LLMs.

5 Discussion

FLAN-T5-XL We observe FLAN-T5-XL results in poor performance on Contract-NLI from Figure 2, despite it can perform adequately well on the other three generic-domain NLI datasets. We conduct an ablation study with FLAN-T5-XL for ICL to investigate the potential reasons and report in Appendix F. Given the ablation study results, we hypothesize several possible reasons: 1) FLAN-T5-XL falls short of properly interpreting long text sequences; 2) FLAN-T5-XL was not fine-tuned to elevate the ability to interpret ICL demonstrations, and 3) FLAN-T5-XL lacks the necessary domain knowledge to solve the Contract-NLI task.

ICS-Related Work A very recent work attempts multiple ICL methodologies to investigate whether LLMs can beat domain-specific fine-tuned models in the medical domain (Nori et al., 2023). The *Choice Shuffling Ensemble* technique in their proposed ensemble methodology shares a similar concept with our proposed ICS paradigm, but the authors only focus on shuffling the answer choices for selecting robust predictions. Nevertheless, we believe that ICS depicts vast prospects and potential to exploit the capabilities of LLMs.

6 Conclusion

This work presents In-context Sampling (ICS), a novel In-Context Learning paradigm for probing confident predictions by sampling demonstration candidates and augmenting different ICL prompt inputs. Our experiments show that even ICS with the random strategy can lead to consistent accuracy improvement compared with the traditional ICL approach, and further illustrate the additional helpfulness provided by three fundamental but effective data similarity-based sampling strategies with ICS. Our work lays the foundation for implementing ICL-based applications to support non-expert users in the real world, as they do not know how to write a single perfect prompt to get their work done but often write multiple prompt inputs (Zamfirescu-Pereira et al., 2023). Our method aligns well with such user scenarios.

7 Limitations

The primary focus of this paper is to propose and demonstrate the effectiveness of our ICS pipeline compared with the traditional ICL approach. We further illustrate the potential of three proposed similarity-based ICS strategies, which, despite fundamental, can further exploit LLM's capability and boost the prediction performance. Thus, we do not compare with other prompting strategies, such as Chain-of-Thoughts. Our experiments showed that ICS can improve the model's performance (in prediction accuracy) even with a random strategy.

However, despite extensive experiments with different n and k combinations, several potential variables require further analysis. For instance, although we considered four datasets of different difficulties and each ICL combination is arbitrary, all four datasets are NLI tasks. The generalizability of the ICS paradigm to other types of tasks goes beyond the scope of this paper, and we are working on this interesting and substantial research question as a follow-up work.

Besides, we only implement and evaluate the same three strategies for both steps of sampling demonstration candidates and augmenting prompt inputs in ICS because the data similarity-based strategies are model agnostic and generally require fewer computing resources than model-based strategies. We are also aware that the optimal strategy for demonstration candidate sampling may not be optimal for prompt input augmentations, and we leave the analysis of strategy optimization for future work.

In addition, we do not perform an in-depth analysis of optimizing time consumption and reducing computing resources in this work, though we are aware that ICS may require more time than the traditional ICL approach. Lastly, our experiment comprises three open-source LLMs as the original plan but excludes Llama2 due to its over inclination to predict the "neutral" category (Appendix E). We identify that there are still a variety of other instructional-finetuned LLMs we do not include in this work, such as InstructGPT (Ouyang et al., 2022). We also do not include close-sourced and commercial-oriented LLMs such as GPT-4 (OpenAI, 2023) in this work.

References

Naoki Abe and Hiroshi Mamitsuka. 1998. Query learning strategies using boosting and bagging. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, page 1–9, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. 2011. Multi-strategy approaches to active learning for statistical machine translation. In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

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. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, pages 1877–1901, Red Hook, NY, USA. Curran Associates Inc.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Jiaju Chen, Yuxuan Lu, Shao Zhang, Bingsheng Yao, Yuanzhe Dong, Ying Xu, Yunyao Li, Qianwen Wang, Dakuo Wang, and Yuling Sun. 2023a. Fairytalecqa: Integrating a commonsense knowledge graph into children's storybook narratives. *arXiv preprint arXiv:2311.09756*.

