Optimizing Demonstration Selection for In-Context Learning using Data Maps

2023 - 2024

Master Computing Science, Utrecht University

Thesis Project

Student name: Silin Chen

Student number: 7217439

First examiner: Dong Nguyen

Second examiner: Albert Gatt

Daily supervisor: Yupei Du

Acknowledgments

I would like to express my sincere gratitude to the following individuals for their valuable contributions to this research.

First of all, I would like to thank my first supervisor, Nguyen Dong, for introducing me to this topic and guiding me into the fascinating field of in-context learning. Her insightful feedback on each draft of my research plan and thesis has been invaluable throughout this year.

Second, I am also grateful to Albert Gatt, my second supervisor, for his thoughtful suggestions on the final versions of my research plan and thesis, as well as his support during my defense. His expertise and attention to detail have greatly improved my work.

Finally, I would like to thank my daily supervisor, Yupei Du, for his continuous support. His suggestions on my experimental code and recommendations for relevant literature have been immensely helpful. I am also deeply thankful for his meticulous feedback and revisions on every draft of my research plan and thesis.

Their guidance has been crucial to my academic development, and I am deeply grateful for their dedication and encouragement. It has been an honor to be mentored by them.

Abstract

In-context learning is a technique in which a model leverages demonstrations provided in the input context to perform tasks, without requiring parameter updates. However, existing selection methods that require the selection of a specialised demonstration set for each query impose significant computational overhead. Inspired by Data Maps (Swayamdipta et al., 2020), this thesis proposes an alternative approach to improve in-context learning by categorizing the dataset into three regions: easy-to-learn, ambiguous, and hard-to-learn. Results indicate that demonstrations from the ambiguous region offer more effective and stable support for in-context learning. Moreover, this approach reduces the size of the dataset to 33% of its original volume while retaining high-quality demonstrations, thereby improving efficiency without compromising performance.

Contents

1 Introduction

Large Language Models (LLMs) have shown remarkable performance in different downstream tasks (Brown et al., 2020; Dong et al., 2024). However, the traditional fine-tuning method requires updating the parameters of the whole model, which is both time-consuming and resource-intensive (Isik et al., 2024). This process requires powerful computational power, often requiring high-performance GPU clusters, which can take hours to days even with sufficient resources, delaying the practical application of the model. In addition, the traditional fine-tuning method requires a substantial amount of labeled data to achieve good results, as fine-tuning relies on a large and high-quality dataset, which can be expensive to obtain (Wang et al., 2020; Liu et al., 2023; Ziegler et al., 2019). Moreover, fine-tuning has limitations in enabling models to handle a diverse range of tasks, highlighting the importance of following exemplars and instructions for real-world applications (Gao et al., 2023; Xi et al., 2023; Liu et al., 2022a). With the expansion of pre-training datasets and model parameters (Brown et al., 2020), LLMs have increasingly gained the ability to perform In-Context Learning (ICL), enabling them to infer effectively in different tasks based on provided context.

In-Context Learning In-context learning, first introduced by Brown et al. (2020), is a technique in which models learn to perform tasks by observing selected demonstrations provided directly within the prompt. First, the human-designed template will turn pairs of text and label from the training corpus into a uniform format. Next, demonstrations are selected from the training dataset using various methods and concatenated with a test query (an instance from the test dataset). Finally, the LLMs will perform inference based on the demonstrations and make further predictions for the test query. As illustrated in Figure 1, given a Natural Language Inference classification dataset **Recognizing Textual Entailment (RTE)** (Wang et al., 2018), we put instances into template (e.g., A computer system failure ... Based on that information, is the claim: The Tokyo Stock ... "True" or "False"? Answer: True.); Secondly, we take k instances with selection methods like KATE (Liu et al., 2022b); Finally, we combined the demonstrations with the templated query (Yet, we now are ... Based on that information, is the claim: Bacteria is winning the war against antibiotics."True" or "False"? Answer:) and feed them into our LLMs. By mimicking the template of demonstrations, LLMs can directly infer the relationship between two sentences (True) instead of explanations or meaningless answers (The claim is "false"), which also helps reduce the likelihood of hallucination (Huang et al., 2024).

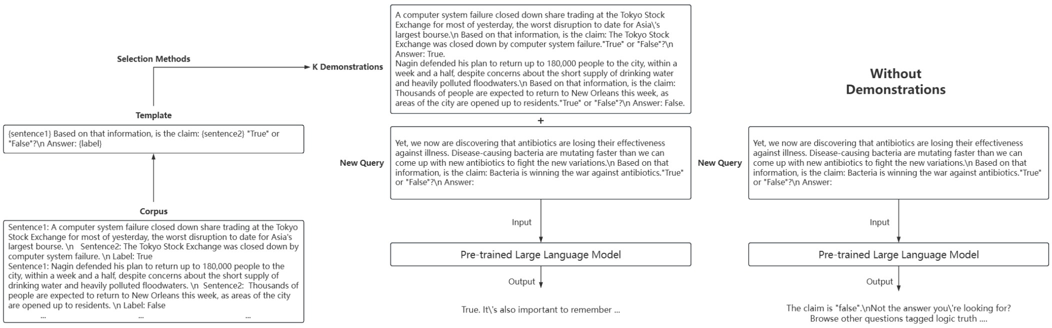

Figure 1: Illustration of in-context learning

Formally, given a pre-trained large language model f_M, a set of candidate answers $Y = \{y_1, ..., y_m\}$ and k demonstration examples: $C = \{s(x_1, y_1), s(x_k, y_k)\}$, where x_i and y_i are the ground truth text and label in the demonstration set, and s is the template according to the task (Dong et al., 2024). For example:

***s(x)**: No Weapons of Mass Destruction Found in Iraq Yet. Based on that information, is the claim Weapons of Mass Destruction Found in Iraq. "True" or "False"? **s(y)**: Answer: False*

Then the final predicted label y for instance x will be the candidate answer with the highest probability:

$$y = \arg\max_{y_j \in Y} f_M(y_j, C, x) \tag{1}$$

Because in-context learning does not require changing the parameters of LLMs, this paradigm has made it possible for LLMs to quickly adapt to new tasks and has been proven successful in many practical applications (Gao et al., 2023; Xi et al., 2023; Liu et al., 2022a).

How to select demonstrations for in-context learning (ICL)? The performance of ICL is affected by many factors, such as the number of demonstrations (Liu et al., 2022b), the order of demonstrations (Lu et al., 2022; Liu et al., 2024), label imbalances (Zhao et al., 2021; Chen et al., 2023a), and the templates of demonstrations (Sorensen et al., 2022). Among these, Work found that the selection of demonstrations has the greatest impact on ICL performance, outweighing the influence of labels and templates. Additionally, selecting a few high-quality demonstrations as a prompt for LLMs is more efficient, as it reduces both inference time and the costs associated with long contexts (Chen et al., 2023b).

Most current demonstration selection methods operate at the instance level, where a unique set of demonstrations is selected for each query. While corpus-level selection strategies are more efficient, as they eliminate the need to make individual selection for each query, identifying a single set of demonstrations that performs well across all queries is still a huge challenge and relatively unexplored (Dong et al., 2024).

This thesis aims to address this challenge by selecting a subset of instances that consistently improve the model's ICL capability on unseen data. Identifying these traits, the model could perform ICL robustly across all queries, even with randomly sampled demonstrations from this subset. Furthermore, for instance-level selection, this approach has the potential to narrow the search space, making the process more efficient.

Data Maps This paper explores the use of Data Maps (Swayamdipta et al., 2020). Data Maps are constructed by first fine-tuning models on the training dataset and then categorizing the data into three regions based on training dynamics (Swayamdipta et al., 2020): easy-to-learn, ambiguous, and hard-to-learn. These regions are used to analyze which one facilitates the model's generalization ability - the capacity of the model to perform well on unseen data beyond the training set. However, Swayamdipta et al. (2020) focused on the fine-tuning stage, and this paper investigates the usability of Data Maps for in-context learning. A detailed explanation of why Data Maps hold potential for use in ICL is provided in Appendix A.1.

This thesis aims to explore whether a data map can serve as a tool for data selection in in-context learning (ICL). To address this, I investigated two sub-questions: **SQ1: How to quantify the contribution of each instance towards the ICL performance?** As training dynamics could be seen as two features of datapoints, it does make sense to explore the potential relationship between training dynamics and the performance of ICL. However, as the demonstrations in one set will interact with each other (Chen et al., 2023b) and the accuracy of one-shot learning (ICL with only one demonstration) ignore the impact from other demonstrations, the accuracy of one-shot learning is not sufficient to evaluate the influence of one demonstration. Therefore, I introduced DataModels (Nguyen and Wong, 2023) to calculate the influence scores of each instance on the ICL.

Secondly, **SQ2: How do the three regions in Data Map affect in-context learning?** Understanding the impact of different regions within a Data Map on ICL is also useful for optimizing demonstration selection. By analyzing each region's contribution to ICL, I can identify which types of data offer the most reliable support. I will sample a large number of demonstration sets from each region and then calculate the average accuracy to represent each region's influence on ICL.

2 Related Work

There are two main kinds of methods to select demonstrations: unsupervised methods (Liu et al., 2022b; Sorensen et al., 2022; Nguyen and Wong, 2023; Gonen et al., 2023; Sun et al., 2024) and supervised methods (Rubin et al., 2022; Li et al., 2023; Ye et al., 2023; Wang et al., 2023). Unsupervised methods do not need to update the parameters of LLMs, they leverage various metrics (e.g., distance to the query, perplexity, diversity) of the text to select the demonstrations. In contrast, supervised methods generally fine-tune two retrievers (also called encoders — one for encoding candidate demonstrations from the training dataset, and one for encoding query texts) through a customised loss function. This loss function often includes specific metrics designed to guide the retrievers toward focusing on particular features of the demonstrations. After the retrievers are fine-tuned, demonstrations are then selected by the unique features used in the above loss function. For example, as shown in Figure 2, it first used an unsupervised method (BM25 (Robertson et al., 2009)) to retrieve a set of candidate training instances, then uses a scoring model to categorize the data into positive and negative. Finally, it trained the query encoder and the demonstrations encoder through contrastive learning (Oord et al., 2018) to learn to retrieve demonstrations that are most similar to the query. Since this paper focuses on unsupervised approaches, this section will only review related work in unsupervised methods.

