LEARNING-TO-CONTEXT SLOPE: EVALUATING IN-CONTEXT LEARNING EFFECTIVENESS BEYOND PERFORMANCE ILLUSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

In-context learning (ICL) has emerged as an effective approach to enhance the performance of large language models (LLMs). However, its effectiveness varies significantly across models and tasks, posing challenges for practitioners to determine when ICL reliably improves performance. Current evaluation approaches, reliant on performance change after applying ICL, suffer from low reliability, poor attribution, and impracticality in data-insufficient scenarios. We propose the **Learning-to-Context Slope (LCS)**, a novel metric that quantifies ICL effectiveness by modeling the slope between *learning gain* (loss decrease from demonstrations) and *contextual relevance* (demonstration-input relevance). LCS addresses key limitations of performance-based metrics: *(i)* it captures continuous loss changes even when outputs are incorrect, improving reliability; *(ii)* its formulation attributes ICL failures to weak contextual alignment (inability to adapt inputs to demonstrations) or strong output calibration (self-verification of correctness); and *(iii)* it minimizes reliance on labeled data via synthetic evaluation. Extensive experiments demonstrate that LCS strongly correlates with performance improvements in labeled settings and reliably reflects true effectiveness in biased or data-scarce scenarios. Further analysis reveals actionable thresholds for LCS and identifies model capabilities critical to ICL success[1].

1 INTRODUCTION

In-context learning (ICL) has emerged as a popular and effective paradigm for enhancing large language model (LLM) performance across diverse tasks, as it eliminates the need to retrain the LLMs (Brown et al., 2020; Dong et al., 2024). By incorporating task-specific demonstrations into the input, ICL enables LLMs to adapt to specific tasks and generate more accurate outputs without parameter updates. Recently, several efforts have been made on unveiling the underlying mechanisms of ICL (Zhou et al., 2024; Edelman et al., 2024a; Park et al., 2025) and exploring methods to further boost the ICL performance (Wang et al., 2023b; Rubin et al., 2022; Agarwal et al., 2024).

However, as illustrated in Figure 1a, even on the models with strong ICL capability like Llama3.1 (Grattafiori et al., 2024), ICL fails to enhance, and in some cases even harms, the performance (DeepSeek-AI et al., 2025; Huang & Wang, 2025; Zheng et al., 2025), showing different ICL effectiveness across different models. This observation raises a critical question: **How can practitioners reliably determine whether ICL is effective for a given model on a specific task**? This uncertainty poses practical challenges in the real-world deployment of ICL:

- For tasks *with labeled data*, practitioners often attempt to evaluate ICL effectiveness by observing performance changes after applying the selected demonstrations. However, this approach suffers from two critical limitations. *(i) Low Reliability*: Performance fluctuations may stem from various factors like the quality of the instruction and selected demonstrations, making it difficult to isolate whether ICL itself is ineffective. *(ii) Poor Attribution*: Disentangling the impact of individual factors requires costly repeated evaluations, hindering actionable analysis and insights.
- For tasks *without labeled data*, there is no direct way to assess whether adding demonstrations for ICL actually improves outcomes, leaving practitioners without clues for improvements.

[1]Our code and data will be released upon acceptance.

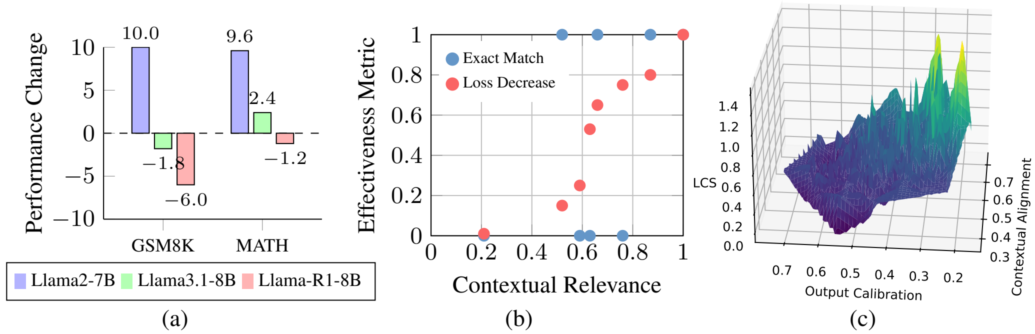


Figure 1: (a) Performance change of different models before and after applying ICL, where ICL exhibits varying effectiveness across different models on the same dataset. (b) Comparisons between metrics based on exact match and loss decrease. Each dot denotes an example data of MATH using Llama3.1-8b with different demonstrations. Performance-based metrics with only binary values fail to quantify the varying contributions of different demonstrations to achieving correct results. In contrast, metrics based on loss decrease yield continuous values, enabling better reliability on measuring ICL effectiveness. (c) The impact of the contextual alignment and output calibration capabilities of the model on the LCS metric.

In light of these challenges, we propose a novel metric, named **L**earning-to-**C**ontext **S**lope (**LCS**), which quantifies the ICL effectiveness by capturing the *slope between the loss decrease by demonstrations (**learning gain**) and the demonstration relevance to the user input (**contextual relevance**)*. Specifically, LCS is grounded in the perspective of the loss decrease in ICL (Wang et al., 2024b; Yang et al., 2024). For a given model and task, it evaluates how the learning gain varies with demonstrations of different contextual relevance. This metric explicitly captures the two most important elements in *in-context learning*: *learning* and *context* (Dong et al., 2024). When ICL effectiveness is high, even demonstrations with low relevance can yield a significant loss decrease. Conversely, when ICL effectiveness is low, the change of learning gain with demonstration relevance is marginal.

Compared to performance-based measurement, LCS offers the following advantages: *(i) Higher Reliability*: As shown in Figure 1b, even when ICL fails to produce correct answers for user inputs, LCS can still capture continuous changes in model loss, providing a more reliable reflection of ICL effectiveness. *(ii) Better Attribution*: LCS is grounded in an intuitive mathematical formulation, enabling clearer analysis of how different factors influence ICL effectiveness. As shown in Figure 1c, ICL tends to be ineffective when 1) the model fails to recognize the relevance of the demonstration to the input (*i.e.*, the *contextual alignment capability*), or 2) the model can independently verify the correctness of the output to the user input without adding demonstrations (*i.e.*, the *output calibration capability*). *(iii) Reduced Reliance on Labeled Evaluation Data*: We theoretically show that LCS derived from synthetic data is consistently lower than that obtained from real data, and empirically identify a threshold value of LCS indicative of effective ICL. Even in data-insufficient scenarios, LCS can still offer actionable insights into ICL effectiveness.

Our contributions can be summarized as follows:

- We propose a novel metric, namely Learning-to-Context Slope (LCS), to measure the ICL effectiveness by capturing the two most important elements in ICL, including the learning gain and the contextual relevance of the demonstrations.
- To validate the effectiveness of LCS, we conduct extensive experiments on eight mainstream datasets covering mathematics, code, reasoning, and domain-specific tasks (*e.g.*, finance and e-commerce). The results validate a strong positive correlation between LCS and task performance improvements in scenarios where abundant labeled data enables reliable performance-based evaluation. When labeled data exhibits inherent biases that distort performance-based metrics, LCS consistently reflects true ICL effectiveness, underscoring its reliability. Even without labeled data, LCS provides actionable insights into ICL effectiveness by leveraging synthetic data.
- Further analysis identifies two key factors in LLMs that hinder ICL effectiveness: 1) weak contextual alignment capability to adapt inputs to task-specific demonstrations, and 2) strong output calibration capability to independently verify the correctness of outputs.

2 Proposed Metric: Learn-to-Context Slope

We introduce a novel metric, named Learn-to-Context Slope (LCS), to measure the ICL effectiveness. First, we interpret the ICL effectiveness by measuring the loss decrease brought by using demonstrations based on the Bayesian model (§2.1). Then, we present our LCS metric to measure the ICL effectiveness, based on which we discuss two main factors that influence the ICL effectiveness (§2.2). Further, we discuss the relationship between the metric using synthetic data and real data, aiding the application under the data-insufficient scenario (§2.3). We discuss why the analysis is based on conditional probability in Appendix E.4.

2.1 Interpreting ICL Effectiveness via Loss Decrease

Motivated by previous studies (Wang et al., 2024b; Yang et al., 2024), the ICL effectiveness of a given predictive distribution p with the parameter θ on a specific task with the task $\mathcal{C} = (Q, X, D)$ can be measured by the generation loss, *i.e.,* negative log-likelihood:

$$\mathbb{L}_\theta(X|Q; D) = -\log p(X|Q; D), \tag{1}$$

where Q denotes the user input, X represents the labeled output corresponding to Q, and D denotes the demonstration, which are the random variables in the sampling spaces $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{D}$, respectively. Based on the Bayesian model (Zhang et al., 2025; Jesson et al., 2025), this loss can be obtained by:

$$\mathbb{L}_\theta(X|Q; D) = \mathbb{L}_\theta(X|Q) - (\log p(D|Q; X) - \log p(D|Q)), \tag{2}$$

where $\mathbb{L}_\theta(X|Q)$ represents the loss of zero-shot generation, which is fixed given the model and task. The proof of Equation 2 is presented in Appendix C.1. It can be observed that only the second term, *i.e.*, $\log p(D|Q; X) - \log p(D|Q)$, is relevant to the demonstrations, specifically the reduction in loss brought about by the demonstrations. Intuitively, this term also measures the information of the user output X that helps to decide the demonstration D.