Jiuhai Chen, Lichang Chen, Chen Zhu, and Tianyi Zhou. 2023b. How many demonstrations do you need for in-context learning?

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Ido Dagan and Sean P Engelson. 1995. Committee-based sampling for training probabilistic classifiers. In *Machine Learning Proceedings 1995*, pages 150–157. Elsevier.

Yue Deng, KaWai Chen, Yilin Shen, and Hongxia Jin. 2018. Adversarial active learning for sequences labeling and generation. In *IJCAI*, pages 4012–4018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A Survey on In-context Learning.

Pinar Donmez, Jaime G Carbonell, and Paul N Bennett. 2007. Dual strategy active learning. In *Machine Learning: ECML 2007: 18th European Conference on Machine Learning, Warsaw, Poland, September 17-21, 2007. Proceedings 18*, pages 116–127. Springer.

Sean P. Engelson and Ido Dagan. 1996. Minimizing manual annotation cost in supervised training from corpora. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 319–326, Santa Cruz, California, USA. Association for Computational Linguistics.

Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen-Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Béatrice Joyeux-Prunel, and Marie-Catherine de Marneffe. 2019. Practical, efficient, and customizable active learning for named entity recognition in the digital humanities. 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 2223–2234, Minneapolis, Minnesota. Association for Computational Linguistics.

Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

Yifan Fu, Xingquan Zhu, and Bin Li. 2013. A survey on instance selection for active learning. *Knowledge and information systems*, 35:249–283.

Atsushi Fujii, Kentaro Inui, Takenobu Tokunaga, and Hozumi Tanaka. 1999. Selective sampling for example-based word sense disambiguation. *arXiv preprint cs/9910020*.

Gholamreza Haffari and Anoop Sarkar. 2009. Active learning for multilingual statistical machine translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 181–189, Suntec, Singapore. Association for Computational Linguistics.

Jiabang He, Lei Wang, Yi Hu, Ning Liu, Hui Liu, Xing Xu, and Heng Tao Shen. 2023. Icl-d3ie: In-context learning with diverse demonstrations updating for document information extraction.

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. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Siddharth Karamcheti, Ranjay Krishna, Li Fei-Fei, and Christopher Manning. 2021. Mind your outliers! investigating the negative impact of outliers on active learning for visual question answering. 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 7265–7281, Online. Association for Computational Linguistics.

Hyuhng Joon Kim, Hyunsoo Cho, Junyeob Kim, Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2022. Self-Generated In-Context Learning: Leveraging Autoregressive Language Models as a Demonstration Generator.

Yuta Koreeda and Christopher Manning. 2021a. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuta Koreeda and Christopher Manning. 2021b. ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anders Krogh and Jesper Vedelsby. 1994. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems*, 7.

Itay Levy, Ben Bogin, and Jonathan Berant. 2023. Diverse demonstrations improve in-context compositional generalization.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023. Unified Demonstration Retriever for In-Context Learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4644–4668, Toronto, Canada. Association for Computational Linguistics.

Ray Liere and Prasad Tadepalli. 1997. Active learning with committees for text categorization. In *AAAI/IAAI*, pages 591–596. Citeseer.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. The Unlocking Spell on Base LLMs: Rethinking Alignment via In-Context Learning.

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.

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.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Andrew McCallum, Kamal Nigam, et al. 1998. Employing em and pool-based active learning for text classification. In *ICML*, volume 98, pages 350–358. Citeseer.

Prem Melville and Raymond J Mooney. 2004. Diverse ensembles for active learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 74.

Tai Nguyen and Eric Wong. 2023. In-context example selection with influences.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

Fredrik Olsson. 2009. A literature survey of active machine learning in the context of natural language processing.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Ameya Prabhu, Charles Dognin, and Maneesh Singh. 2019. Sampling bias in deep active classification: An empirical study. 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 4058–4068, Hong Kong, China. Association for Computational Linguistics.

Kun Qian, Poornima Chozhiyath Raman, Yunyao Li, and Lucian Popa. 2020. Learning structured representations of entity names using Active Learning and weak supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6376–6383, Online. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

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

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. 2021. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40.