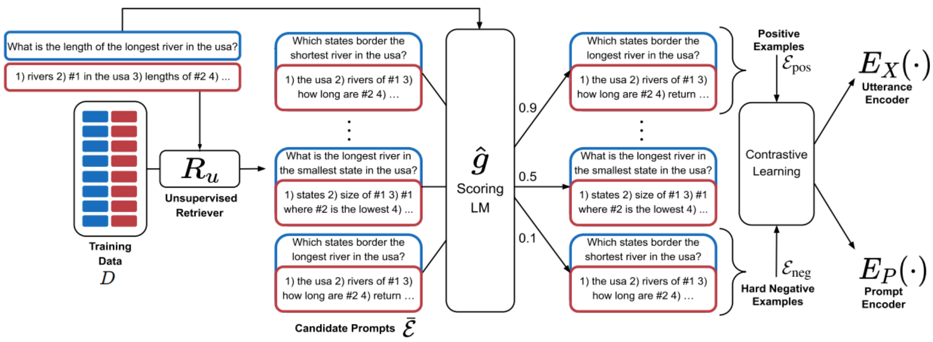

Figure 2: An overview of one supervised method, cited from Rubin et al. (2022)

Liu et al. (2022b) found that random selection of demonstration sets often leads to significant variability in ICL accuracy, as different demonstration sets can lead to large differences in ICL performance. They also proposed an unsupervised method called KATE (**K**nn-**A**ugmented in-con**T**ext **E**xample selection), which is based on the distance (Euclidean Distance or Cosine Similarity) between query and training instances to select demonstrations. Based on their result, Sun et al. (2024) proposed a data compression-based approach to selecting demonstrations with the aim of efficiently selecting examples relevant to the query input while retaining sufficient information from the training dataset. Firstly, they applied BM25 (Robertson et al., 2009) to retrieve a set of relevant demonstrations. Next, they used the Influence Function (Yang et al.,

2022) to estimate the parameter change caused by re-weighting an example for the training dataset. Finally, they re-ranked demonstrations by combining relevance and impact scores.

Different from KATE, which assumes that the more similar the demonstration is to the test query, the better in-context learning performs, it's also crucial to consider the diversity of demonstrations set. Levy et al. (2022) thought that when models are tested on queries that are out of distribution from the training set, selecting similar demonstrations is insufficient; instead, diverse demonstrations can enhance generalization. They applied Determinantal Point Process (DPP) (Kulesza et al., 2012) to choose sets that contain relevant and diverse demonstrations and performed experiments on semantic parsing datasets. Their results proved that DPP method outperformed Top-k (select k most similar demonstrations based on the distance metric (Liu et al., 2022b)).

In addition to similarity and diversity, Gonen et al. (2023) assumed that demonstrations with lowest perplexity conform to the syntax of the used LLMs, i.e., the better it is understood by that LLM leading to the better performance. They performed experiments on 2 tasks: word-level translation and classification tasks and looked at two measures: (a) the confidence of the correct label given by LLMs, averaged across 1,000 test instances; (b) the accuracy on the task, computed over the 1,000 test instances. They found that the prompts with the lowest perplexity often performed the better than than the prompts with higher perplexity.

Unlike the methods above, which make various assumptions about the properties of demonstrations and their contribution to ICL, Nguyen and Wong (2023) selected demonstrations based on the changes they caused to accuracy on the validation set. First, they randomly selected N (a hyperparameter) subsets S_i of instances from the training dataset as demonstration sets and recorded the accuracy y_i for each demonstration set S_i. Secondly, they fit a Linear Lasso Model (Tibshirani, 1996) g_θ on the dataset D of $\{(S_i, y_i)\}$ pairs:

$$g_\theta(s_i) = \theta \cdot 1_{s_i}^T + \theta_0 \tag{2}$$

where s_i is i-th demonstrations set and 1_{s_i} is a binary indicator vector of length equal to the size of the original training set. A value of 1 at position j indicates that the instance j in training set is included in the demonstrations and a value of 0 is the opposite. By training this linear model on D, the parameter θ_j could be seen as influence estimates on ICL (the j-th value in the vector is the influence score of the j-th instance).

They performed experiments in classification tasks and multi-choice tasks and compared their methods with random choosing, best set (demonstrations with the highest accuracy in validation set), one-shot (also one demonstration with the highest accuracy in validation set), KATE (Liu et al., 2022b) and perplexity (Gonen et al., 2023). For positive demonstrations (influence score is positive), their methods outperform other unsupervised methods (KATE (Liu et al., 2022b) and

Perplexity (Gonen et al., 2023)). Furthermore, in most cases, selecting instances with negative influence scores as demonstrations for ICL tends to result in lower performance compared to using instances with either positive or neutral influence scores. This indicates influence-based selection methods can consistently identify helpful/harmful demonstrations. In addition, they also found different LLMs (GPT-NeoX, LLaMA, OPT) will not share the high-influence demonstrations and the accuracy of LLMs tends to increase as the number of demonstrations increases. Finally, they studied the relationship between the position of demonstrations (the order) and the performance of ICL. They computed the influences of each position in 4-shot ICL and the results showed that influence scores of demonstrations increased as their position moved closer to the query in the order.

All of the above methods examined the sensitivity of their approaches to the number of demonstrations used. Chen et al. (2023b) was the first to investigate the universal impact of demonstration quantity on reasoning tasks and found: a single positive demonstration (for each test query, the demonstration leading to the correct answer in one-shot ICL is positive) outperforms using eight positive demonstrations. They suggested that multiple demonstrations may introduce redundant information, which can confuse the model. Their experiments also revealed that adding more positive demonstrations reduces performance, while adding negative demonstrations (with stricter rules) improves it. This indicates interactions among demonstrations in in-context learning and suggests that carefully selected negative demonstrations may enhance ICL performance more effectively than positive ones.

3 Methodology

This section describes two main concepts of my methodology. In the first section, I will introduce the concept of Data Maps. In Section 3.2, I will focus on DataModels and explain how they can be leveraged to analyze demonstrations in in-context learning.

3.1 Data Maps

Swayamdipta et al. (2020) proposed a method called Data Maps to categorize and diagnose datasets, which is based on the idea that each instance in the training dataset contributes differently to the model. They fine-tuned a large language model, RoBERTa (Liu et al., 2019), over multiple epochs on the training set and defined two **Training Dynamics**: **Confidence** and **Variability**, where confidence refers to the mean model probability of the true label (y_i^*) for the i-th instance across E epochs:

$$\tilde{\mu_i} = \frac{1}{E}\sum_{e=1}^{E} p_{\theta^{(e)}}(y_i^*|x_i) \tag{3}$$

where $p_{\theta^{(e)}}$ denotes the model's probability with parameters $\theta^{(e)}$ at the end of the e-th epoch. Variability measures the standard deviation of the i-th instance's $p_{\theta^{(e)}}(y_i^*|x_i)$ across epochs:

$$\tilde{\sigma_i} = \sqrt{\frac{\sum_{e=1}^{E}(p_{\theta^{(e)}}(y_i^*|x_i) - \tilde{\mu_i})^2}{E}} \tag{4}$$

As shown in Figure 3, the authors concluded that there are three main regions in the training dataset: easy-to-learn, ambiguous and hard-to-learn. Specifically, easy-to-learn instances are those consistently predicted with high confidence across epochs, while hard-to-learn instances show consistently low confidence, and ambiguous datapoints are instances with high variability. Their experimental results show the datapoints in the ambiguous region make the greatest contribution to improving out-of-distribution generalization. Secondly, easy-to-learn region plays an important role in model optimization. While Swayamdipta et al. (2020) focused only on the fine-tuning setting, the potential value of Data Maps for in-context learning remains understudied. Since ICL can be explained as a form of implicit fine-tuning (Dai et al., 2023) (see Appendix A.1 for details), I hypothesize that using instances from the ambiguous region as demonstrations could also enhance ICL performance on unseen instances.

Most work using DataMaps have focused on the fine-tuning setting (e.g., Ethayarajh et al. (2022); Karamcheti et al. (2021)). One exception is that Liu et al. (2022a) used it to generate a new dataset based on MultiNLI. They selected the demonstrations from the most ambiguous (25%) region of the dataset and leveraged GPT-3's in-context learning capabilities to generate

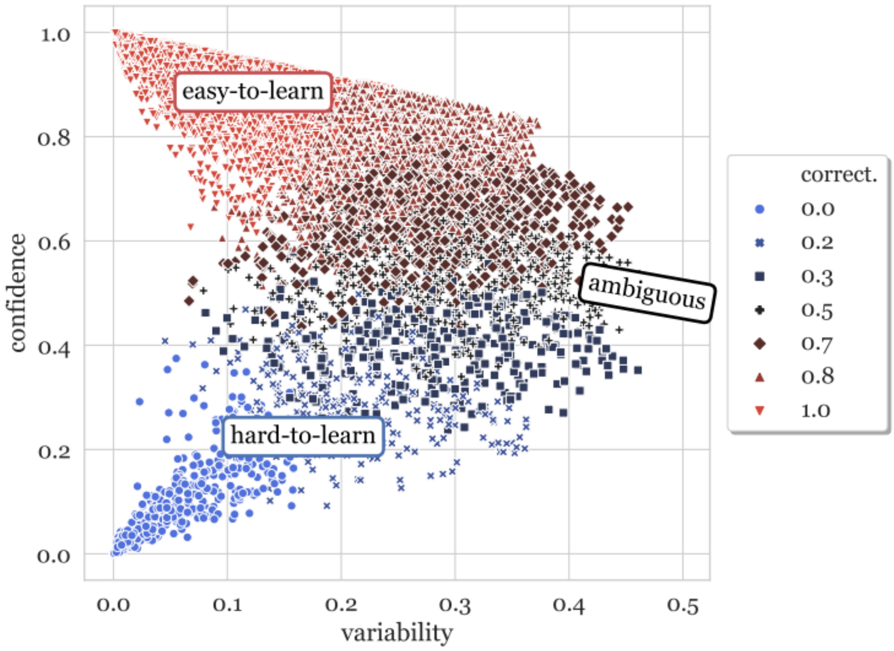

Figure 3: Data Map for SNLI training dataset, taken from (Swayamdipta et al., 2020)

new instances that have a high probability of also belonging to the ambiguous region. In this way, compared with the datasets from the same task, their new datasets have more instances in the ambiguous region. Their experimental results proved that the models trained on their new dataset perform better on test datasets than the same models trained on other datasets from the same task, including the original dataset. However, their aim is not to enhance ICL directly, but rather to utilize ICL for the generation of datasets containing more ambiguous instances, intended for further fine-tuning.

3.2 DataModels

Nguyen and Wong (2023) proposed an influence-based demonstrations selection that invokes DataModels (Ilyas et al., 2022) framework. Firstly, they randomly selected a subset of instances from the training dataset to create a demonstration set s_i, repeating this process N times. For each demonstration set s_i, they recorded its corresponding accuracy y_i on the validation set. Secondly, consider the combined dataset $D = \{S, Y\}$, where $S = \{s_1, ..., s_N\}$ and $Y = \{y_1, ..., y_N\}$ such that s_i represents a demonstrations set and y_i represents the validation accuracy corresponding to subset s_i. To predict validation accuracy, they fit a Linear Lasso Model (Tibshirani,

1996) g_θ on the dataset D of $\{(s_i, y_i)\}$ pairs:

$$g_\theta(s_i) = \theta \cdot 1_{s_i}^T + \theta_0 \tag{5}$$

where s_i is i-th demonstrations set, $g_\theta(s_i)$ predicts the validation accuracy given s_i, and 1_{s_i} is a binary indicator vector of length equal to the size of the original training set. A value of 1 at position j indicates that the instance j on the training set is included in the demonstrations and a value of 0 is the opposite.

Lasso model is chosen for its ability to produce sparse models by shrinking less important feature weights to zero, which allows the identification of the most influential training instances. By training the model on D, the parameter θ_j represents the influence score of the j-th training instance, while the bias term θ_0 indicates the baseline accuracy of the model's predictions. The output $g_\theta(s_i)$ provides a predicted accuracy for the given demonstration set.