2.2 Metric of ICL Effectiveness: LCS

For simplicity, we denote the **Learning Gain** brought by the demonstrations as $I_p(X \to D|Q) = p(D|Q; X) - p(D|Q)$ to reflect the decrease in loss, we discuss it in detail in Appendix E.5. To evaluate the overall effectiveness of the given specific model and task in ICL, we propose measuring effectiveness by assessing how the learning gain varies with demonstrations of different relevance. The motivation is that even demonstrations with low relevance to the user question can still lead to significant learning gains for tasks and models where ICL is highly effective. Conversely, when the ICL effectiveness is low, the change in learning gain with demonstration relevance is marginal. We measure the **Contextual Relevance** of the demonstration to the user question as $I_p(D \to X|Q) = p(X|Q; D) - p(X|Q)$. The contextual relevance is quantified by how much information for inferring the output X can be learned from the demonstration D in the context. To demonstrate that the contextual relevance defined by probability can reflect the demonstration relevance, we also compare it with other relevance measurement of the demonstration to the user question in Appendix F.2.

We have that the learning gain $I_p(X \to D|Q)$ and the contextual relevance $I_p(D \to X|Q)$ satisfy:

Theorem 1.

$$I_p(X \to D|Q) = \frac{p(D|Q)}{p(X|Q)} I_p(D \to X|Q)$$

The proof of Theorem 1 is presented in Appendix C.2. According to the theorem, learning gain and contextual relevance are positively correlated with a certain slope. A larger slope indicates a greater decrease in loss when increasing the information relevant to the user question of the demonstrations, thereby making ICL more effective.

In practice, let $\hat{p}$ represent the empirical probability distribution and $C = \{(q_i, x_i, d_i)\}_n$ be the sampling on $\mathcal{C}$, we calculate the slope of Theorem 1 ($r_{\hat{p}}$) on C with the least squares method (Wang et al., 2018):

$$r_{\hat{p}} = \frac{\sum_{i=1}^{n}(t_i - \bar{t})(s_i - \bar{s})}{\sum_{i=1}^{n}(t_i - \bar{t})^2}, \texttt{where}$$
$$s_i = I_{\hat{p}}(x_i \to d_i | q_i), t_i = I_{\hat{p}}(d_i \to x_i | q_i) \tag{3}$$
$$\bar{s} = \frac{1}{n}\sum_{i=1}^{n} s_i, \bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i.$$

We use $\boldsymbol{r_{\hat{p}}}$ as the metric to measure the ICL effectiveness, which we call the **Learning-to-Context Slope (LCS)**. Although $r_p = \frac{p(D|Q)}{p(X|Q)}$, considering that $\hat{p}$ has the error compared with p, $r_{\hat{p}} \neq \frac{\hat{p}(D|Q)}{\hat{p}(X|Q)}$. In Appendix C.3, we discuss the impact of error and prove that the impact of the error on $r_{\hat{p}}$ is less than $\frac{\hat{p}(D|Q)}{\hat{p}(X|Q)}$. We discuss how to calculate our metric in detail in Appendix D.2.

Based on Theorem 1, it can be observed that there are two main factors influencing the ICL effectiveness: ***the contextual alignment capability*** that learn the question-relevant information from the demonstrations ($\hat{p}(D|Q)$), and ***the output calibration capability*** that verify the correctness of the output to the given input ($\hat{p}(X|Q)$). Therefore, given a specific model and task, the reasons for poor ICL effectiveness can be attributed to two aspects: *(i) Low Contextual Alignment Ability*: The model fails to adequately comprehend the task-relevant information in the provided demonstrations. *(ii) High Output Calibration Capability*: The model possesses a strong inherent ability to verify the input consistency with the given output. We further discuss the meaning of the contextual alignment capability and the output calibration ability in detail in Appendix E.1.

2.3 ICL Effectiveness without Labeling

Since the calculation of LCS in § 2.2 relies on labeled data, its application to new tasks in data-insufficient scenarios is limited. Prior work has shown that the resource requirements for obtaining task questions are lower than those for obtaining the labels (Shen et al., 2019; Tan et al., 2024). Therefore, in this section, we discuss the relationship of LCS with synthetic data and real data, using only the labeled input, which satisfies the following:

Theorem 2. *Let $\hat{D}, D^*$ denote two demonstration satisfying that, for all $X \sim \mathcal{X}$ and $Q \sim \mathcal{Q}$:*

$$\hat{p}(\hat{D} \mid Q; X) \;\leq\; \hat{p}(D^* \mid Q; X).$$

The above condition means that demonstration D^* is better than $\hat{D}$ to help to generate the correct answer. Then, we can derive that:

$$\frac{\hat{p}(\hat{D}|Q)}{\hat{p}(\hat{X}|Q)} \leq \frac{\hat{p}(D^*|Q)}{\hat{p}(X^*|Q)}$$

The conclusion in Theorem 2 suggests that the more demonstrations that can help the model make correct predictions, the larger the corresponding LCS. Considering that previous work has shown that the quality of synthesized data is generally lower than annotated data (Ashok & May, 2025; Gulati et al., 2023), we can consider $\hat{D}$ as a synthesized demonstration and D^* as an annotated demonstration. Therefore, we can observe that LCS fitted with synthetic data is consistently smaller than that using real data. Consequently, while fitting synthetic data can reflect the ICL effectiveness to some extent, the magnitude of the effectiveness derived is lower than that of the real effectiveness.

3 Experiment

In this section, we empirically investigate three research questions about the ICL effectiveness: **RQ1**. How to reliably evaluate the ICL effectiveness? **RQ2**. How do different factors influence the ICL effectiveness? **RQ3**. Can synthetic data accurately reflect the ICL effectiveness?

Dataset	Llama2-7b Δ	Llama2-7b LCS	Llama3.1-8b Δ	Llama3.1-8b LCS	Llama-R1-8b Δ	Llama-R1-8b LCS
GSM8K	+10.0	0.32	−1.8	0.07	−6.0	0.05
MATH	+9.6	1.03	+2.4	0.34	−1.2	0.09
HumanEval	−0.6	0.07	−2.5	0.10	−3.0	−0.11
MBPP	−0.5	0.05	+0.8	0.15	−6.4	0.07
ARC-C	+11.6	0.74	−1.9	−0.54	−0.3	0.08
MMLU-Pro	+5.5	0.64	+2.6	0.52	−5.5	−0.04
FinQA	+7.3	0.63	+4.9	0.82	−1.8	0.04
Amazon	+0.5	0.07	+5.0	0.94	+11.8	0.37

Table 1: Performance and LCS across different models and datasets. Δ denotes the performance change of 1-shot compared to 0-shot. Results in green indicate a significant improvement with ICL, while those in red indicate no improvement or performance drop. Detailed performance is presented in Appendix F.1.

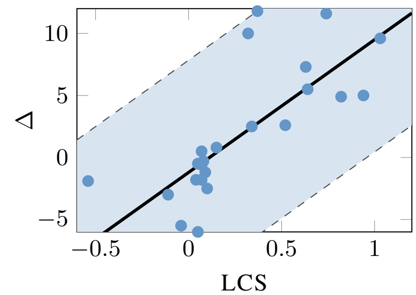


Figure 2: The performance improvement Δ brought by ICL (y-axis) with different LCS (x-axis) on different models and datasets. The solid line in the graph represents the fitted line for all data points. The Pearson correlation coefficient is 0.737.

3.1 Experiment Setup

Dataset We conduct experiments on four mainstream tasks: math (GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021)), code (HumanEval (Chen et al., 2021a), MBPP (Austin et al., 2021)), reason (ARC-Challenge (Yadav et al., 2019), MMLU-Pro (Wang et al., 2024d)), and domain-specific (FinQA (Chen et al., 2021b), Amazon Review (Ni et al., 2019)). We introduce the above dataset, as well as the split of the demonstrations and the test data, in Appendix D.1.

Metric For datasets of math, reasoning, and domain-specific, we use Exact Match (EM) (Cobbe et al., 2021) as the evaluation metric. For the datasets of code, we use Pass@1 (Chen et al., 2021a) as the metric. To prove that LCS is a better metric than performance change to reflect the ICL effectiveness, our main experiment includes two parts: *(i)* In §3.2.1, we evaluate that when performance change can reflect the effectiveness of ICL, LCS can also reflect the effectiveness of ICL. *(ii)* In §3.2.2, we present that LCS can still reflect the ICL effectiveness, even when the provided demonstrations do not lead to performance improvements.

Model We conduct our experiments on three mainstream LLMs: Llama2-7b (Touvron et al., 2023), Llama3.1-8b (Grattafiori et al., 2024), and DeepSeek-R1-Distill-Llama-8b (Llama-R1-8b) (DeepSeek-AI et al., 2025), which cover different ICL capabilities to fully evaluate whether our metric can reflect the ICL effectiveness. We also conduct experiments on models of other scales and series in Appendix F.3, further validating the effectiveness of LCS. We discuss how to adapt LCS to black-box LLMs in Appendix F.6.