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.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Sunita Sarawagi and Anuradha Bhamidipaty. 2002. Interactive deduplication using active learning. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, page 269–278, New York, NY, USA. Association for Computing Machinery.

Christopher Schröder and Andreas Niekler. 2020. A survey of active learning for text classification using deep neural networks. *arXiv preprint arXiv:2008.07267*.

Christopher Schröder, Andreas Niekler, and Martin Potthast. 2022. Revisiting uncertainty-based query strategies for active learning with transformers. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2194–2203, Dublin, Ireland. Association for Computational Linguistics.

Ozan Sener and Silvio Savarese. 2017. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*.

Burr Settles. 2009. Active learning literature survey.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *proceedings of the 2008 conference on empirical methods in natural language processing*, pages 1070–1079.

H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. 1992. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294.

Dan Shen, Jie Zhang, Jian Su, Guodong Zhou, and Chew-Lim Tan. 2004. Multi-criteria-based active learning for named entity recognition. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 589–596, Barcelona, Spain.

Seongjin Shin, Sang-Woo Lee, Hwijeen Ahn, Sungdong Kim, HyoungSeok Kim, Boseop Kim, Kyunghyun Cho, Gichang Lee, Woomyoung Park, Jung-Woo Ha, et al. 2022. On the effect of pretraining corpora on in-context learning by a large-scale language model. *arXiv preprint arXiv:2204.13509*.

Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. 2019. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5972–5981.

Min Tang, Xiaoqiang Luo, and Salim Roukos. 2002. Active learning for statistical natural language parsing. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 120–127, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Katrin Tomanek and Udo Hahn. 2009. Reducing class imbalance during active learning for named entity annotation. In *Proceedings of the fifth international conference on Knowledge capture*, pages 105–112.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and Efficient Foundation Language Models.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023c. Llama 2: Open Foundation and Fine-Tuned Chat Models.

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

Boshi Wang, Xiang Deng, and Huan Sun. 2022. Iteratively prompt pre-trained language models for chain of thought. *arXiv preprint arXiv:2203.08383*.

Chenguang Wang, Laura Chiticariu, and Yunyao Li. 2017. Active learning for black-box semantic role labeling with neural factors. In *IJCAI*, pages 2908–2914.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.

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

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2023. Mental-llm: Leveraging large language models for mental health prediction via online text data.

Zhao Xu, Kai Yu, Volker Tresp, Xiaowei Xu, and Jizhi Wang. 2003. Representative sampling for text classification using support vector machines. In *Advances in Information Retrieval: 25th European Conference on IR Research, ECIR 2003, Pisa, Italy, April 14–16, 2003. Proceedings 25*, pages 393–407. Springer.

Bingsheng Yao, Ishan Jindal, Lucian Popa, Yannis Katsis, Sayan Ghosh, Lihong He, Yuxuan Lu, Shashank Srivastava, Yunyao Li, James Hendler, and Dakuo Wang. 2023. Beyond labels: Empowering human annotators with natural language explanations through a novel active-learning architecture.

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. *arXiv preprint arXiv:2205.12685*.

Yue Yu, Rongzhi Zhang, Ran Xu, Jieyu Zhang, Jiaming Shen, and Chao Zhang. 2023. Cold-start data selection for better few-shot language model fine-tuning: A prompt-based uncertainty propagation approach. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2499–2521, Toronto, Canada. Association for Computational Linguistics.

J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can't prompt: How non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Xiangkai Zeng, Sarthak Garg, Rajen Chatterjee, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Empirical evaluation of active learning techniques for neural MT. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 84–93, Hong Kong, China. Association for Computational Linguistics.

Shujian Zhang, Chengyue Gong, Xingchao Liu, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. 2022a. ALLSH: Active learning guided by local sensitivity and hardness. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1328–1342, Seattle, United States. Association for Computational Linguistics.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022b. Active Example Selection for In-Context Learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022c. A survey of active learning for natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6166–6190, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K Tsou. 2008. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1137–1144, Manchester, UK. Coling 2008 Organizing Committee.