In Data Maps (Swayamdipta et al., 2020), each of the three regions had a different effect on the performance of the fine-tuned model, so I created separate DataModel for each region in this work. Since a Data Map divides the dataset into three regions, comparing the influence scores of instances from each region directly would not be meaningful due to different bias of each DataModels. To address this, I introduce a metric called the Influence-Bias Score. This score combines the bias of the DataModel with the influence score of the i-th demonstration, providing a measure of how the demonstration adjusts the baseline accuracy after receiving the effect from the sampling region. The Influence-Bias Score is computed as the sum of the bias (θ_0) and the influence score for a given demonstration:

$$IBS_i = \theta_0 + \theta_i \tag{6}$$

4 Experiments

This section describes the experimental steps of my methodology. In Section 4.1, I will present the experimental configuration: the large language model and the NLI dataset I choose. In Section 4.2 and 4.3, I will describe my main research findings.

4.1 Setup

Models. For the Data Map, I will follow the original paper by Swayamdipta et al. (2020), using the RoBERTa models (Liu et al., 2019) to build the data map. Compared with larger LLMs such as LLaMa-2 (Touvron et al., 2023), RoBERTa models are more lightweight. Furthermore, Du et al. (2023) found that the overall structure of data maps remains consistent across different models. Therefore, I will use the RoBERTa-base model with 125M parameters to build data maps.

For ICL, since Brown et al. (2020) first proposed the concept of in-context learning with GPT, most studies have used the GPT family for experiments. Nguyen and Wong (2023) used a different large language model family, LLaMa-2 (Touvron et al., 2023), and demonstrated its strong ICL abilities. In this thesis, I will perform ICL experiments using the LLaMa-2-13b model from Hugging Face. All experiments were conducted using an NVIDIA A100 80 GB GPU.

Dataset. I use one **Natural Language Inference** dataset: RTE (Recognizing Textual Entailment) dataset (Wang et al., 2018), which is designed to determine the semantic/entailment relationship between two sentences. RTE has been extensively studied in previous work on Data Maps (Swayamdipta et al., 2020) and will be the primary focus of my experiments. The goal of RTE is to predict whether the first sentence entails the second sentence and classify each pair into two categories: entailment and not entailment. In previous studies, this dataset has been used to analyze how to select optimal demonstrations for ICL (Lu et al., 2022; Nguyen and Wong, 2023; Li et al., 2023).

Previous studies generally used the training set for demonstration selection and the development and test sets for prediction in their experiments, e.g., (Nguyen and Wong, 2023). Following previous studies, I used the entire training set for demonstration selection and performed in-context learning (ICL) on the complete development set, evaluating performance with accuracy as the metric. The data split statistics are shown in Table 1:

Dataset	Train	Dev
RTE	2.49k	277

Table 1: Data split for the datasets

Examples from the RTE dataset and the prompt template are shown in Tables 2 and 3.

Premise	Hypothesis	Label
No Weapons of Mass Destruction Found in Iraq Yet.	Weapons of Mass Destruction Found in Iraq.	not_entailment
Edward VIII became King in January of 1936 and abdicated in December.	King Edward VIII abdicated in December 1936.	entailment

Table 2: Examples from RTE, cited from Wang et al. (2018)

Dataset	Template
RTE	"text1" Based on that information, is the claim "text2" "True" or "False"? Answer:

Table 3: Templates of RTE dataset

In my experiments, I also used *Stanford Natural Language Inference (SNLI)* (Bowman et al., 2015) and *Multi-Genre Natural Language Inference (MultiNLI)* (Williams et al., 2018) datasets. The SNLI dataset comprises 570k sentence pairs labeled as entailment, contradiction, or neutral, while the MultiNLI dataset contains 433k sentence pairs spanning ten distinct genres. Although SNLI and MultiNLI are well-established benchmarks for natural language inference, the performance of Llama2-13b on these datasets was unexpectedly low, nearly approaching random guessing. Given that these are three-class classification tasks, the model's accuracy hovered at approximately 34%, slightly above the random guessing baseline of 33.3%. In contrast, the model performed more reliably on the RTE dataset, likely due to its simpler two-class classification structure. This suggests that task complexity and class structure significantly impact model performance. Consequently, my thesis focuses on the RTE dataset.

Configurations. Firstly, for each shot, I randomly sampled demonstration sets from each region and calculated the accuracy of each set on the validation set separately, configured as shown in Table 4. I also sampled a number of demonstration sets from the entire training dataset and compared it to different regions to explore which region retained more demonstrations that positively impact ICL. Secondly, to compute influence scores of instances in different regions under different shots, I constructed Lasso linear models using binary vectors representing demonstration indices (with each dimension indicating whether a demonstration was included (1) or not (0)) and the corresponding accuracy. The weight of each demonstration in DataModel can then be seen as the influence score of this demonstration for ICL in the region.

shot	dataset	the number of demonstration sets
3	Entire	5200
	Easy-to-learn	1800
	Ambiguous	1800
	Hard-to-learn	1800
5	Entire	2000
	Easy-to-learn	700
	Ambiguous	700
	Hard-to-learn	700
8	Entire	2000
	Easy-to-learn	700
	Ambiguous	700
	Hard-to-learn	700
12	Entire	1300
	Easy-to-learn	450
	Ambiguous	450
	Hard-to-learn	450
16	Entire	1000
	Easy-to-learn	350
	Ambiguous	350
	Hard-to-learn	350
20	Entire	800
	Easy-to-learn	300
	Ambiguous	300
	Hard-to-learn	300

Table 4: The number of demonstration sets for different regions under different shots

For the calculation of influence scores, the official code by Nguyen and Wong (2023) sampled 400 data points from the original training set. However, since I need to calculate influence scores for all 2,490 instances, I expanded the number of demonstration sets. This expansion increased the number of indices-accuracy pairs, thereby providing a better fit for the lasso model.

In the inference process, to accelerate in-context learning, I utilized the **vLLM** (Kwon et al., 2023) library. For each set of experiments, I configured the following parameters: *temperature* = 1, *top_k* = 10, *repetition_penalty* = 1, *max_tokens* = 5 and *random_seed* = 42. See Appendix A.2 for specific configuration instructions.

4.2 Main Results

Data Maps. Figure 4 demonstrates the Data Map for the training set of the RTE dataset constructed by RoBERTa-base. For RTE dataset, datapoints with high confidence (i.e., the mean probability of the true label of a data point across epochs) constitute the majority of the data points. The variability values (i.e., the standard deviation of the true label of a data point across epochs) are concentrated in the middle range, around 0.2. I have also listed a few examples from different regions in Table 5. The first example is categorized as easy-to-learn because it involves different subjects (compensation vs. shares) and different quantifiers (549 million vs. 30 million). The ambiguous example can be interpreted in two ways: one interpretation is that if the discussion is about when humans left Africa, the hypothesis is unrelated to the premise (as they cover different subjects); the other interpretation is that the premise, which states humans left Africa one million years ago, implies humans existed 10,000 years ago, since one million is greater than 10,000. The hard-to-learn example is more complex: although the premise contains all the words of the hypothesis, there is no causal relationship between them.

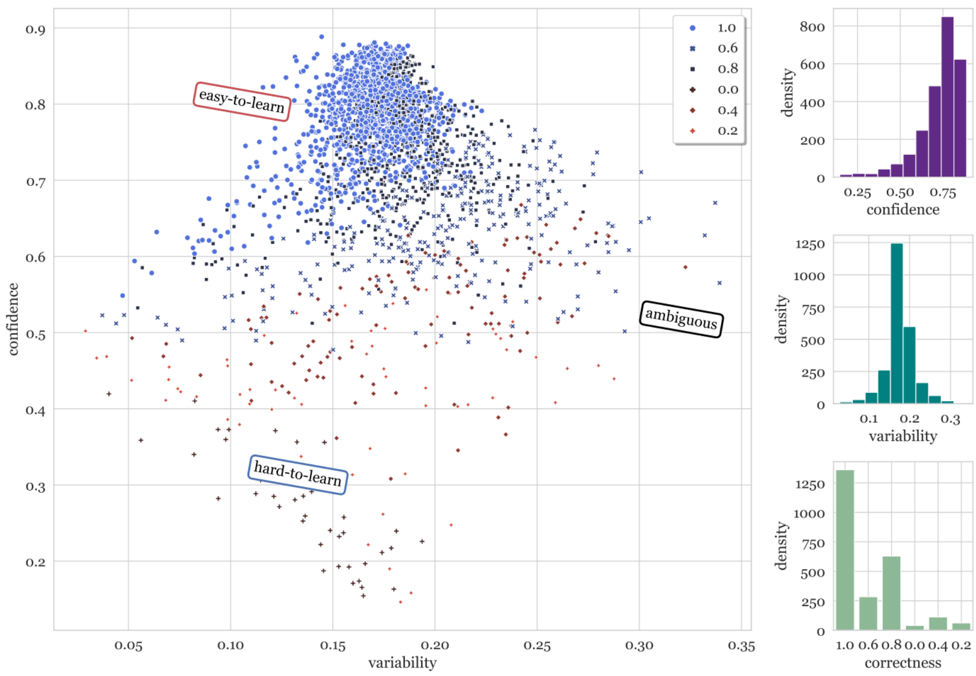

Figure 4: Data Map for RTE training dataset built by roberta-base

Premise	Hypothesis	Label	Region
With $549 million in cash as of June 30, Google can easily afford to make amends.	Some 30 million shares have been assigned to the company's workers.	not_entailment	easy-to-learn
About one million years ago, these people began to slowly leave Africa.	Humans existed 10,000 years ago.	entailment	ambiguous
The Chicago White Sox are a major league baseball team based in Chicago, Illinois.	The Bulls basketball team is based in Chicago, Illinois.	not_entailment	hard-to-learn

Table 5: Example from each region in the RTE dataset

I experimented with various numbers of demonstrations and observed consistent results across different shots. I therefore present the experimental results for 12 shots in the main text and the experimental results for 3-shot, 5-shot, 8-shot, 16-shot and 20-shot can be found in the

Appendix A.3. Following the configuration of Table 4, I sampled 1300 demonstration sets from the entire training dataset and 450 demonstration sets from each of three regions for 12-shot. For each region, I calculated the accuracy corresponding to each demonstration sets to construct the DataModels.

Reliability of DataModels. In order to explore the potential relationship between training dynamics and the performance of ICL, I performed DataModels to evaluate the impact of the specific demonstration. I trained DataModels with the demonstration set and corresponding true accuracy pairs obtained in the previous step. The fitted DataModels closely approximate the true accuracy: there is a strong linear correlation between true accuracy and predicted accuracy by DataModels of the entire training dataset (Figure 5). I also calculated the **Pearson Correlation Coefficient** (Mukaka, 2012), which has a value greater than 0.759. Overall, when the influence score of a demonstration is greater than zero (increases the predicted accuracy), it will positively impact the in-context learning performance. Similarly, I also created three DataModels for each of the three regions of Data Maps, and the experimental results show that there is a strong linear relationship between prediction accuracy and true accuracy, as shown in Table 6.