Implementation Details We evaluate the performance on all datasets under 0-shot and 1-shot settings, using BM25 to select the demonstrations for each user input. We also discuss the performance and ICL effectiveness under different shots in §3.3.3. Following DeepSeek-AI et al. (2025), we set the maximum generation length to 32,768. Our experiments are conducted on a single A100-80G, with an average computation time of approximately 20 minutes on each dataset and model.

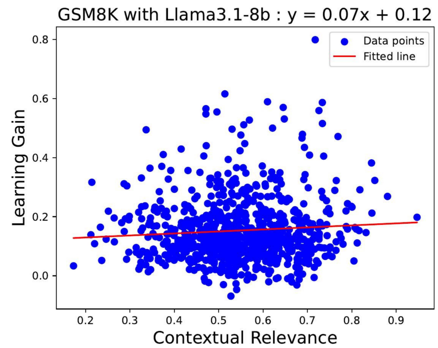
3.2 RQ1. How to Reliably Evaluate the Effectiveness of ICL?

First, we discuss that LCS can accurately reflect the effectiveness of ICL. Subsequently, we provide experimental evidence demonstrating that performance improvement is insufficient for accurately reflecting the ICL effectiveness. In addition, we present that LCS can reflect the performance improvement brought by ICL to a certain extent.

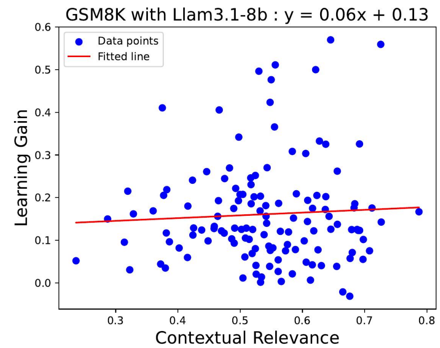
3.2.1 LCS Reliably Reflects the ICL Effectiveness

According to the main experimental results shown in Table 1, there are several notable observations:

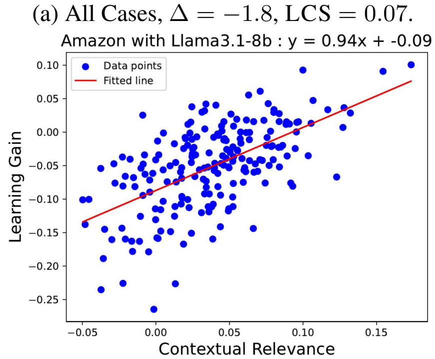
The ICL effectiveness is independent of dataset difficulty. For Llama2-7b, ICL is effective on the MATH dataset but fails on the easier Amazon Review dataset. Conversely, for Llama-R1-



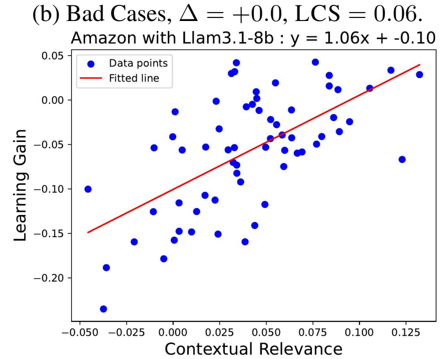
(a) All Cases, $\Delta = -1.8$, LCS $= 0.07$.



(b) Bad Cases, $\Delta = +0.0$, LCS $= 0.06$.



(c) All Cases, $\Delta = +5.0$, LCS $= 0.94$.



(d) Bad Cases, $\Delta = +0.0$, LCS $= 1.06$.

Figure 3: The experimental results of using Llama3.1-8b on GSM8K and Amazon under the full set and the bad cases of ICL. Δ denotes the performance change of ICL compared with zero-shot.

8b, ICL is ineffective on MATH but performs well on Amazon Review. This discrepancy arises because, for more difficult datasets, the model struggles to comprehend the relationships between demonstrations, answers, and user questions, leading to a decline in both the ICL ability and the answer verification ability. Consequently, it is uncertain whether LCS increases or decreases on more difficult datasets, supporting that ICL effectiveness is irrelevant to the difficulty of the dataset.

The ICL effectiveness is independent of model capability. In Amazon Review, Llama-R1-8b demonstrates a significant improvement with ICL, whereas the less capable Llama2-7b does not exhibit a noticeable performance improvement. This discrepancy arises because, as the model capability increases, both the contextual alignment capability and the output calibration capability increase simultaneously, making it uncertain whether LCS rises or falls.

The performance improvement brought by ICL is positively related to LCS. To evaluate whether LCS effectively reflects the efficacy of ICL, we analyze the performance improvement with different LCS, as illustrated in Figure 2. A high LCS suggests that the model achieves higher learning gain as the contextual relevance increases, demonstrating that LLMs learn how to solve the task from the demonstrations, thereby improving performance. In contrast, a low LCS indicates that the learning gain from the demonstrations remains relatively constant regardless of the contextual relevance, implying limited learning from the demonstrations. Notably, the relationship between the change in EM and LCS is not strictly linear. Since the factors influencing EM are complex and difficult to formalize, in this paper, we only conclude that LCS is positively correlated with the change in EM. We discuss the empirical threshold of the ICL effectiveness using LCS in Appendix E.2.

3.2.2 The Performance Change Cannot Reflect the ICL Effectiveness

In §3.1, we assume that whether performance improves or not can genuinely reflect the ICL effectiveness. However, in practical applications, the quality of demonstrations or instructions can impact performance, causing no performance improvement even for models and tasks where ICL is effective. To demonstrate that LCS can still reflect the ICL effectiveness even when performance does not improve, we plot $r_{\hat{p}}$ on the bad cases after using ICL, as shown in Figure 3. It can be observed

Table 2: The performance of ICL with different demonstration selection methods using Llama3.1-8b. The best performance of each setting is marked in **bold**.

Method	**GSM8K**	**MATH**	**ARC-C**	**MMLU-Pro**	**FinQA**	**Amazon**
Zero-Shot	86.4	48.4	82.1	50.4	49.7	63.5
BM25 (Robertson & Zaragoza, 2009)	84.2	50.8	80.2	53.0	54.6	68.5
GTR (Luo et al., 2023)	82.1	50.8	80.5	53.5	55.0	68.5
Yang et al. (2023)	84.2	50.2	80.9	53.3	55.0	69.0
Influence (Nguyen & Wong, 2023)	83.9	51.0	81.3	52.4	54.6	69.5
IDS (Qin et al., 2024)	85.3	50.4	82.4	52.4	54.6	68.0
Ours	**86.4**	**51.2**	**82.5**	**54.3**	**55.1**	**70.0**

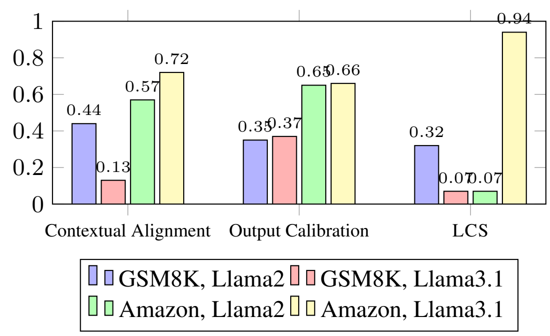


Figure 4: The results of the contextual alignment capability ($\hat{p}(D|Q)$), the output calibration capability ($\hat{p}(X|Q)$) and LCS of Llama2-7b and Llama3.1-8b on GSM8K and Amazon Review. $\hat{p}(D|Q)$ and $\hat{p}(X|Q)$ are calculated as the average value on all test data.

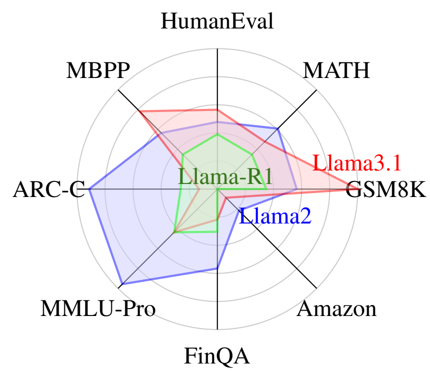


Figure 5: The intercept of the fitted line on each dataset and each model. We also compare the intercepts of Llama3.1 under different scales in Appendix F.4.

that: *(i)* Even on data where ICL does not improve performance, LCS still reveals the ICL effectiveness, proving the higher reliability of our metric compared with the performance-based metric. *(ii)* LCS is higher reliability, unlike performance which is susceptible to factors like the instruction, as it directly evaluates $p(X|Y)$ by using Y as input and X as output without relying on instructions (Appendix D.2), thus providing a more faithful reflection of the ICL effectiveness.