A Experiment Setup

All four datasets included in our work (e-SNLI, Multi-NLI, ANLI, and Contract-NLI) are of the same natural language inference task. Thus, we leverage the same instruction narrative across all the experiments: **Determine whether a hypothesis is *entailment, neutral, contradiction* giving a premise.** For Contract-NLI, the original dataset only consists of annotations for the "entailment" and "contradiction" categories. Thus, we only evaluate the performance of those data.

All the experiments are computed on one of two resources: 1) an NVIDIA A100 40G graphic card or 2) an NVIDIA 3090 24G graphic card. For Llama2 and Mistral-7B, we load both models in fp16 precision to fit them in both graphic cards and limit to generate a maximum of 10 tokens.

B Dataset Statistics

Dataset	Train	Validation	Test
e-SNLI Camburu et al. (2018)	$549,367$	$9,842$	$9,824$
Multi-NLI Williams et al. (2017)	$392,702$	$9,815$	$9,832$
ANLI Nie et al. (2019)	$16,946$	$1,000$	$1,000$
Contract-NLI Koreeda and Manning (2021b)	$3,999$	555	$1,113$

Table 2: Datasets involved in our experiment. Contract-NLI only comprises annotations of "entailment" and "contradiction" categories.

C Complete Evaluation Results

Here, we report the complete results of our evaluation (Section 4) in Table 5 and Table 6 for FlanT5-XL (Chung et al., 2022) and Mistral-7B (Jiang et al., 2023), correspondingly. We acquire an average prediction accuracy score over 10 trials of each setting. n denotes the amount of demonstration candidate data we sampled, and k denotes the number of ICL combinations for each test data.

D Results Diagrams

Additionally, we plot the LLM's prediction accuracy and the standard deviation across 10 experiment trials for different settings in Figure 3, 4, 5, 6 on e-SNLI, Multi-NLI, ANLI, and Contract-NLI. We can observe that the ICS strategy can consistently improve LLMs' performance compared with the traditional ICL baseline; in addition, FLAN-T5-XL is much less sensitive than Mistral toward the improvement provided by the ICS strategy. From the diagrams, $k = 10$ and $n = 100$ are the best-performing parameters that maximize the performance improvement and minimize the standard deviations.

E Analysis on Llama2

Llama2	**Inst. 1**	**Inst. 2**	**Inst. 3**	**Ground-truth**
entailment	75	202	151	334
neutral	808	668	785	333
contradiction	117	130	64	333

Table 3: Analysis of Llama2 performance on ANLI.

We conducted an initial inference experiment with Llama2 (Touvron et al., 2023d) on ANLI utilizing three different natural language instructions:

i Determine whether a hypothesis is *entailment, neutral, contradiction* giving a premise.

ii Classifying a pair of premise and hypothesis sentences into three classes: *entailment, neutral, contradiction*

iii Predict the relationship between the premise and hypothesis by *entailment, neutral, contradiction*

The results are reported in Table 3. We can easily observe that Llama2 tends to overly predict "neutral" over the other two categories despite changing instruction narratives, whereas the ground-truth distribution is even across categories. Thus, we omit Llama2 in our work. There could be different reasons contributing to this issue; for example, Llama2 was overfitted to the NLI task or similar tasks that share the same set of targeting categories: "entailment", "neutral", and "contradiction".

F Ablation on FLAN-T5-XL with Contract-NLI

We design and conduct an ablation study with FLAN-T5-XL for ICL to verify our hypothesis. The experiment is conducted on the Contract-NLI dataset. Specifically, we start with the zero-shot setting to examine whether FLAN-T5-XL can properly solve the task without demonstrations. Then, we experiment with both ICS and ICL approaches and gradually increase the number of demonstrations from 1 to 3. The demonstrations are randomly selected from the training split, and each ICL setting is repeated 3 times to acquire the average score.

Setting	zero-shot	1-shot	2-shot	3-shot
ICL	2.48	19.39	23.80	22.88
ICS	/	20.03	24.54	23.34

Table 4: ICL ablation experiment of FLAN-T5-XL on Contract-NLI.