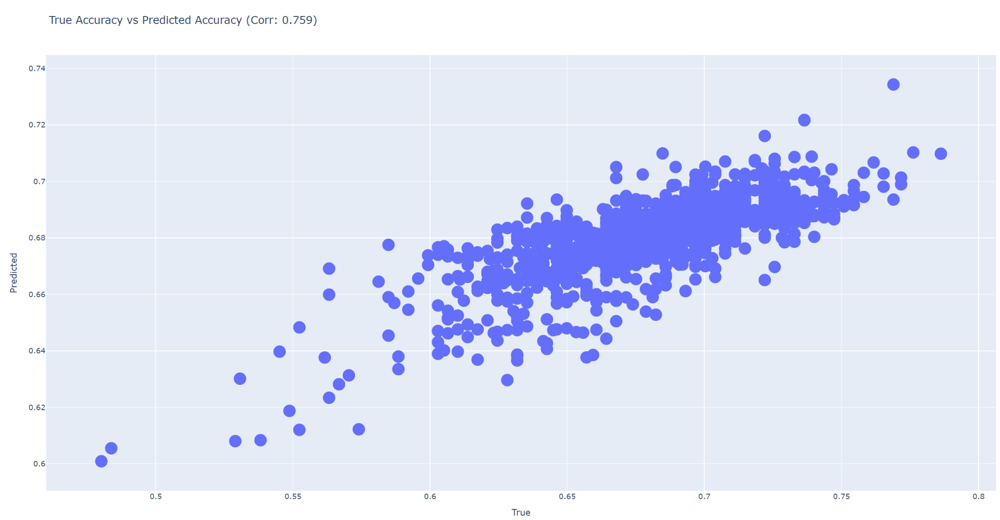

Figure 5: The predicted versus true accuracy of the linear in-context DataModel for the entire training dataset under 12 shots. Pearson Correlation Coefficient describes the strength and direction of the linear relationship between true accuracy and predicted accuracy

Methods	Pearson Correlation Coefficient
the entire dataset	0.759
easy-to-learn	0.930
ambiguous	0.904
hard-to-learn	0.956

Table 6: The Pearson Correlation Coefficient of four methods under 12 shots

Training Dynamics and Influence Scores. In order to explore the direct relationship between training dynamics and in-context learning, I examined the scatter plot between confidence/variability and influence scores, as shown in Figure 6 and Figure 7. Most results are mixed and there is no explicit relationship between training dynamics and influence score for the entire training dataset. However, as shown in Figure 6, the influence scores for samples with low confidence scores are usually 0. This indicates that, when sampling from the entire dataset as demonstrations, hard-to-learn instances usually have no effect on ICL.

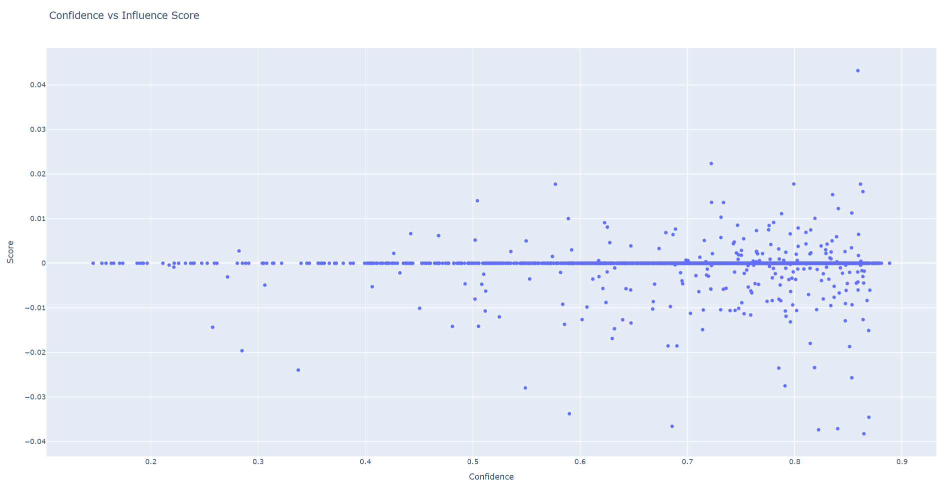

Figure 6: Scatter Plot of Confidence vs. Influence Score for the entire training dataset under 12 shots

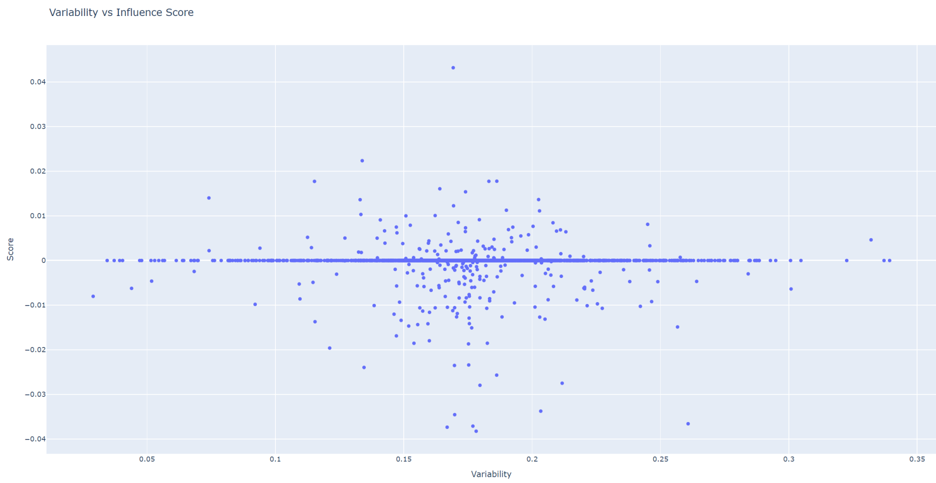

Figure 7: Scatter Plot of Variability vs. Influence Score for the entire training dataset under 12 shots

Likewise, there is no explicit relationship between training dynamics and influence score for three separate regions (Figure 12, Figure13 and Figure 14 in Appendix A.3). Overall, the connection between these factors is more complex and may be influenced by other variables, such as similarity, which warrants further exploration in future work.

Accuracy of Different Regions. In order to measure the impact of different regions on ICL, I computed the average accuracy and variance of the demonstration sets from different regions, as configured in Table 4 of Section 4.1. Table 7 shows the average accuracy and variance for the entire dataset as well as for the three regions. The analysis reveals that ambiguous demonstration sets achieve a similar average accuracy (0.681) to the entire training dataset but with lower variance, indicating more consistent performance. The average accuracy of easy-to-learn demonstration sets was slightly lower, coming in second. In contrast, hard-to-learn demonstration sets had the lowest average accuracy and the highest variance. This indicates that most hard-to-learn instances perform worse than ambiguous or easy-to-learn instances.

Methods	Average Accuracy	Variance
the entire dataset	0.681	0.001348
easy-to-learn	0.676	0.001316
ambiguous	0.681	0.001088
hard-to-learn	0.668	0.001540

Table 7: The accuracy (mean and variance) of four methods under 12 shots

Impact of Sampling Regions on ICL Performance. In order to evaluate the impact of one demonstration in different regions (e.g., when it is combined with only ambiguous instances and when it is combined with instances from the entire training dataset), I used influence-bias scores (IBS) to represent how sampling the same instance from different sets impacts ICL.

Figure 8 shows the distinct effects of different sampling regions on the IBS, where **Easy (whole)** denotes the IBS of the easy instances when other demonstrations are sampled from the whole dataset, whereas **Easy (region)** refers to the performance of the ICL when other demonstrations are only sampled from the easy-to-learn region. Firstly, when only **easy-to-learn** instances were used as demonstrations, the upper fence ($Q_3 + 1.5 \times IQR$, where Q_3 is the third quartile, marking the 75-th percentile of the data, and IQR is the interquartile range, calculated as $Q_3 - Q_1$, representing the range between the first and third quartiles) of the IBS of the instances didn't change but the mean and lower fence ($Q_1 - 1.5 \times IQR$) of IBS became lower compared with Easy (whole). This indicates that too many easy-to-learn demonstrations can actually be detrimental to ICL. Secondly, for the **hard-to-learn** instances, the overall IBS significantly decreases when demonstrations were only sampled from the hard-to-learn regions. Finally, the mean IBS of **ambiguous** instances improves when ambiguous demonstration was only combined with ambiguous demonstration. The upper fence of IBS increased while the lower fence also decreased. Ambiguous demonstrations usually have a more positive impact on ICL than demonstrations from other regions.

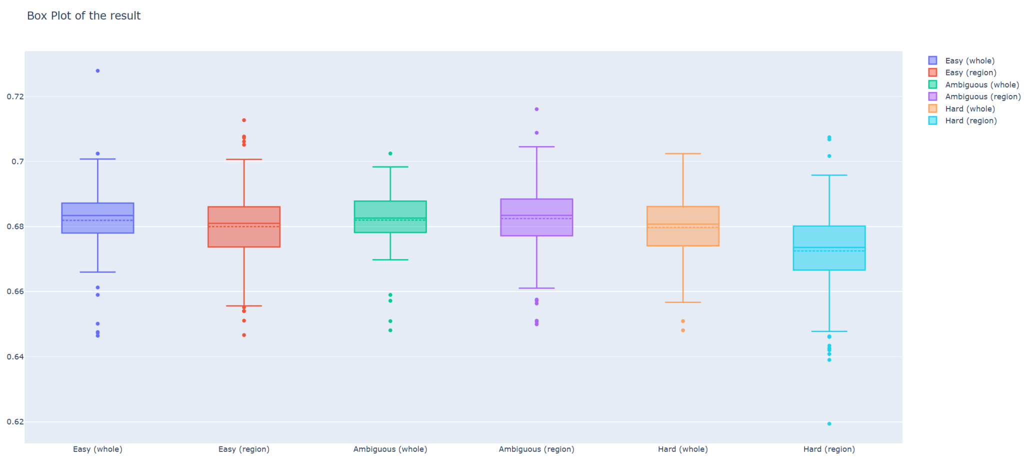

Figure 8: The Influence-Bias Scores of three regions under 12 shots; **Easy (whole)** represents the IBS of easy instances when other demonstrations are sampled from the entire dataset, while **Easy (region)** refers to the performance when other demonstrations are only sampled from the easy-to-learn region for ICL; the same applies to Ambiguous and Hard.

4.3 Overall Performance for all shots

Figure 9: The mean (left) and variance (middle) accuracy of four methods under different shots, along with average accuracy of ambiguous region minus average accuracy of the entire dataset (right)

Figure 10: The average influence-bias scores of three regions under different shots; **Easy (whole)** represents the IBS of easy instances when other demonstrations are sampled from the entire dataset, while **Easy (region)** refers to the performance when other demonstrations are only sampled from the easy-to-learn region for ICL; the same applies to Ambiguous and Hard.

Average Accuracy and Variance across different shots. For all shots, selecting demonstrations exclusively from the ambiguous region results in both improved performance and greater consistency in in-context learning (ICL). In Table 7, the average accuracy of the ambiguous region for 12 shots is slightly lower than that of the entire dataset. However, as shown in Figure 9, for other shots, ambiguous demonstration sets consistently achieve the highest average accuracy (slightly exceeding that of the entire dataset) and the lowest variance, while the hard-to-learn region shows the opposite trend. In terms of both mean and variance, the easy-to-learn region and the entire training dataset show no significant difference, with the average accuracy of the entire dataset being only about 0.2% higher.