3.2.3 The Learning Gain is a Good Metric for Demonstration Selection

Enhancing ICL performance has been a topic of significant interest. Although this paper does not primarily focus on improving ICL performance, the discussions in §3.2.1 reveal several potential avenues for improvement. We observe that while there is a general positive correlation between the decrease in loss and the information learned from demonstrations, there also exist cases where demonstrations with rich information yield a low decrease in loss. To address this, we propose a method that first generates a preliminary answer $\hat{X}$ for the user question and then selects the demonstrations with a high learning gain. As shown in Table 2, our method outperforms other baselines, demonstrating the effectiveness of the learning gain-based method. In addition, on the datasets with low ICL effectiveness (e.g., GSM8K, ARC-Challenge), all methods have not brought significant improvement, which is consistent with our conclusion in §3.2.1.

3.3 RQ2. How Do Different Factors Influence the ICL Effectiveness?

3.3.1 The Main Factors that Influence the ICL Effectiveness

In §2.2, we discuss the main factors influencing the ICL effectiveness, including the contextual alignment capability ($\hat{p}(D|Q)$) and the output calibration capability ($\hat{p}(X|Q)$). In this section, we conduct experiments to analyze these conclusions further. We calculate the average values of $\hat{p}(D|Q)$ and $\hat{p}(X|Q)$ with Llama2-7b and Llama3.1-8b on GSM8K and Amazon Review, as shown in Figure 4.

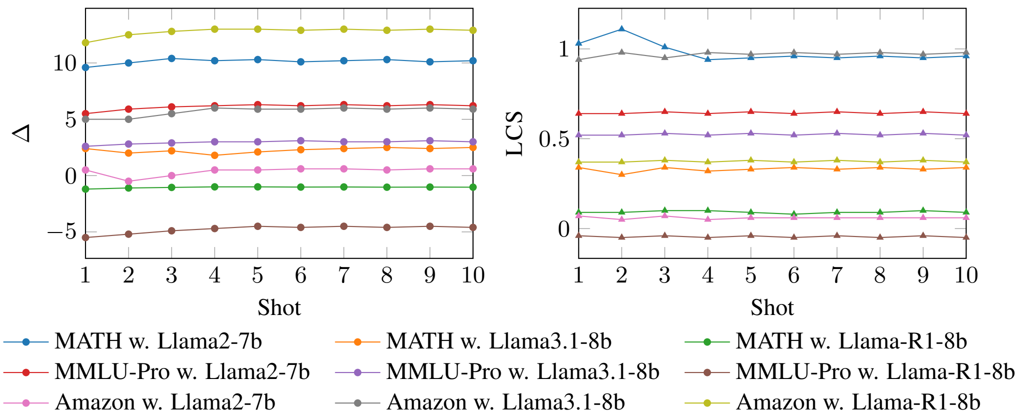


Figure 6: The performance change (Y-axis, left figure) and LCS (Y-axis, right figure) on MATH, MMLU-Pro, and Amazon Review with different shots (X-axis). The lines of the same color denote the results under the same setting.

From the figure, we can observe the following: *(i)* The results of Llama2-7b on Amazon Review indicate that the model is unable to effectively learn the information relevant to the user input from the provided demonstration D, *i.e.*, the contextual alignment capability is low, which leads to poor ICL effectiveness; *(ii)* The results of Llama3.1-8b on GSM8K show that although the ICL ability of the model is high, the model can accurately assess the relationship between input and output, *i.e.,* the output calibration capability also diminishes the ICL effectiveness; *(iii)* LCS is not equal to $\frac{\hat{p}(D|Q)}{\hat{p}(X|Q)}$, due to the error between p and $\hat{p}$, as discussed in detail in Appendix C.3.

3.3.2 It is Harder for ICL to Improve the Learning Gain on Stronger Model

Apart from the slope, the intercept of the fitted line also reflects the effectiveness of ICL under different settings. We examine the intercept under different datasets and models, which is shown in Figure 5. From the figure, we observe that as model capacity increases, the corresponding intercepts decrease, indicating that: *(i)* From the perspective of the learning gain, the intercept reflects the overall magnitude of learning gain attributed to demonstrations for a given model and task, where a smaller intercept suggests less learning gain. *(ii)* From the perspective of error estimation (Appendix C.3), a smaller intercept implies a smaller discrepancy between p and $\hat{p}$, meaning that the empirical predictor more closely approximates the oracle predictor. In summary, as model capacity increases, model predictions become more aligned with the oracle predictor, but the overall learning gain from demonstrations also diminishes.

3.3.3 More Shots Improve the ICL performance but not Effectiveness

To observe the differences in the ICL effectiveness under varying shot numbers, we conduct experiments with different shot numbers. Since Theorem 1 can calculate the influence of only a single demonstration, we divide the k-shot into k data points to calculate LCS. The experimental results are shown in Figure 6. From the figure, we can observe that: *(i)* As the number of shots increases, the overall performance change shows an upward trend. However, LCS does not generally increase or decrease with the number of shots but rather exhibits some degree of fluctuation. This is because the value of LCS is related to the inherent ICL effectiveness on a given model and dataset, while increasing the shot number cannot affect the ICL effectiveness. *(ii)* Relatively, the fluctuation of LCS gradually decreases as the number of shots increases. As discussed in Appendix C.3, increasing the number of shots can reduce computational errors, making the calculated result of LCS more stable and a more accurate reflection of the ICL effectiveness.

3.4 RQ3. Can Synthetic Data Accurately Reflect the ICL Effectiveness?

To verify the conclusions regarding the computation of LCS for synthetic data presented in §2.3, we conduct experiments to calculate LCS using synthetic data. During synthesis, we follow the procedure in Wang et al. (2025b) by inputting the task definition to generate corresponding demon-

Table 3: Performance change (Δ) and LCS using labeled and synthetic demonstrations.

Dataset	**Type**	**Llama2-7b**		**Llama3.1-8b**		**Llama-R1-8b**	
		Δ	LCS	Δ	LCS	Δ	LCS
MATH	Labeled	+9.6	1.03	+2.4	0.34	−1.2	0.09
	Synthetic	+6.0	0.75	+1.3	0.16	−1.8	0.05
ARC-C	Labeled	+11.6	0.74	−1.9	−0.54	−0.3	0.08
	Synthetic	+8.2	0.53	−1.5	−0.56	−0.2	0.06
MMLU-Pro	Labeled	+5.5	0.64	+2.6	0.52	−5.5	−0.04
	Synthetic	+3.6	0.33	+2.0	0.32	−4.0	−0.12
Amazon	Labeled	+0.5	0.07	+5.0	0.94	+11.8	0.37
	Synthetic	+0.0	0.0	+4.0	0.42	+9.0	0.25

strations. In each iteration, we provide the model with the task definition and the synthetic results from the previous iteration (empty in the first iteration), ask the model to generate demonstrations. The prompt we used is shown in Appendix D.3. We set the temperature to 0.9 and top_p to 0.9, sampling 8 demonstrations per iteration. A multi-round iterative process is used to ensure the diversity and quality of the synthesized demonstrations. Considering the computational resource limit, we only adapted experiments on four datasets, which are shown in Table 3. From the table, we observe the following: *(i)* The trend of LCS using synthetic data is consistent with that derived from labeled data, demonstrating that synthetic data can effectively reflect the ICL effectiveness. *(ii)* Compared to labeled data, the values of LCS obtained from synthetic data are relatively smaller, which supports the conclusion of Theorem 2.

4 Related Work

In-Context Learning In-context learning guides the LLM reasoning process by providing several task-relevant demonstrations in the input, thereby improving performance (Brown et al., 2020; Dong et al., 2024; Zhao et al., 2025). Existing ICL research can be broadly categorized into two main areas: constructing high-quality demonstrations and improving demonstration selection performance. For demonstration construction, many works focus on the offline enhancement of demonstration quality. This includes methods aimed at increasing demonstration diversity, for instance, by generating synthetic data tailored to a given task or by selecting diverse demonstrations to improve compositional generalization (Wang et al., 2024a; 2025a; Chen et al., 2023a; Su et al., 2024; Levy et al., 2022). Another key aspect of offline construction is synthesizing or augmenting reasoning steps within existing demonstrations to better guide the inference process (Li et al., 2024; Zelikman et al., 2022; ZHAO et al., 2023). Other methods focus on the online synthesis of demonstrations, where demonstrations are generated or rewritten dynamically based on the user input to enhance reasoning performance, sometimes even leveraging the LLM itself to create these demonstrations (He et al., 2024; Chang & Fosler-Lussier, 2023; Kim et al., 2022). In the domain of demonstration selection, research primarily explores how to choose demonstrations most relevant to the user query, with some approaches also incorporating active learning principles to identify the most informative demonstration (Luo et al., 2024; Vu et al., 2023). Selection strategies include those based on n-grams (Li et al., 2023), semantic similarity using embeddings (Yang et al., 2023; Luo et al., 2023), or hybrid methods that combine multiple diverse strategies for retrieval and ranking (Wan et al., 2025; Wang et al., 2024c; Hao et al., 2022).