From table 4, we can observe that FLAN-T5-XL can hardly interpret the dataset and solve it with a zero-shot setting. Since we leverage the same prompt narrative as the one for the other NLI tasks that FLAN-T5-XL performs relatively well, we can imply that the lack of domain knowledge might be the primary reason for such low performance. Nevertheless, we can observe that the 1-shot setting can significantly improve the model performance, although the overall accuracy is still very low. It is worth noticing that the improvement becomes relatively trivial once we add more demonstrations to the prompt inputs, which implies that FLAN-T5-XL falls short of interpreting longer and more complex ICL format, possibly due to its relatively short training input length limit. Moreover, our random ICS strategy can still outperform the ICL baseline across all settings.

FlanT5-XL		k=3	k=5	k=10	k=20
$n = 50$	ICL	83.82	83.82	83.77	83.79
	ICS_{ours}	83.99	84.00	84.06	84.10
$n = 100$	ICL	83.79	83.77	83.77	83.73
	ICS_{ours}	83.91	83.99	84.04	84.03
$n = 250$	ICL	83.80	83.76	83.80	83.79
	ICS_{ours}	83.94	83.95	84.06	84.04
$n = 500$	ICL	83.75	83.79	83.81	83.64
	ICS_{ours}	83.90	84.06	84.09	84.11

(a) e-SNLI

FlanT5-XL		k=3	k=5	k=10	k=20
$n = 50$	ICL	69.58	69.43	69.47	69.64
	ICS_{ours}	69.87	69.92	69.97	70.13
$n = 100$	ICL	69.58	69.66	69.53	69.51
	ICS_{ours}	69.91	70.05	70.13	70.04
$n = 250$	ICL	69.71	69.54	69.54	69.69
	ICS_{ours}	70.02	70.06	70.11	70.10
$n = 500$	ICL	69.59	69.51	69.41	69.58
	ICS_{ours}	69.90	70.00	70.08	70.16

(b) Multi-NLI

FlanT5-XL		k=3	k=5	k=10	k=20
$n = 50$	ICL	40.38	39.95	40.47	40.15
	ICS_{ours}	40.52	40.40	40.56	40.31
$n = 100$	ICL	40.68	40.75	40.05	40.12
	ICS_{ours}	40.68	40.76	40.18	40.39
$n = 250$	ICL	40.35	40.25	40.02	40.10
	ICS_{ours}	40.36	40.28	40.50	40.38
$n = 500$	ICL	40.10	40.22	40.65	40.11
	ICS_{ours}	40.41	40.49	40.66	40.58

(c) ANLI

FlanT5-XL		k=3	k=5	k=10	k=20
$n = 50$	ICL	24.44	23.98	23.16	23.62
	ICS_{ours}	24.72	24.44	24.72	25.18
$n = 100$	ICL	25.00	24.35	22.33	23.06
	ICS_{ours}	26.56	25.09	22.33	25.55
$n = 250$	ICL	24.44	23.98	24.08	23.34
	ICS_{ours}	26.19	25.55	25.00	25.64
$n = 500$	ICL	23.80	23.52	23.89	23.25
	ICS_{ours}	25.18	24.72	25.09	25.64

(d) Contract-NLI

Table 5: Evaluation result for FlanT5-XL.

Mistral-7B		k=3	k=5	k=10	k=20
$n = 50$	ICL	64.26	64.61	65.01	64.71
	ICS_{ours}	68.82	70.28	71.97	73.82
$n = 100$	ICL	65.15	65.59	65.11	64.30
	ICS_{ours}	69.17	71.95	72.57	74.94
$n = 250$	ICL	65.09	65.02	64.25	64.73
	ICS_{ours}	69.76	71.82	73.23	74.95
$n = 500$	ICL	64.76	64.97	64.66	64.97
	ICS_{ours}	69.21	71.36	73.56	75.06

(a) e-SNLI

Mistral-7B		k=3	k=5	k=10	k=20
$n = 50$	ICL	57.38	57.58	57.82	56.62
	ICS_{ours}	60.31	61.50	61.90	60.56
$n = 100$	ICL	57.26	57.43	56.68	57.24
	ICS_{ours}	60.15	61.13	61.90	63.42
$n = 250$	ICL	57.32	56.53	57.01	56.70
	ICS_{ours}	60.26	60.57	62.82	62.62
$n = 500$	ICL	56.88	59.38	56.96	56.70
	ICS_{ours}	59.78	60.93	62.11	61.96