Average Influence-Bias Score across different shots. The result shows that only ambiguous (region) achieves a higher average influence-bias score compared to ambiguous (whole) for in-context learning in all shots. Figure 10 illustrates the average influence-bias scores (IBS) of demonstrations sampled either from specific regions or the entire dataset. For the other two regions, the average IBS of easy (region) is exceeded by that of easy (whole) at 12 shots and the average IBS of hard (region) is exceeded by that of hard (whole) when shot is 8. This indicates that as the number of demonstrations increases, sampling exclusively from the easy or hard regions becomes less effective for ICL compared to incorporating demonstrations from other regions.

5 Conclusions

Since different demonstrations are selected for each query, most current unsupervised selection methods for ICL are computationally expensive. Inspired by Data Maps, this thesis introduces a method to classify and diagnose the training dataset using a smaller model (RoBERTa). This approach effectively reduces the search space for demonstrations while maintaining their quality. Furthermore, I demonstrate that instances in the ambiguous region can positively influence ICL performance.

Importantly, my findings reveal that there is no straightforward linear relationship between individual training dynamics and in-context learning. However, the results suggest that the ambiguous region provides more effective and robust demonstrations for ICL compared to other regions. By sampling solely from the ambiguous region, which contains the top 33% of instances with the highest variability, Data Maps can reduce the search space for unsupervised methods to just 33% of the original dataset.

In addition, demonstrations from the easy-to-learn region perform slightly underperform compared to those sampled from the entire training dataset, while the hard-to-learn region proves to be the least effective and most unstable. Notably, across all shots, demonstrations solely from the ambiguous region outperform those combined with the other two regions. For the easy-to-learn and hard-to-learn regions, combining demonstrations within the same region only yields positive effects when the number of demonstrations (shots) is small.

6 Future Work

In this section, I'd like to discuss the limitations of the current work and outline several potential avenues for future research. Firstly, Swayamdipta et al. (2020) used 33% as threshold for each region: for each region, they selected the top 33% of data with the largest or smallest confidence/variability. For RTE dataset, the number of instances with low confidence is relatively low and the instances in the hard-to-learn region exhibit significant variation in confidence score. Similarly, the ambiguous region also contains some instances with low variability. Therefore, dynamically adjusting the threshold for different datasets may lead to improved ICL performance.

Secondly, this paper applied Data Maps in the RTE dataset with the LLaMa-2-13b model. It remains to be seen whether models with different structures or different parameter sizes can achieve similar results on the RTE dataset using Data Maps. Moreover, since this paper only applied Data Maps to the RTE dataset for ICL, future work could evaluate the performance of Data Maps across various Natural Language Inference datasets. Given that Swayamdipta et al. (2020) applied Data Maps specifically to the NLI task, which is a classification task, extending its use to ICL in generation tasks presents an additional challenge. Finally, even within the same region of the RTE dataset, the performance of different instances varies considerably. Therefore, it would be valuable to explore how to select the most effective set of demonstrations from one region to maximize the performance of ICL.

Bibliography

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.746. URL `https://aclanthology.org/2020.emnlp-main.746`.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.64. URL `https://aclanthology.org/2024.emnlp-main.64`.

Berivan Isik, Natalia Ponomareva, Hussein Hazimeh, Dimitris Paparas, Sergei Vassilvitskii, and Sanmi Koyejo. Scaling laws for downstream task performance of large language models. *arXiv preprint arXiv:2402.04177*, 2024.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

Jiaxing Liu, Chaofeng Sha, and Xin Peng. Improving fine-tuning pre-trained models on small source code datasets via variational information bottleneck. In *2023 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 331–342. IEEE, 2023.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates, December 2022a. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.508.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi, editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In Eneko Agirre, Marianna Apidianaki, and Ivan Vulić, editors, *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, May 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.deelio-1.10. URL https://aclanthology.org/2022.deelio-1.10.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, November 2024. ISSN 1046-8188. doi: 10.1145/3703155. URL https://doi.org/10.1145/3703155. Just Accepted.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR, 2021.

Yanda Chen, Chen Zhao, Zhou Yu, Kathleen McKeown, and He He. On the relation between sensitivity and accuracy in in-context learning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 155–167, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.12. URL `https://aclanthology.org/2023.findings-emnlp.12`.

Taylor Sorensen, Joshua Robinson, Christopher Rytting, Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. An information-theoretic approach to prompt engineering without ground truth labels. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 819–862, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.60. URL `https://aclanthology.org/2022.acl-long.60`.

What Makes In-Context Learning Work. Rethinking the role of demonstrations: What makes in-context learning work?

Jiuhai Chen, Lichang Chen, Chen Zhu, and Tianyi Zhou. How many demonstrations do you need for in-context learning? In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11149–11159, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.745. URL `https://aclanthology.org/2023.findings-emnlp.745`.

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

Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. Demystifying prompts in language models via perplexity estimation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10136–10148, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.679. URL `https://aclanthology.org/2023.findings-emnlp.679`.

ZhongXiang Sun, Kepu Zhang, Haoyu Wang, Xiao Zhang, and Jun Xu. Effective in-context example selection through data compression. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 871–877, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.50. URL `https://aclanthology.org/2024.findings-acl.50`.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *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, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.191. URL `https://aclanthology.org/2022.naacl-main.191`.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. Unified demonstration retriever for in-context learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4644–4668, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.256. URL `https://aclanthology.org/2023.acl-long.256`.

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

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Shuo Yang, Zeke Xie, Hanyu Peng, Min Xu, Mingming Sun, and Ping Li. Dataset pruning: Reducing training data by examining generalization influence. *arXiv preprint arXiv:2205.09329*, 2022.

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

Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4005–4019, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.247. URL `https://aclanthology.org/2023.findings-acl.247`.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with

\

mathcal {V} $-usable information. In$ *International Conference on Machine Learning*, $pages 5988--6008. PMLR$, 2022.

Siddharth Karamcheti, Ranjay Krishna, Li Fei-Fei, and Christopher Manning. Mind your outliers! investigating the negative impact of outliers on active learning for visual question answering. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.564. URL `https://aclanthology.org/2021.acl-long.564`.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. *arXiv preprint arXiv:2202.00622*, 2022.

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

Yupei Du, Albert Gatt, and Dong Nguyen. Ftft: efficient and robust fine-tuning by transferring training dynamics. *arXiv preprint arXiv:2310.06588*, 2023.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL `https://aclanthology.org/D15-1075`.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *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, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL `https://aclanthology.org/N18-1101`.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, 2023.

Mavuto M Mukaka. A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal*, 24(3):69–71, 2012.

Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. *Natural language processing with transformers.* " O'Reilly Media, Inc.", 2022.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

A Appendix

A.1 Explanation

Dai et al. (2023) explains language models as meta-optimizers and understands ICL as a kind of implicit fine-tuning:

$$\begin{aligned}\tilde{F}_{ICL}(q) &= W_{ZSL}q + W_V X'(W_K X')^T q \\ &= W_{ZSL}q + \Delta W_{ICL}q \\ &= (W_{ZSL} + \Delta W_{ICL})q\end{aligned} \tag{7}$$

where W_{ZSL} is the the initialized parameters of zero shot learning, X' denotes the input representations of the demonstration tokens, q is the attention query vector and W_K, W_V are the projection matrices for computing the attention keys and values respectively;

As shown in the above equation, the attention to the demonstration tokens is equivalent to parameter updates ΔW_{ICL} that take effect on W_{ZSL}. The authors explained in-context learning as a process of meta-optimization: (1) a pre-trained GPT model serves as a meta-optimizer; (2) it produces meta-gradients according to the demonstration examples through forward computation; (3) through attention mechanism, the meta-gradients are applied to the original language model to perform ICL (Dai et al., 2023). Therefore, I hypothesize that the instances which can improve the model's generalization ability during fine-tuning stage may also enhance the ICL ability of LLMs on unseen instances.

A.2 Inference Configuration

Temperature: This hyperparameter adjusts the randomness of predictions by scaling the logits before applying softmax. A value of 1 implies no additional scaling, maintaining the model's original distribution, while lower values reduce randomness and higher values increase it (Tunstall et al., 2022). In this paper, temperature is set to 1.

Top K: This parameter limits the number of candidate tokens considered for sampling. The model selects from the top K most likely tokens based on their likelihood scores, introducing a degree of controlled randomness by sampling from the most probable options (Holtzman et al., 2019). In this paper, Top K is set to 10.

Repetition Penalty: This parameter influences the model's tendency to repeat words or phrases during text generation. A value of 1 indicates no additional penalty, while values greater than 1 discourage repetition, enhancing the diversity of generated text (Tunstall et al., 2022).

In this paper, Repetition penalty is set to 1.

Max Tokens: This hyperparameter defines the maximum number of new tokens generated in the output. For efficiency and brevity, I set it to 5, which helps to limit the length of generated text and accelerate the inference process (Tunstall et al., 2022). In this paper, max tokens is set to 5.

In this configuration, temperature = 1 maintains the model's original predictive distribution while Top-k ensures that the model outputs the most likely words and ensures randomness, and repetition penalty is set to 1 so that the predicted label will not be influenced by the labels in the demonstrations.

A.3 Full Results

This Appendix will provide the results of all my experiments including 3-shot, 5-shot, 8-shot, 16-shot and 20-shot. Because the pearson correlation coefficient of 3-shot is relatively low, the experimental results of bias and influence-bias scores will not be shown here.

Methods	Average Accuracy	Variance
the entire dataset	0.661	0.0018024
easy-to-learn	0.659	0.0018528
ambiguous	0.661	0.0017332
hard-to-learn	0.655	0.0021904

Table 8: The mean and variance accuracy of four methods under 3 shots

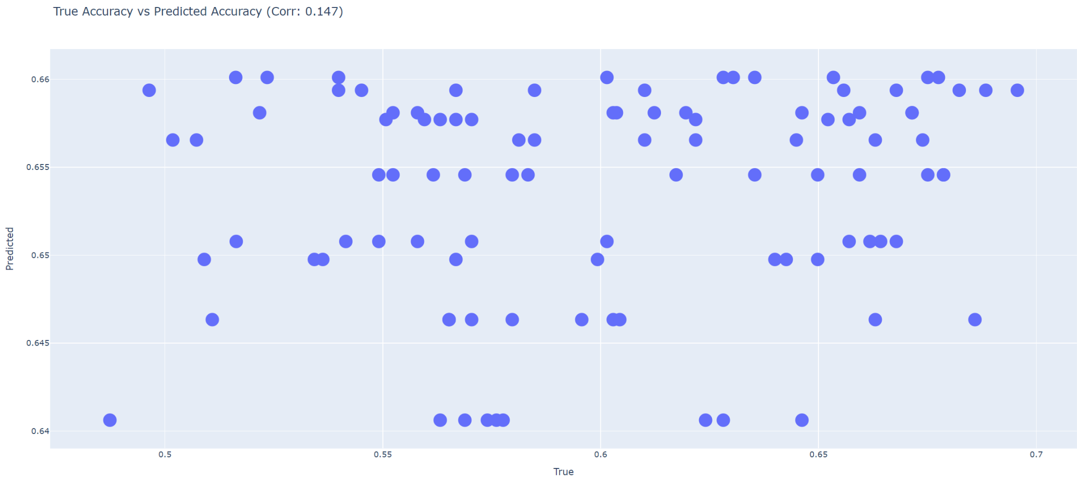

Figure 11: The predicted versus true accuracy of the linear in-context DataModel for the entire training dataset under 3 shots. Pearson Correlation Coefficient describes the strength and direction of the linear relationship between true accuracy and predicted accuracy