Mechanism Analysis of In-Context Learning Many studies have investigated the mechanisms underlying ICL for improving reasoning performance (Zhou et al., 2024; Dong et al., 2024). One line of research explores the mechanism of ICL by controlling the types of tasks used during pre-training (Edelman et al., 2024b; Han et al., 2023). Current mainstream work suggests that ICL ability arises from task diversity rather than data scale, with models gradually generalizing from solving in-domain tasks to solving out-of-domain tasks (Raventos et al., 2023). Additionally, some studies find that the modules responsible for knowledge acquisition and ICL ability are functionally independent (Nguyen & Reddy, 2025). Increasing the amount of data primarily enhances the knowledge-related components, while improvements in ICL depend more on the diversity of tasks encountered during training. Another line of work focuses on ICL reasoning, aiming to discover the ICL mechanism by examining the relationship between provided demonstrations and the user question (Park et al., 2025; Li et al., 2025; Min et al., 2022; Wang et al., 2023a). Some studies argue that models perform ICL by learning the mapping between inputs and labels in the demonstrations, thereby improving task-solving performance (Kossen et al., 2024). Other research suggests that models learn the reasoning process embedded in the demonstrations and enhance reasoning performance by understanding and mimicking these processes (Lampinen et al., 2022).

However, the aforementioned studies mainly focus on improving the ICL performance or explaining the mechanism of ICL, often presuming that ICL is inherently effective. In contrast, recent studies have shown that ICL does not lead to performance improvement on certain tasks and models (DeepSeek-AI et al., 2025; Huang & Wang, 2025; Zheng et al., 2025). In this work, we investigate the main factors influencing the ICL effectiveness and propose the metric to evaluate the ICL effectiveness, to inform and inspire future research.

5 Conclusion

In this paper, we propose a novel metric LCS, to evaluate the ICL effectiveness. LCS overcomes the low reliability and poor attribution issues of performance-based metrics by measuring the variation in the learning gain with the contextual relevance. Based on LCS, we first discuss two primary factors that contribute to poor ICL effectiveness: poor contextual alignment capability and strong output calibration capability, demonstrating the strong attribution of LCS. Analytical experiments show that LCS can effectively reflect the effectiveness of ICL even on demonstrations where ICL does not lead to performance improvements, indicating high reliability. Furthermore, we present that the results of LCS on synthetic data are lower than those on real data, to inspire the application of LCS in data-insufficient scenarios.

6 Reproducibility

We have provided all proofs of this paper in Appendix C.1, Appendix C.2 and Appendix C.4. We will release the experimental and pre-processed data and code upon the paper being accepted.

References

Rishabh Agarwal, Avi Singh, Lei M Zhang, Bernd Bohnet, Luis Rosias, Stephanie C.Y. Chan, Biao Zhang, Aleksandra Faust, and Hugo Larochelle. Many-shot in-context learning. In *ICML 2024 Workshop on In-Context Learning*, 2024. URL https://openreview.net/forum?id=goi7DFHlqS.

Dhananjay Ashok and Jonathan May. A little human data goes a long way. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 381–413, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-252-7. doi: 10.18653/v1/2025.acl-short.30. URL https://aclanthology.org/2025.acl-short.30/.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models, 2021. URL https://arxiv.org/abs/2108.07732.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(null):1137–1155, March 2003. ISSN 1532-4435.

Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse, and Geoffrey Zweig. Syntactic clustering of the web. *Computer Networks and ISDN Systems*, 29(8):1157–1166, 1997. ISSN 0169-7552. doi: https://doi.org/10.1016/S0169-7552(97)00031-7. URL https://www.sciencedirect.com/science/article/pii/S0169755297000317. Papers from the Sixth International World Wide Web Conference.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Shuaichen Chang and Eric Fosler-Lussier. Selective demonstrations for cross-domain text-to-SQL. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14174–14189, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.944. URL https://aclanthology.org/2023.findings-emnlp.944/.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021a.

Wei-Lin Chen, Cheng-Kuang Wu, Yun-Nung Chen, and Hsin-Hsi Chen. Self-ICL: Zero-shot in-context learning with self-generated demonstrations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15651–15662, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.968. URL https://aclanthology.org/2023.emnlp-main.968/.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*, 2023b. ISSN 2835-8856. URL https://openreview.net/forum?id=YfZ4ZPt8zd.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. Finqa: A dataset of numerical reasoning over financial data. *Proceedings of EMNLP 2021*, 2021b.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context, 2019. URL https://arxiv.org/abs/1901.02860.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang

Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL `https://arxiv.org/abs/2501.12948`.

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 (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 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/`.

Ezra Edelman, Nikolaos Tsilivis, Benjamin L. Edelman, eran malach, and Surbhi Goel. The evolution of statistical induction heads: In-context learning markov chains. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL `https://openreview.net/forum?id=qaRT6QTIqJ`.

Ezra Edelman, Nikolaos Tsilivis, Benjamin L. Edelman, eran malach, and Surbhi Goel. The evolution of statistical induction heads: In-context learning markov chains. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL `https://openreview.net/forum?id=qaRT6QTIqJ`.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Samaksh Gulati, Anshit Verma, Manoj Parmar, and Palash Chaudhary. Efficacy of machine-generated instructions, 2023. URL https://arxiv.org/abs/2312.14423.

Qi Guo, Leiyu Wang, Yidong Wang, Wei Ye, and Shikun Zhang. What makes a good order of examples in in-context learning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14892–14904, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.884. URL https://aclanthology.org/2024.findings-acl.884/.

Xiaochuang Han, Daniel Simig, Todor Mihaylov, Yulia Tsvetkov, Asli Celikyilmaz, and Tianlu Wang. Understanding in-context learning via supportive pretraining data. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12660–12673, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.708. URL https://aclanthology.org/2023.acl-long.708/.

Yaru Hao, Yutao Sun, Li Dong, Zhixiong Han, Yuxian Gu, and Furu Wei. Structured prompting: Scaling in-context learning to 1,000 examples, 2022. URL https://arxiv.org/abs/2212.06713.

Wei He, Shichun Liu, Jun Zhao, Yiwen Ding, Yi Lu, Zhiheng Xi, Tao Gui, Qi Zhang, and Xuanjing Huang. Self-demos: Eliciting out-of-demonstration generalizability in large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3829–3845, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.243. URL https://aclanthology.org/2024.findings-naacl.243/.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=7Bywt2mQsCe.

Donghao Huang and Zhaoxia Wang. Explainable sentiment analysis with deepseek-r1: Performance, efficiency, and few-shot learning, 2025. URL https://arxiv.org/abs/2503.11655.

Andrew Jesson, Nicolas Beltran-Velez, and David Blei. Can generative AI solve your in-context learning problem? a martingale perspective. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bcynT7s2du.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Masahiro Kaneko, Youmi Ma, Yuki Wata, and Naoaki Okazaki. Sampling-based pseudo-likelihood for membership inference attacks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and

Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8894–8907, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.465. URL https://aclanthology.org/2025.findings-acl.465/.

Gyeonghary Kim, Won Ik Cho, Seok Min Shin, Sang-goo Lee, and Jamin Kim. Self-generated in-context learning: Leveraging auto-regressive language models as a demonstration generator. *arXiv preprint arXiv:2206.08082*, 2022.

Jannik Kossen, Yarin Gal, and Tom Rainforth. In-context learning learns label relationships but is not conventional learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=YPIA7bgd5y.

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. Can language models learn from explanations in context? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 537–563, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.38. URL https://aclanthology.org/2022.findings-emnlp.38/.

Noah Lee, Na Min An, and James Thorne. Can large language models capture dissenting human voices? In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4569–4585, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.278. URL https://aclanthology.org/2023.emnlp-main.278/.

Omer Levy, Gabriel Poesia, Sewon Min, Romain Paulus, Luke Zettlemoyer, and Mike Lewis. Diverse demonstrations improve in-context compositional generalization. *arXiv preprint arXiv:2211.12703*, 2022.

Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Eric Tang, Sumanth Hegde, Kourosh Hakhamaneshi, Shishir G. Patil, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Llms can easily learn to reason from demonstrations structure, not content, is what matters!, 2025. URL https://arxiv.org/abs/2502.07374.

Junlong Li, Jinyuan Wang, Zhuosheng Zhang, and Hai Zhao. Self-prompting large language models for zero-shot open-domain QA. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 296–310, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.17. URL https://aclanthology.org/2024.naacl-long.17/.

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 (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 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/.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatGPT really correct? rigorous evaluation of large language models for code generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=1qvx610Cu7.

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 (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 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/.

Man Luo, Xin Xu, Zhuyun Dai, Panupong Pasupat, Mehran Kazemi, Chitta Baral, Vaiva Imbrasaite, and Vincent Y Zhao. Dr.icl: Demonstration-retrieved in-context learning, 2023. URL https://arxiv.org/abs/2305.14128.

Man Luo, Xin Xu, Yue Liu, Panupong Pasupat, and Mehran Kazemi. In-context learning with retrieved demonstrations for language models: A survey. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=NQPo8ZhQPa. Survey Certification.

Aman Madaan, Katherine Hermann, and Amir Yazdanbakhsh. What makes chain-of-thought prompting effective? a counterfactual study. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1448–1535, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.101. URL https://aclanthology.org/2023.findings-emnlp.101/.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759/.

Alex Nguyen and Gautam Reddy. Differential learning kinetics govern the transition from memorization to generalization during in-context learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=INyi7qUdjZ.