(b) Multi-NLI

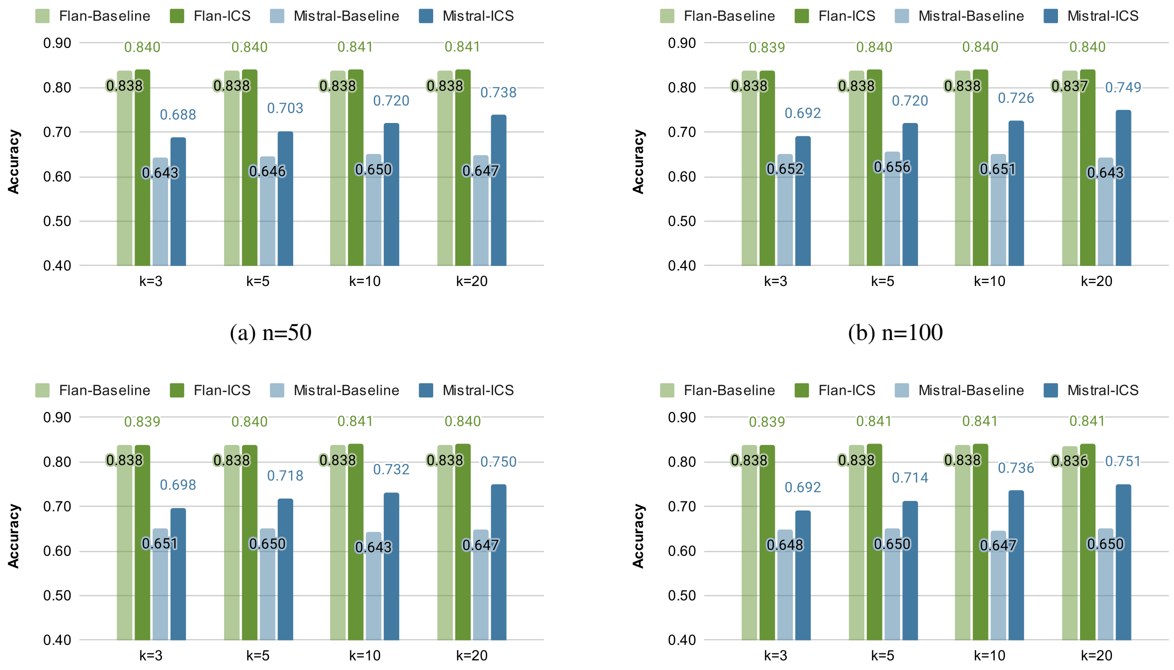
Mistral-7B		k=3	k=5	k=10	k=20
$n = 50$	ICL	40.57	40.25	40.71	41.27
	ICS_{ours}	42.03	42.18	42.31	42.77
$n = 100$	ICL	40.23	41.01	40.47	40.22
	ICS_{ours}	41.56	42.59	43.13	43.25
$n = 250$	ICL	40.81	40.51	40.14	41.28
	ICS_{ours}	42.09	41.90	43.26	43.27
$n = 500$	ICL	40.52	40.57	41.24	39.86
	ICS_{ours}	40.95	42.20	42.85	42.65

(c) ANLI

Mistral-7B		k=3	k=5	k=10	k=20
$n = 50$	ICL	78.90	80.25	79.97	77.05
	ICS_{ours}	80.97	85.01	88.23	89.37
$n = 100$	ICL	79.92	79.33	78.83	80.18
	ICS_{ours}	82.02	85.05	88.54	90.20
$n = 250$	ICL	79.62	78.64	79.12	78.02
	ICS_{ours}	89.63	84.99	88.43	90.83
$n = 500$	ICL	79.62	78.61	79.15	78.70
	ICS_{ours}	81.86	84.29	88.03	90.38

(d) Contract-NLI

Table 6: Our evaluation result for Mistral-7B.