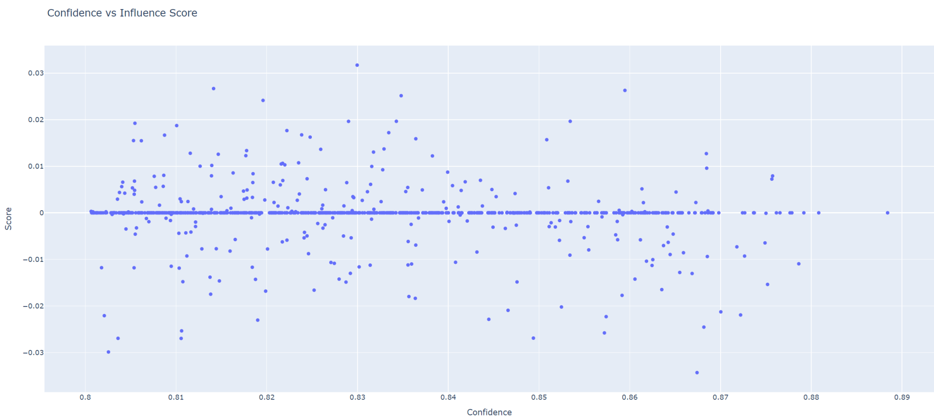

Figure 12: Scatter Plot of Confidence vs. Influence Score for the easy-to-learn region under 12 shots

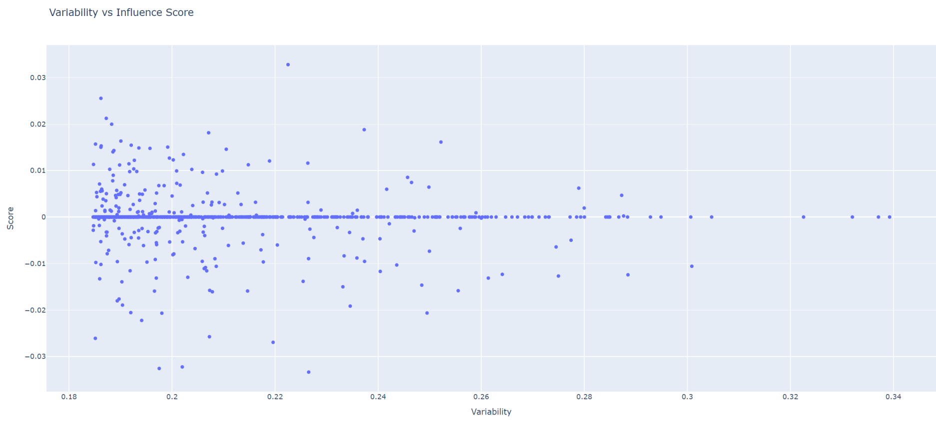

Figure 13: Scatter Plot of Variability vs. Influence Score for the ambiguous region under 12 shots

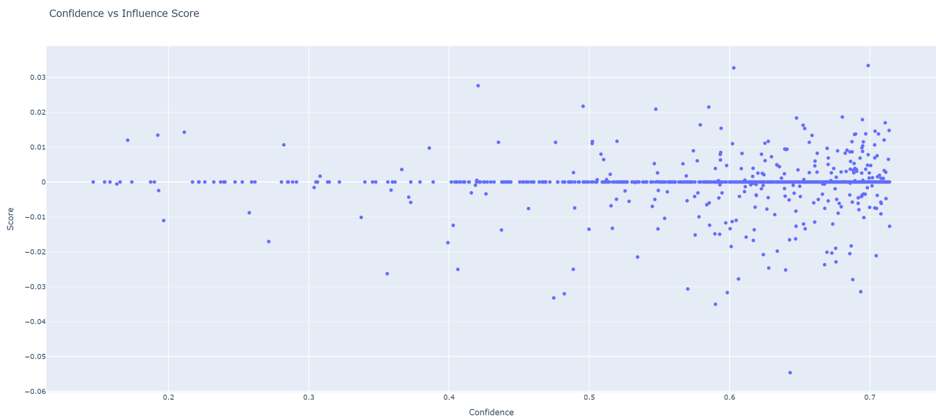

Figure 14: Scatter Plot of Confidence vs. Influence Score for the hard-to-learn region under 12 shots

Methods	Average Accuracy	Variance
the entire dataset	0.668	0.0019587
easy-to-learn	0.666	0.0019102
ambiguous	0.670	0.0017032
hard-to-learn	0.659	0.0027713

Table 9: The mean and variance accuracy of four methods under 5 shots

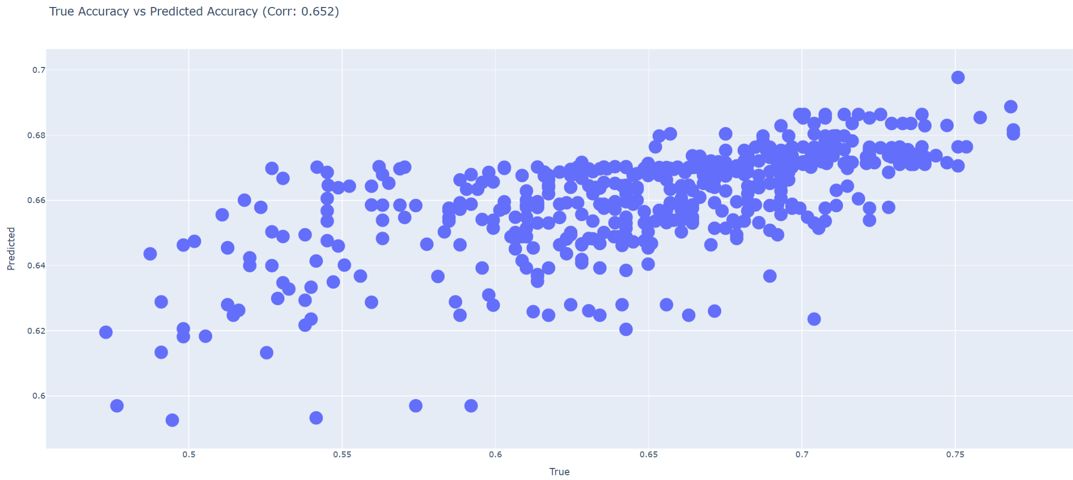

Figure 15: The predicted versus true accuracy of the linear in-context DataModel for the entire training dataset under 5 shots. Pearson Correlation Coefficient describes the strength and direction of the linear relationship between true accuracy and predicted accuracy

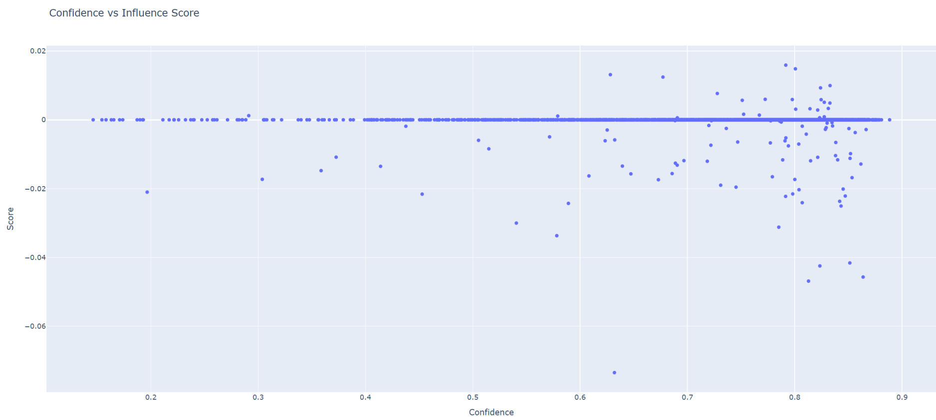

Figure 16: Scatter Plot of Confidence vs. Influence Score for the entire training dataset under 5 shots

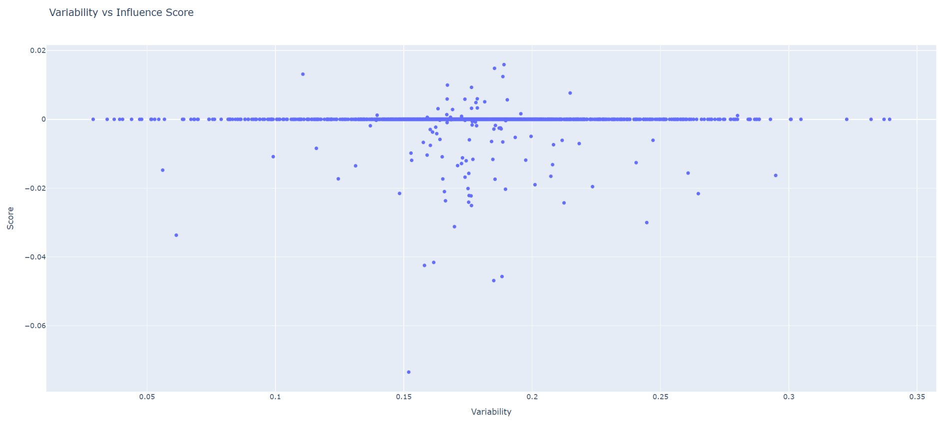

Figure 17: Scatter Plot of Variability vs. Influence Score for the entire training dataset under 5 shots

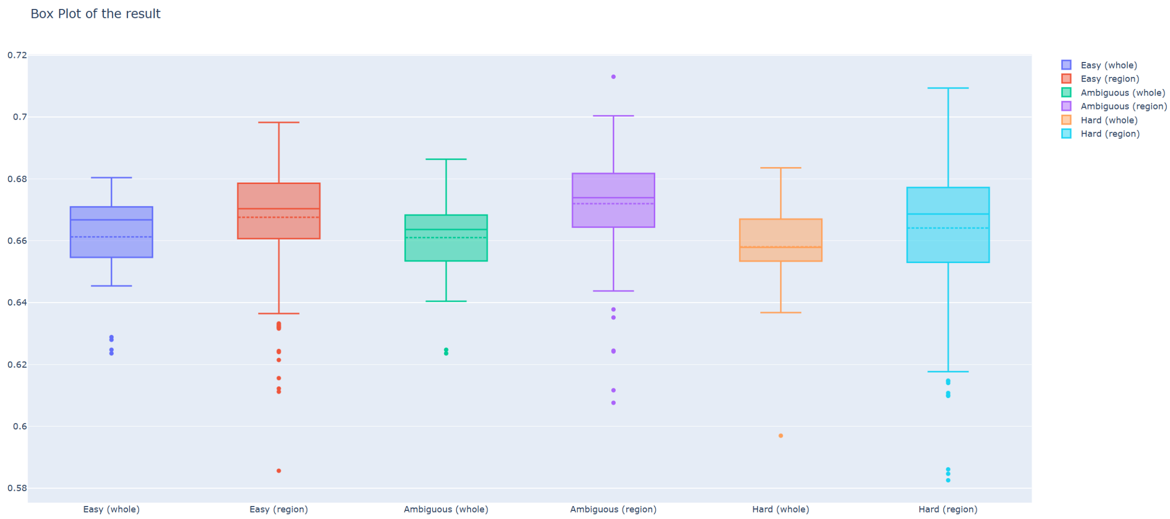

Figure 18: The Influence-Bias Scores of three regions under 5 shots; **Easy (whole)** represents the IBS of easy instances when other demonstrations are sampled from the entire dataset, while **Easy (region)** refers to the performance when other demonstrations are only sampled from the easy-to-learn region for ICL; the same applies to Ambiguous and Hard.