Tai Nguyen and Eric Wong. In-context example selection with influences, 2023. URL https://arxiv.org/abs/2302.11042.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 188–197, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1018. URL https://aclanthology.org/D19-1018/.

OpenAI. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/, 2025.

Core Francisco Park, Ekdeep Singh Lubana, and Hidenori Tanaka. Competition dynamics shape algorithmic phases of in-context learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=XgH1wfHSX8.

Kha Pham, Hung Le, Man Ngo, and Truyen Tran. Rapid selection and ordering of in-context demonstrations via prompt embedding clustering. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=1Iu2Yte5N6.

Chengwei Qin, Aston Zhang, Chen Chen, Anirudh Dagar, and Wenming Ye. In-context learning with iterative demonstration selection. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7441–7455, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.438. URL https://aclanthology.org/2024.findings-emnlp.438/.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL `https://arxiv.org/abs/2412.15115`.

Allan Raventos, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=BtAz4a5xDg`.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, April 2009. ISSN 1554-0669. doi: 10.1561/1500000019. URL `https://doi.org/10.1561/1500000019`.

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 (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 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/`.

Yikang Shen, Shawn Tan, Alessandro Sordoni, and Aaron Courville. Ordered neurons: Integrating tree structures into recurrent neural networks. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=B1l6qiR5F7`.

Amit Singhal and I. Google. Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin*, 24, 01 2001.

Yi Su, Yunpeng Tai, Yixin Ji, Juntao Li, Yan Bowen, and Min Zhang. Demonstration augmentation for zero-shot in-context learning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14232–14244, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.846. URL `https://aclanthology.org/2024.findings-acl.846/`.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pp. 3104–3112, Cambridge, MA, USA, 2014. MIT Press.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation and synthesis: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 930–957, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.54. URL `https://aclanthology.org/2024.emnlp-main.54/`.

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. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Tu Vu, Heming Liu, David Dohan, and Denny Zhou. Active learning principles for in-context learning with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 4891–4901. Association for Computational Linguistics, 2023.

Xingchen Wan, Han Zhou, Ruoxi Sun, and Sercan O Arik. From few to many: Self-improving many-shot reasoners through iterative optimization and generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=JBXO05r4AV.

Dingzirui Wang, Longxu Dou, Xuanliang Zhang, Qingfu Zhu, and Wanxiang Che. Improving demonstration diversity by human-free fusing for text-to-SQL. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1193–1207, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.65. URL https://aclanthology.org/2024.findings-emnlp.65/.

Dingzirui Wang, Xuanliang Zhang, Qiguang Chen, Longxu Dou, Xiao Xu, Rongyu Cao, YINGWEI MA, Qingfu Zhu, Wanxiang Che, Binhua Li, Fei Huang, and Yongbin Li. In-context transfer learning: Demonstration synthesis by transferring similar tasks, 2025a. URL https://openreview.net/forum?id=ptTt8mhS7n.

Dingzirui Wang, Xuanliang Zhang, Keyan Xu, Qingfu Zhu, Wanxiang Che, and Yang Deng. V-synthesis: Task-agnostic synthesis of consistent and diverse in-context demonstrations from scratch via v-entropy, 2025b. URL https://arxiv.org/abs/2506.23149.

George Wang, Matthew Farrugia-Roberts, Jesse Hoogland, Liam Carroll, Susan Wei, and Daniel Murfet. Loss landscape geometry reveals stagewise development of transformers. In *High-dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning*, 2024b. URL https://openreview.net/forum?id=2JabyZjM5H.

Guorong Wang, Yimin Wei, and Sanzheng Qiao. *Generalized Inverses: Theory and Computations*, volume 53 of *Developments in Mathematics*. Springer, Singapore, 2018. ISBN 978-981-13-0145-2 (Print), 978-981-13-0146-9 (eBook). doi: 10.1007/978-981-13-0146-9.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Label words are anchors: An information flow perspective for understanding in-context learning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023a. URL https://openreview.net/forum?id=OkQD6RMUK5.

Liang Wang, Nan Yang, and Furu Wei. Learning to retrieve in-context examples for large language models. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1752–1767, St. Julian's, Malta, March 2024c. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.105/.

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*, 2023b. URL https://openreview.net/forum?id=BGvkwZEGt7.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024d. URL https://openreview.net/forum?id=y10DM6R2r3.

Yanzheng Xiang, Hanqi Yan, Lin Gui, and Yulan He. Addressing order sensitivity of in-context demonstration examples in causal language models. In Lun-Wei Ku, Andre Martins, and

Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6467–6481, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.386. URL https://aclanthology.org/2024.findings-acl.386/.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. Quick and (not so) dirty: Unsupervised selection of justification sentences for multi-hop question answering. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2578–2589, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1260. URL https://aclanthology.org/D19-1260.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Tong Yang, Yu Huang, Yingbin Liang, and Yuejie Chi. In-context learning with representations: Contextual generalization of trained transformers. In *ICML 2024 Workshop on Theoretical Foundations of Foundation Models*, 2024. URL https://openreview.net/forum?id=fShWHkLX3o.

Zhao Yang, Yuanzhe Zhang, Dianbo Sui, Cao Liu, Jun Zhao, and Kang Liu. Representative demonstration selection for in-context learning with two-stage determinantal point process. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5443–5456, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.331. URL https://aclanthology.org/2023.emnlp-main.331/.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STar: Bootstrapping reasoning with reasoning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_3ELRdg2sgI.

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1684–1692. PMLR, 03–05 May 2025.

Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Is in-context learning sufficient for instruction following in LLMs? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=STEEDDv3zI.

Jiachen ZHAO, Zonghai Yao, zhichao Yang, and hong yu. SELF-EXPLAIN: Teaching large language models to reason complex questions by themselves. In *R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023. URL https://openreview.net/forum?id=nN8pCTVQZD.

Tianshi Zheng, Yixiang Chen, Chengxi Li, Chunyang Li, Qing Zong, Haochen Shi, Baixuan Xu, Yangqiu Song, Ginny Y. Wong, and Simon See. The curse of cot: On the limitations of chain-of-thought in in-context learning, 2025. URL https://arxiv.org/abs/2504.05081.

Yuxiang Zhou, Jiazheng Li, Yanzheng Xiang, Hanqi Yan, Lin Gui, and Yulan He. The mystery of in-context learning: A comprehensive survey on interpretation and analysis. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14365–14378, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.795. URL https://aclanthology.org/2024.emnlp-main.795/.

A Limitations and Ethics

A.1 Limitations

(i) The current experimental datasets and models are limited, where future work will validate LCS on a broader range of models and datasets. *(ii)* Although we discuss that contextual alignment and output calibration capabilities are key factors influencing the ICL effectiveness, the underlying factors that affect these two capabilities warrant further investigation.

A.2 Ethics

All datasets and models used in this paper are publicly available, and our usage follows their licenses and terms. We employ AI tools for coding and writing polishing.

B LLM Usage

We have employed the AI tool for coding and writing polishing.

C Prove

C.1 Equation 2

Proof. Suppose $X = (x_1, ..., x_{|X|})$, where x_i is the token of X, we can derive that:

$$\begin{aligned}
\mathbb{L}_p(X|K;D;Q) &= -\log p(X|K;D;Q) \\
&= \sum_{t=0}^{T} (-\log p(x_t|D;Q;x_{1:t-1})) \\
&= \sum_{t=0}^{T} \left(-\log\left(\frac{p(x_t|Q;x_{1:t-1})p(D|Q;x_{1:t})}{p(D|Q;x_{1:t-1})}\right)\right) \\
&= \mathbb{L}_p(X|Q) - \sum_{t=0}^{T} \left(\log\left(\frac{p(D|Q;x_{1:t})}{p(D|Q;x_{1:t-1})}\right)\right) \\
&= \mathbb{L}_p(X|Q) - (\log p(D|Q;X) - \log p(D|Q))
\end{aligned}$$

□

C.2 Theorem 1

Proof.

$$\begin{aligned}
p(X|Q;D) - p(X|Q) &= \frac{p(X,Q,D)}{p(Q,D)} - \frac{p(X,Q)}{p(Q)} \\
&= \frac{p(X,Q,D)p(Q) - p(X,Q)p(Q,D)}{p(Q,D)p(Q)} \\
&= \frac{p(D|Q,X)p(Q,X)p(Q) - p(X,Q)p(D|Q)p(Q)}{p(Q,D)p(Q)} \\
&= \frac{p(X|Q)}{p(D|Q)}\left(p(D|Q,X) - p(D|Q)\right) \\
&= \frac{p(X|Q)}{p(D|Q)} I(X \to D|Q)
\end{aligned}$$