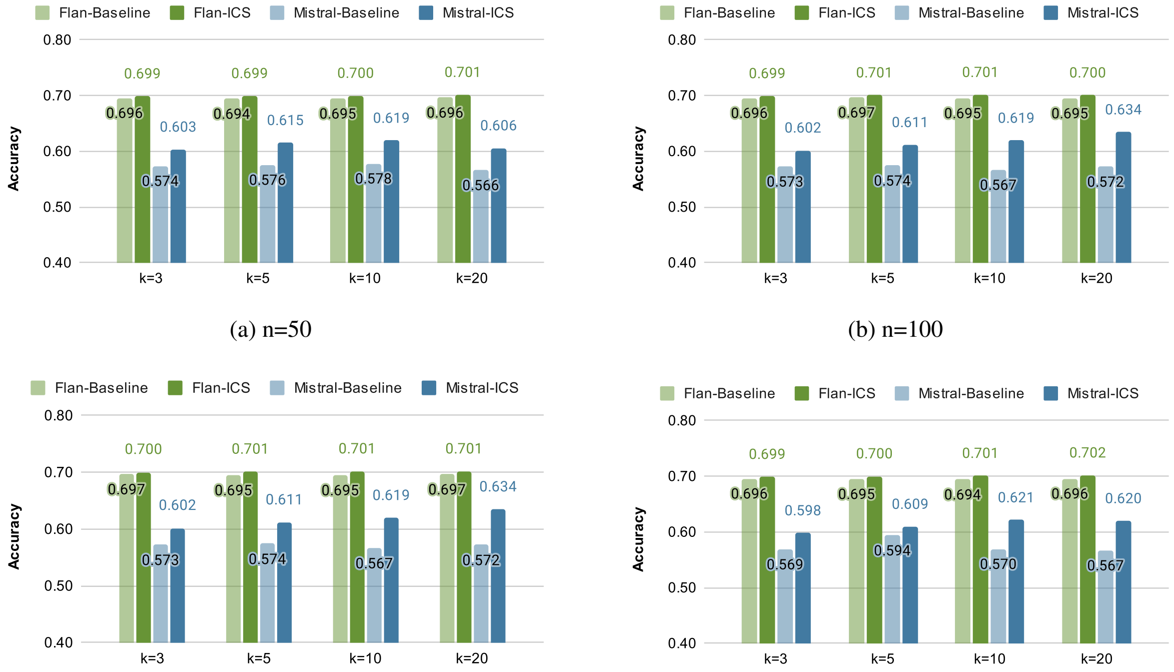
(a) n=50

(b) n=100

(c) n=250

(d) n=500

Figure 3: Evaluation results with FlanT5-XL and Mistral on e-SNLI (Camburu et al., 2018) dataset.