Methods	Average Accuracy	Variance
the entire dataset	0.676	0.0016903
easy-to-learn	0.674	0.0016477
ambiguous	0.678	0.0014744
hard-to-learn	0.665	0.0021193

Table 10: The mean and variance accuracy of four methods under 8 shots

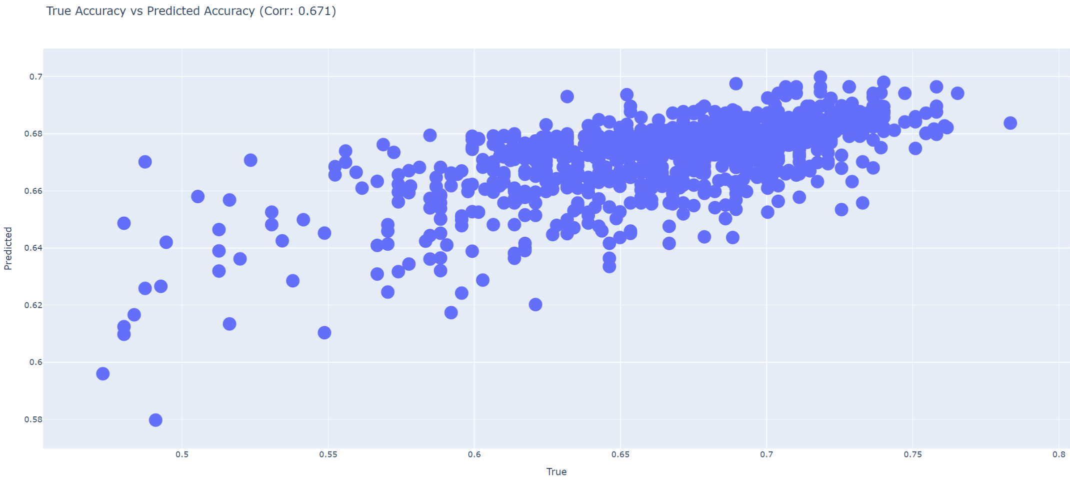

Figure 19: The predicted versus true accuracy of the linear in-context DataModel for the entire training dataset under 8 shots. Pearson Correlation Coefficient describes the strength and direction of the linear relationship between true accuracy and predicted accuracy

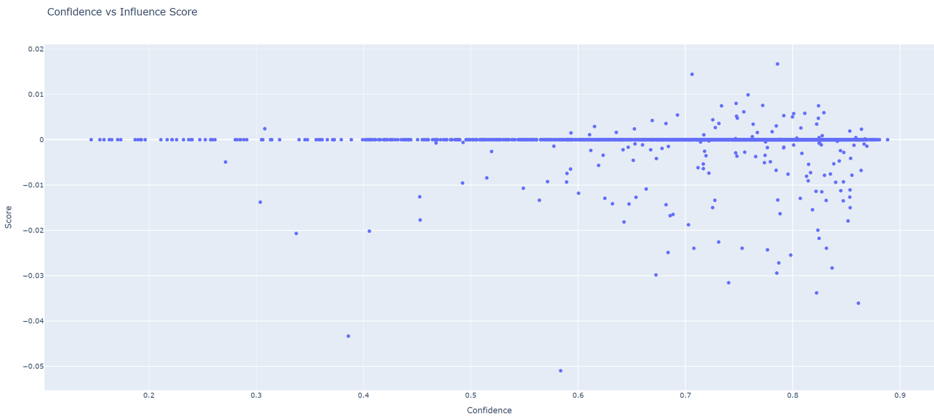

Figure 20: Scatter Plot of Confidence vs. Influence Score for the entire training dataset under 8 shots

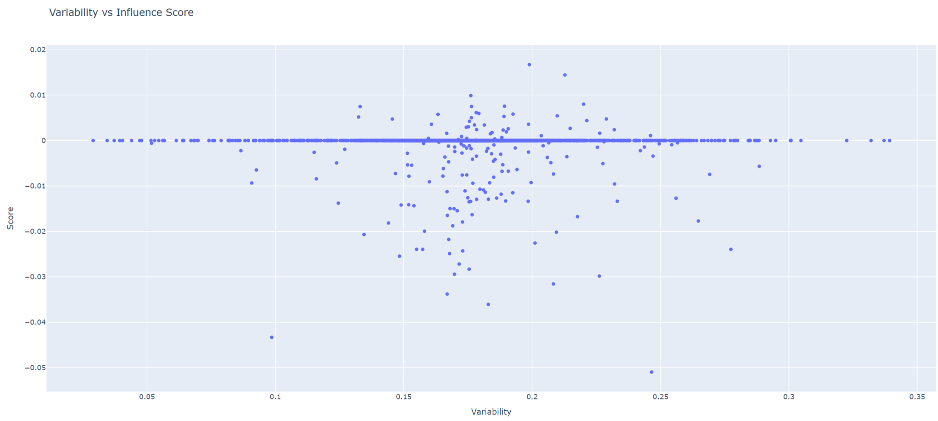

Figure 21: Scatter Plot of Variability vs. Influence Score for the entire training dataset under 8 shots

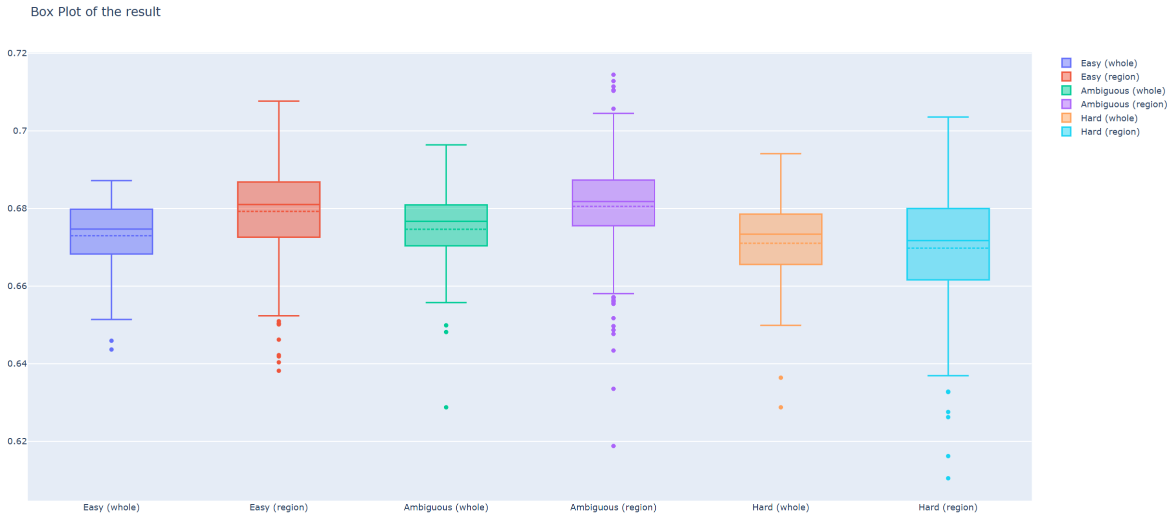

Figure 22: The Influence-Bias Scores of three regions under 8 shots; **Easy (whole)** represents the IBS of easy instances when other demonstrations are sampled from the entire dataset, while **Easy (region)** refers to the performance when other demonstrations are only sampled from the easy-to-learn region for ICL; the same applies to Ambiguous and Hard.

Methods	Average Accuracy	Variance
the entire dataset	0.684	0.0010789
easy-to-learn	0.682	0.0012179
ambiguous	0.685	0.0009928
hard-to-learn	0.669	0.0016272

Table 11: The mean and variance accuracy of four methods under 16 shots

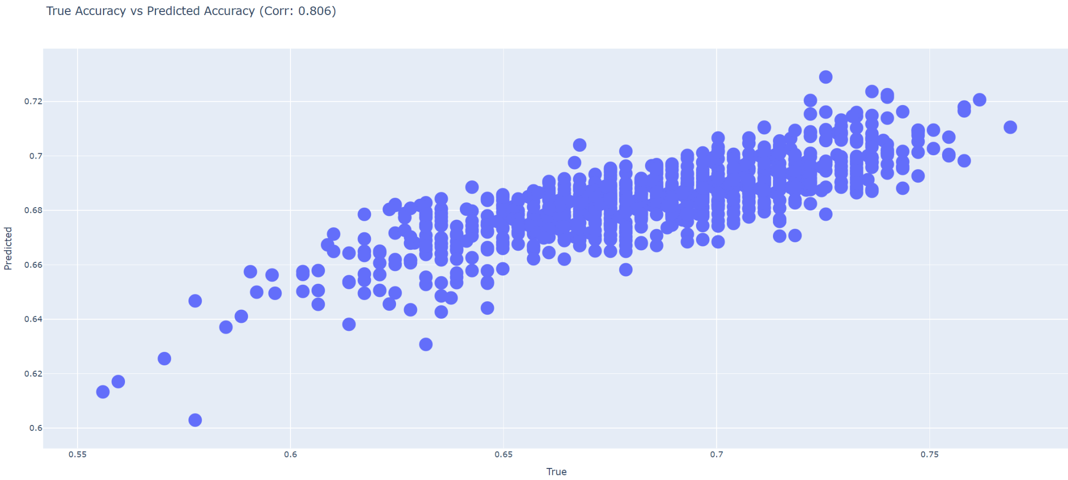

Figure 23: The predicted versus true accuracy of the linear in-context DataModel for the entire training dataset under 16 shots. Pearson Correlation Coefficient describes the strength and direction of the linear relationship between true accuracy and predicted accuracy

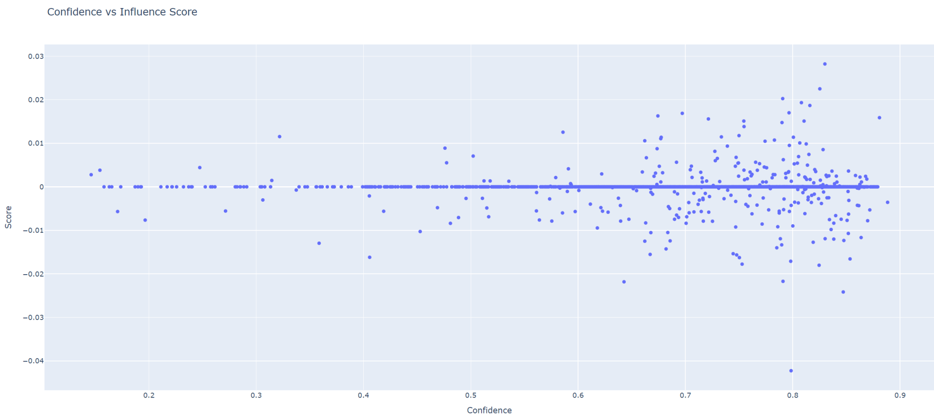

Figure 24: Scatter Plot of Confidence vs. Influence Score for the entire training dataset under 16 shots