Therefore, we can conclude that:

$$I(X \to D|Q) = \frac{p(D|Q)}{p(X|Q)} I(D \to X|Q)$$

□

C.3 Error of Theorem 1

Assuming the error of the empirical predictor relative to the true predictor is $\hat{p}(A|B) = p(A|B) + \varepsilon(A|B)$, where A, B are any random variables. We suppose that $r_p \geq \frac{\varepsilon(D|Q)}{\varepsilon(X|Q)} \geq \frac{\varepsilon(D|Q;X)}{\varepsilon(X|Q;D)}$, i.e., the error growth rate with introduced demonstrations is smaller than that without demonstrations, which is further smaller than the ICL effectiveness. According to Theorem 1, the slope of the fitted line can be approximated as:

$$\frac{I_{\hat{p}}(D|Q;X)}{I_{\hat{p}}(X|Q;D)} = \frac{(p(D|Q;X) - p(D|Q)) + (\varepsilon(D|Q;X) - \varepsilon(D|Q))}{(p(X|Q;D) - p(X|Q)) + (\varepsilon(X|Q;D) - \varepsilon(X|Q))}$$

Direct computation yields:

$$\frac{\hat{p}(D|Q)}{\hat{p}(X|Q)} = \frac{p(D|Q) + \varepsilon(D|Q)}{p(X|Q) + \varepsilon(X|Q)}$$

Thus, we have:

$$\Delta_I := \frac{I_{\hat{p}}(D|Q;X)}{I_{\hat{p}}(X|Q;D)} - \frac{I_p(D|Q;X)}{I_p(X|Q;D)} = \frac{\varepsilon(D|Q;X) - \varepsilon(X|Q;D)r_p}{I_p(X|D;Q)\,(I_p(X|D;Q) + \varepsilon(X|Q;D))}$$

$$\Delta_p := \frac{\hat{p}(D|Q)}{\hat{p}(X|Q)} - \frac{p(D|Q)}{p(X|Q)} = \frac{\varepsilon(D|Q;X) - \varepsilon(X|Q;D)r_p}{p(X|D;Q)\,(p(X|D;Q) + \varepsilon(X|Q;D))}$$

Assuming $I_p(X|D;Q) \leq p(X|D;Q)$, i.e., the information the model learns about D from X is less than the information inherently contained in the model, we have:

$$\Delta_I \leq \Delta_p$$

This implies that using the slope as a metric for ICL effectiveness has a smaller error compared to using $\frac{\hat{p}(D|Q)}{\hat{p}(X|Q)}$.

C.4 Theorem 2

Proof. Since $\hat{X} = \arg\max_{X\sim\mathcal{X}} \hat{p}(X|Q)$, we can conclude that $\hat{p}(\hat{X}|Q) \geq \hat{p}(X^*|Q)$. Based on the total probability theorem, we can draw that:

$$\hat{p}(\hat{D}|Q) = \sum_{X\sim\mathcal{X}} \hat{p}(\hat{D}|Q;X)\hat{p}(X)$$

$$\hat{p}(D^*|Q) = \sum_{X\sim\mathcal{X}} \hat{p}(D^*|Q;X)\hat{p}(X)$$

Considering that $\hat{p}(\hat{D}|Q;X) \leq \hat{p}(D^*|Q;X), \forall X \sim \mathcal{X}, Q \sim \mathcal{Q}$, it can be concluded that $\hat{p}(\hat{D}|Q) \leq \hat{p}(D^*|Q)$. Therefore, we can draw the conclusion that:

$$\frac{\hat{p}(\hat{D}|Q)}{\hat{p}(\hat{X}|Q)} \leq \frac{\hat{p}(D^*|Q)}{\hat{p}(X^*|Q)}$$

□

D Additional Information

D.1 Detail of Benchmarks

In this section, we discuss the datasets we used in this paper in detail. The scale of the test set and the demonstrations of each dataset are shown in Table 4.

Table 4: The scales of test set and demonstrations of each dataset.

Dataset	Test Set	Demonstration
GSM8K	1319	7473
MATH	500	7496
HumanEval	164	596
MBPP	378	596
ARC-Challenge	1172	1119
MMLU-Pro	1000	70
FinQA	1147	6251
Amazon Review	200	1800

GSM8K GSM8K (Cobbe et al., 2021) is a high-quality dataset consisting of grade school level math problems. We directly use the training set as the demonstration pool.

MATH MATH (Hendrycks et al., 2021) is a dataset of high school competition-level math problems covering various domains, such as algebra, probability, and geometry. Following (Lightman et al., 2024), we use a sampled subset of 500 examples for evaluation. We use the training set as the demonstration pool.

HumanEval HumanEval (Chen et al., 2021a) is a Python-based code generation benchmark. We follow the evaluation protocol of (Liu et al., 2023). Since the dataset does not provide a labeled training set, we use demonstrations from MBPP as the demonstration pool.

MBPP MBPP (Austin et al., 2021) is another Python-based code generation benchmark. Compared to HumanEval, it is larger in scale and includes a split between validation and test sets. In this paper, we adapt the evaluation on the test set and use the remaining data as the demonstration pool, following the evaluation protocol of (Liu et al., 2023).

ARC-Challenge ARC-Challenge (Yadav et al., 2019) is a difficult question-answering dataset focusing on scientific knowledge. We directly use the training set as the demonstration pool.

MMLU-Pro MMLU-Pro (Wang et al., 2024d) is a multitask benchmark designed to comprehensively evaluate LLMs on professional domain knowledge and complex reasoning capabilities. As the dataset only provides validation and test sets, we use the validation set as the demonstration pool and evaluate on the test set.

FinQA FinQA (Chen et al., 2021b) is a question-answering dataset in the financial domain. It requires models to perform numerical reasoning and calculations based on given financial tables and textual information. We use the training set as the demonstration pool.

Amazon Review The Amazon Review (Ni et al., 2019) dataset consists of numerous user ratings and textual reviews on products from the Amazon platform, and it is widely used in sentiment analysis and recommendation system research. Due to the large scale of the dataset, we select the *Health and Personal Care* category as the test set and use *All Beauty*, *Digital Music*, and *Software* as the demonstration pool.

D.2 CALCULATION OF LCS

In this section, we present how to calculate LCS, which primarily involves two sequential steps: reasoning process paraphrasing and likelihood calculation. The prompts employed for these computations are detailed in Appendix D.3.

The reasoning process paraphrasing step requires models to restructure human-labeled reasoning processes according to their preferred reasoning style when provided with a given reasoning process. This adaptation is crucial because discrepancies between human-labeled reasoning formats and model-preferred reasoning patterns (e.g., "<think >" tag of Llama-R1 (DeepSeek-AI et al., 2025)) could lead to inflated information gain measurements that reflect stylistic variations rather than knowledge acquisition. To mitigate this confounding factor, we implement reasoning process paraphrasing to eliminate format-induced biases, thereby ensuring that computational results authentically reflect knowledge-derived information learned from demonstrations. Specifically, for each data instance and demonstration, we input the question, answer, and human-labeled reasoning process (if provided), instructing the model to rephrase the output using its preferred reasoning style.

Following the paraphrasing, we calculate the likelihood with paraphrased results. For conditional probabilities expressed as $\hat{p}(A|B)$, we treat B as user input and A as model output, encapsulating them into a formatted string using the model chat template. This composite string is then processed through the model to obtain token-level likelihoods. The joint likelihood of a sequence A is computed by multiplying the probabilities of all constituent tokens. To minimize the confounding effects of sequence length on probability comparisons, we apply length normalization to all computed likelihood values (Dai et al., 2019). This standardized approach ensures a fair comparison across outputs of varying lengths while preserving the probabilistic relationships between different reasoning processes.

D.3 PROMPTS

In this section, we introduce the prompts used in this paper. The reasoning prompts of §3 can be seen in (Chen et al., 2023b; Grattafiori et al., 2024; DeepSeek-AI et al., 2025). The prompts used for the paraphrasing and the synthesis are shown in Table 5 and Table 6.

Table 5: The prompt of the paraphrase.

Prompt of Paraphrasing
<Begin of Task Definition > {definition} <End of Task Definition > <Begin of Input > {question} <End of Input > <Begin of Hint > {hint} <End of Hint > <Begin of Answer > {answer} <End of Answer > Considering the above task definition, generate the reasoning process of the given input and answer with the hint (could be empty).

Table 6: The prompt of the demonstration synthesis.

Prompt of Synthesis
``md {task_definition} ``` Given Question: {question} Based on the above task definition and the given question, synthesize a question and the corresponding answer that is similar to the given question of the task.

E Additional Discussion

E.1 The Factors that Affect ICL Effectiveness

Following the discussion of §2.2, in this section, we delve deeper into the factors that influence the ICL effectiveness, specifically the meaning of $\hat{p}(D|Q)$ and $\hat{p}(X|Q)$. Our primary focus is on different predictors $\hat{p}_1$ and $\hat{p}_2$ applied to the same data, assuming that the answer $X = \arg\max_{X \in \mathcal{X}} p(X|Q)$ is the correct answer, and the demonstration $D = \arg\max_{D \in \mathcal{D}} p(D|Q)$ is the most relevant demonstration to the question Q.

Contextual Alignment Capability $\hat{p}(D|Q)$ If $\hat{p}_1(D|Q) \geq \hat{p}_2(D|Q)$, it indicates that $\hat{p}_1$ has a stronger ability to judge the relevance of demonstrations to the question compared to $\hat{p}_2$, showing that $\hat{p}_1$ is a better demonstration selector. From the perspective of demonstrations, this means that $\hat{p}_1$ is better at understanding the information in the demonstration and determining its relationship with the user question Q, reflecting that $\hat{p}_1$ has a stronger ICL ability than $\hat{p}_2$.