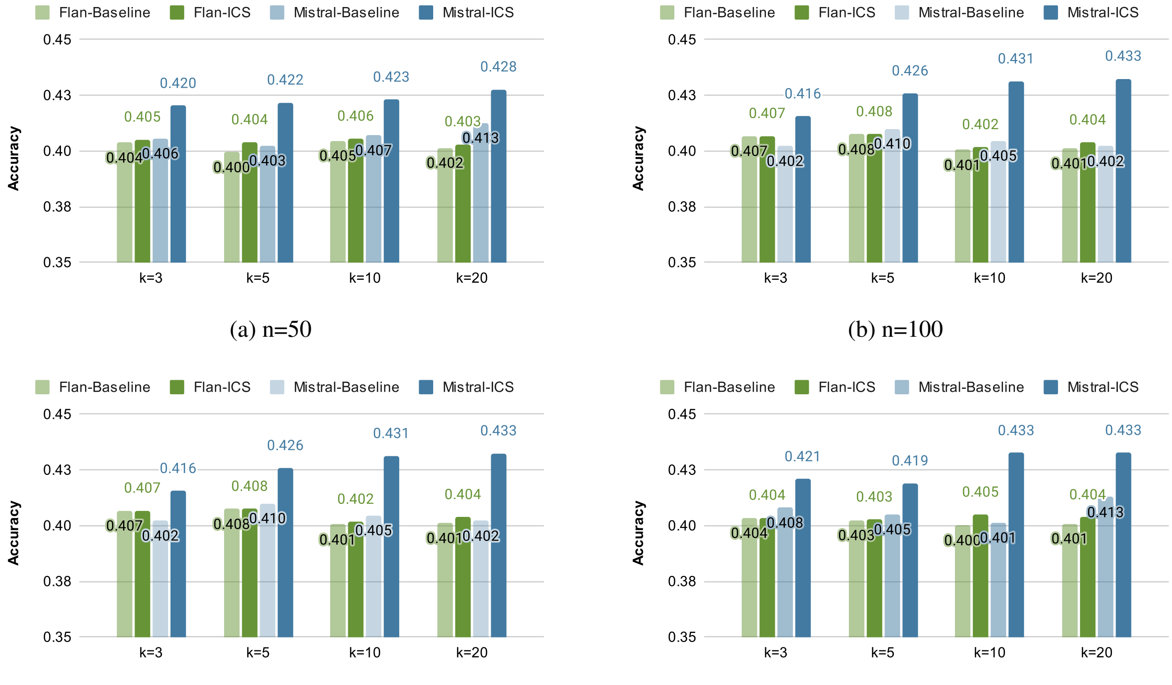
(a) n=50

(b) n=100

(c) n=250

(d) n=500

Figure 4: Evaluation results with FlanT5-XL and Mistral on Multi-NLI (Williams et al., 2017) dataset.



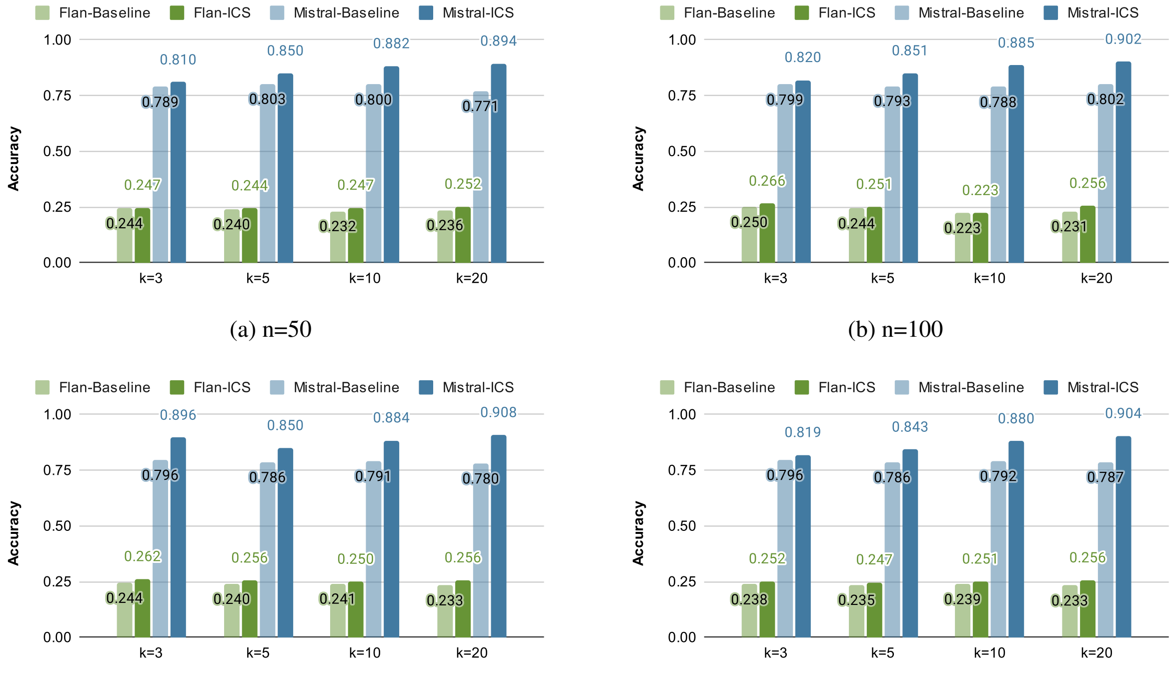
(a) n=50

(b) n=100

(c) n=250

(d) n=500

Figure 5: Evaluation results with FlanT5-XL and Mistral on ANLI (Nie et al., 2019) dataset.



(a) n=50

(b) n=100

(c) n=250

(d) n=500

Figure 6: Evaluation results with Mistral on Contract-NLI (Koreeda and Manning, 2021b) dataset.