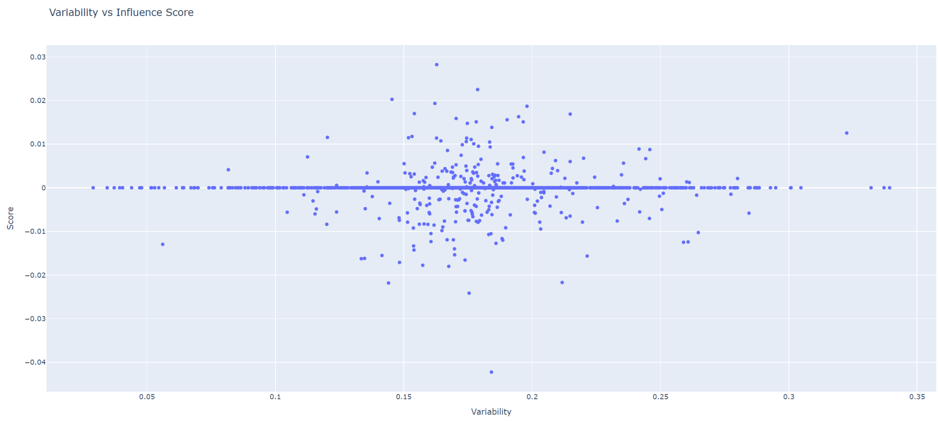

Figure 25: Scatter Plot of Variability vs. Influence Score for the entire training dataset under 16 shots

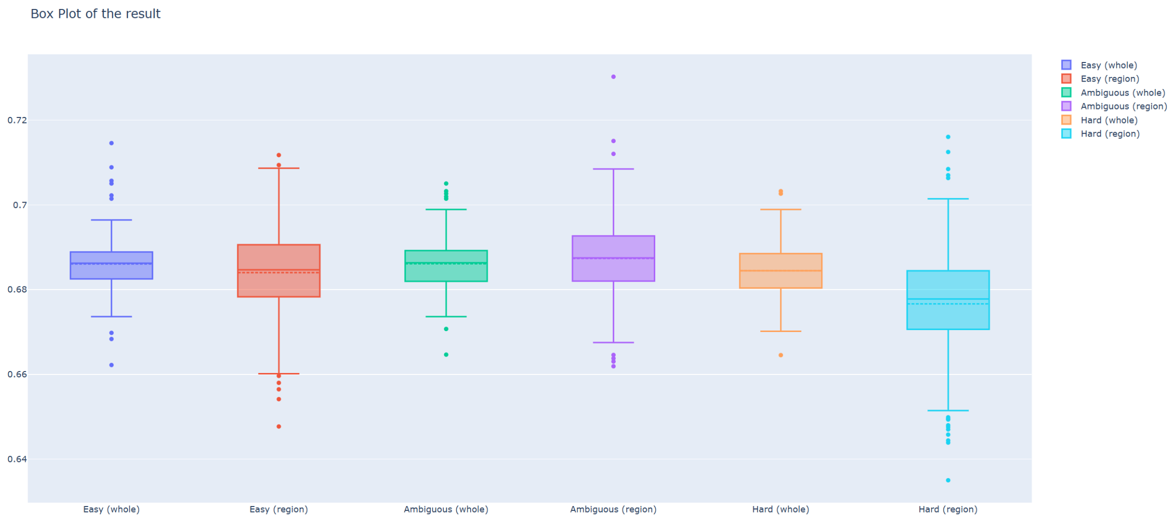

Figure 26: The Influence-Bias Scores of three regions under 16 shots; **Easy (whole)** represents the IBS of easy instances when other demonstrations are sampled from the entire dataset, while **Easy (region)** refers to the performance when other demonstrations are only sampled from the easy-to-learn region for ICL; the same applies to Ambiguous and Hard.

Methods	Average Accuracy	Variance
the entire dataset	0.688	0.0009908
easy-to-learn	0.690	0.00088003
ambiguous	0.689	0.00077313
hard-to-learn	0.675	0.00124258

Table 12: The mean and variance accuracy of four methods under 20 shots

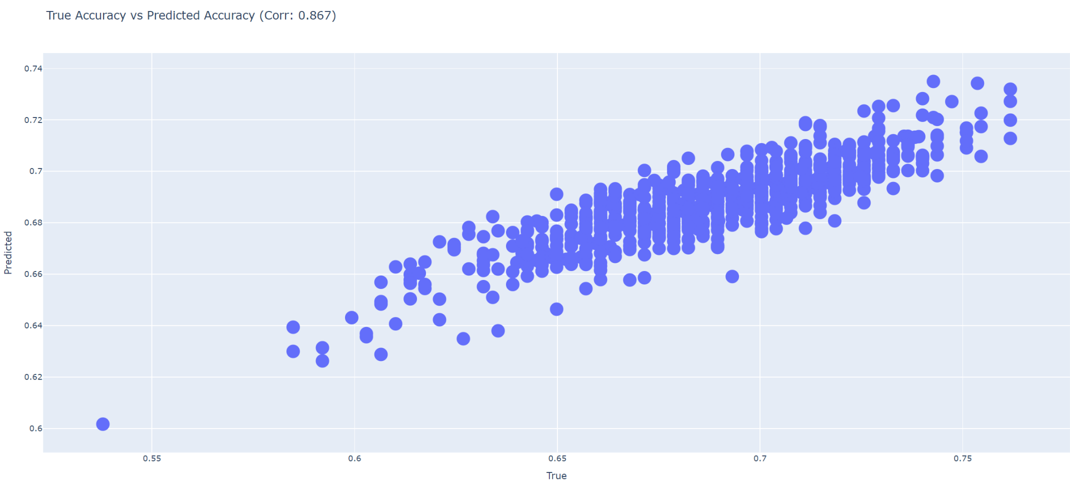

Figure 27: The predicted versus true accuracy of the linear in-context DataModel for the entire training dataset under 20 shots. Pearson Correlation Coefficient describes the strength and direction of the linear relationship between true accuracy and predicted accuracy

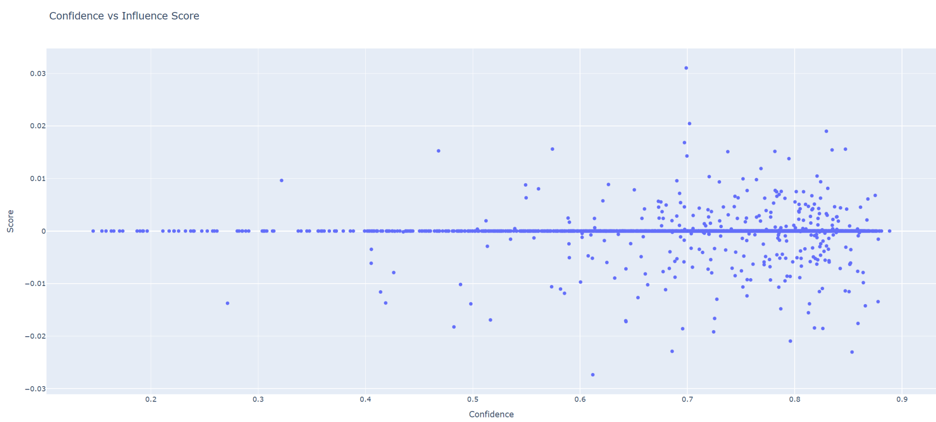

Figure 28: Scatter Plot of Confidence vs. Influence Score for the entire training dataset under 20 shots

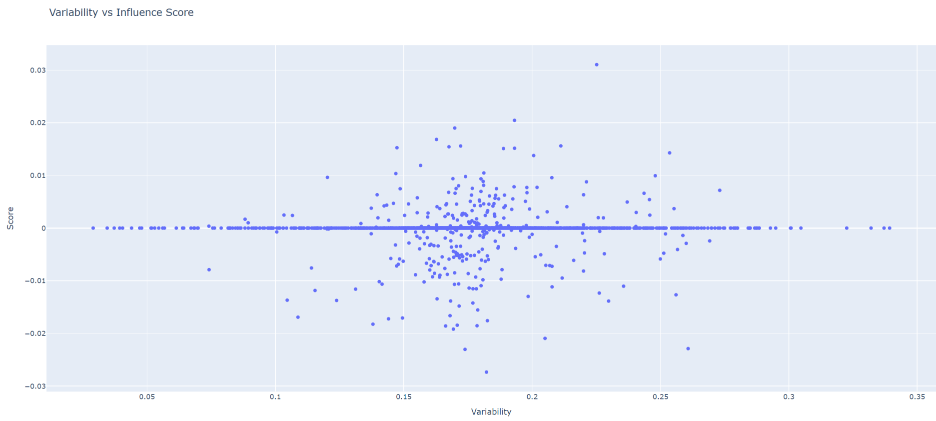

Figure 29: Scatter Plot of Variability vs. Influence Score for the entire training dataset under 20 shots

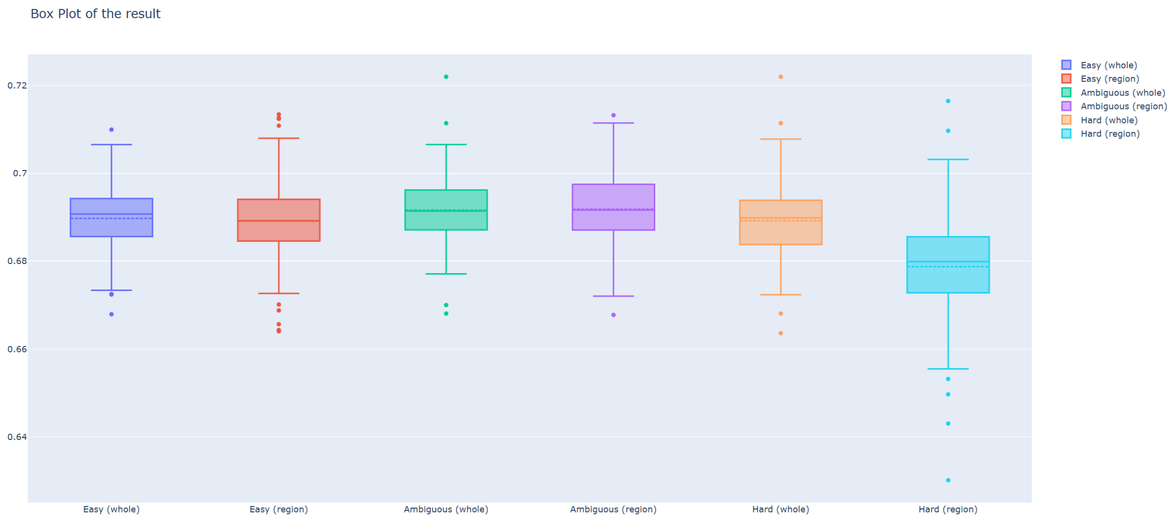

Figure 30: The Influence-Bias Scores of three regions under 20 shots; **Easy (whole)** represents the IBS of easy instances when other demonstrations are sampled from the entire dataset, while **Easy (region)** refers to the performance when other demonstrations are only sampled from the easy-to-learn region for ICL; the same applies to Ambiguous and Hard.