It is worth noting that while both $\hat{p}(D|Q)$ and $I(D \rightarrow X|Q)$ measure the consistency between the demonstration and the user input to some extent, their fundamental perspectives differ. $I(D \rightarrow X|Q)$ primarily focuses on the data perspective, measuring the relevance between the input and the demonstration under the assumption that the model is an oracle. In contrast, $\hat{p}(D|Q)$ primarily focuses on the model perspective, observing whether the model has the capability to gauge the relevance between the input and the demonstration, assuming that the demonstration is highly relevant to the user input.

Output Calibration Capability $\hat{p}(X|Q)$ If $\hat{p}_1(X|Q) \geq \hat{p}_2(X|Q)$, it implies that $\hat{p}_1$ has a stronger ability to judge the correct answer compared to $\hat{p}_2$, meaning that $\hat{p}_1$ is a better answer scorer. It should be noticed that $\hat{p}(X|Q)$ does not directly reflect the model ability to solve the given question. This is because the model generates answers using greedy decoding, which means the generated answer could not be the answer with the highest likelihood. Rather, $\hat{p}(X|Q)$ represents the score the model assigns to a given answer, reflecting the model ability to assess the consistency between the answer and the question.

E.2 Empirical Threshold of LCS

Specifically, based on Figure 2, we can use LCS $= 0.2$ as an empirical threshold, since when LCS ≤ 0.2, the corresponding performance gain is minimal or negative, suggesting that ICL is less effective in the given task and model. This threshold is largely empirical. In practice, users can adjust the sensitivity to ICL effectiveness according to their preferences.

E.3 Efficiency of LCS Calculation

LCS requires calculating four related likelihoods for each data point. Therefore, if we assume the time cost for a model to run one pass of ICL inference on a given dataset is T, the time cost of our method is $4T$. Although our time cost is greater than that of a single inference pass, our primary motivation is to propose an effective method for measuring ICL effectiveness to guide subsequent demonstration annotation and ICL usage, rather than to perform efficient inference. Therefore, we consider the additional time cost to be acceptable.

E.4 Why Our Analysis is Based on Conditional Probability

We acknowledge that a growing body of work has demonstrated that the order of demonstrations in the prompt can affect the outputs of current autoregressive LLMs (Lu et al., 2022; Guo et al., 2024). In this paper, however, we deliberately work with an idealized order-invariant model in which the relevant conditionals are unaffected by permutations of the demonstrations. Prior research suggests that the observed order sensitivity is a limitation of existing models and training procedures, and that an ideal language model should, in principle, be invariant to the ordering of demonstrations (Xiang et al., 2024). Furthermore, several studies indicate that the influence of text order on in-context learning is diminishing as models and training improve (Pham et al., 2025), and many influential analyses of ICL adopt a similar order-invariance assumption when deriving related decompositions (Zhang et al., 2025; Jesson et al., 2025). Under this widely used idealization, Equation 2 is well-defined and yields a decomposition of the loss into a "zero-shot" term and a "demo-dependent" term. We explicitly adopt this assumption in our derivation in Appendix C.1 to isolate the fundamental nature of ICL effectiveness and to obtain a clean measure of it. Our experimental results are consistent with the theoretical predictions obtained under this assumption, supporting the practical reasonableness of applying Equation 2 even when instantiated with standard autoregressive LLMs.

In this work, we define the predictive probability for an output sequence $\mathbf{X} = (x_1, \ldots, x_T)$ as the product of token-wise conditional probabilities:

$$P_\theta(\mathbf{X} \mid \mathbf{Q}) = \prod_{t=1}^{T} P_\theta(x_t \mid x_{<t}, \mathbf{Q}) \,.$$

This construction is widely used in the NLP research (Bengio et al., 2003; Sutskever et al., 2014), which follows directly from the chain rule of probability: the joint distribution of any discrete sequence can be decomposed, without loss of generality, into a left-to-right product of conditional distributions. Hence, as long as the model provides reasonable estimates of each conditional distribution $P_\theta(y_t \mid y_{<t}, \mathbf{x})$, the above product corresponds to the likelihood of the sequence under the model and can be naturally interpreted as its predictive probability, or a joint score.

E.5 Why Removing Logarithm in Learning Gain

In this section, we discuss why, when defining the learning gain $I_p(X \to D \mid Q)$, we remove the log based on Equation 2. This is because, by using the difference in probabilities, we can better analyze its relationship with other factors. We consider this a reasonable transformation, since the transformed expression still reflects the change in loss, which is consistent with our motivation of measuring the effectiveness of ICL via loss change. Specifically, let $k = \log p(D \mid Q; X) - \log p(D \mid Q)$, i.e., the decrease in loss brought by introducing the demonstration. Then we have:

$$p(D \mid Q; X) = e^k p(D \mid Q)$$
$$\Rightarrow p(D \mid Q; X) - p(D \mid Q) = (e^k - 1)p(D \mid Q)$$

Since $p(D \mid Q) > 0$, $p(D \mid Q; X) - p(D \mid Q)$ is positively correlated with $\log p(D \mid Q; X) - \log p(D \mid Q)$. That is, $I_p(X \to D \mid Q)$ is positively correlated with the change in loss.

F Additional Experiments

F.1 Main Experiment Results

Table 7: The performance of different models on different datasets using 0-shot and 1-shot. Δ is the performance change of 1-shot compared with 0-shot. HumanE denotes HumanEval, A-C denotes ARC-Challenge, M-P denotes MMLU-Pro, and Amazon denotes Amazon Review.

Model	Shot	Math		Code		Reason		Domain	
		GSM8K	MATH	HumanE	MBPP	A-C	M-P	FinQA	Amazon
Llama2-7b	0	12.7	5.0	14.0	23.0	34.6	14.1	10.5	28.5
	1	27.7	14.6	13.4	22.5	46.2	19.6	17.8	29.0
	Δ	+10.0	+9.6	−0.6	−0.5	+11.6	+5.5	+7.3	+0.5
Llama3.1-8b	0	86.4	48.4	65.9	54.8	82.1	50.4	49.7	63.5
	1	84.2	50.8	63.4	55.6	80.2	53.0	54.6	68.5
	Δ	−1.8	+2.4	−2.5	+0.8	−1.9	+2.6	+4.9	+5.0
Llama-R1-8b	0	86.1	75.4	70.7	67.2	84.8	58.2	45.2	53.5
	1	80.1	74.2	67.7	60.8	84.5	52.7	43.4	65.0
	Δ	−6.0	−1.2	−3.0	−6.4	−0.3	−5.5	−1.8	+11.5

Overall Performance In this part, we present the performance of 0-shot and 1-shot on different models and datasets, as shown in Table 7.

Figurative Illustrations of Learn-to-Context Slope In this section, we present the variation of $I_{\hat{p}}(X \rightarrow D|Q)$ with respect to $I_{\hat{p}}(D \rightarrow X|Q)$ under different settings, as illustrated in Figure 7, Figure 8, and Figure 9. Considering that the number of data points could vary slightly across different models due to the potential for excessively long responses from certain models (e.g., Llama-R1-8b could persist in "thinking").

F.2 Different Similarity Measurement

In this section, we discuss the impact of replacing the contextual relevance $I_{\hat{p}}(D \rightarrow X|Q)$ with other metrics. We conduct experiments on Llama3.1-8b using the GSM8K, MATH, and Amazon Review dataset, where we replace the similarity measure with n-gram (Broder et al., 1997), BM25 (Robertson & Zaragoza, 2009), and cosine similarity (Singhal & Google, 2001) to evaluate the similarity between the provided demonstration and user input. The experimental results are shown in Figure 10, Figure 11, abd Figure 12. From the figure, we observe the following: *(i)* For effective similarity measures (e.g., BM25, cosine similarity), the observed ICL effectiveness is consistent with using the contextual relevance; *(ii)* However, for metrics with poorer performance (e.g., n-gram), the ICL effectiveness is not accurately reflected, demonstrating that n-gram fails to properly capture the similarity between demonstrations and user inputs.

F.3 Different Model

Table 8: Performance and fitted lines across different models and datasets. ARC-C denotes ARC-Challenge, and Amazon denotes Amazon Review. Δ denotes the performance change of 1-shot relative to 0-shot, where performance gains < 1.0 are marked in red. $r_{\hat{p}}x + b$ represents the fitted line with $I_{\hat{p}}(X \rightarrow D|Q)$ as the x-axis and $I_{\hat{p}}(D \rightarrow X|Q)$ as the y-axis, where $r_{\hat{p}}$ values < 0.2 are highlighted in red.

Model	**MATH**		**FinQA**		**Amazon**	
	Δ	$r_{\hat{p}}x + b$	Δ	$r_{\hat{p}}x + b$	Δ	$r_{\hat{p}}x + b$
Llama3.1-8b	$+2.4$	$0.34x - 0.00$	$+4.9$	$0.82x - 0.06$	$+5.0$	$0.94x - 0.09$
Llama3.1-70b	-3.6	$-0.13x - 0.04$	$+7.6$	$0.77x - 0.07$	$+16.0$	$0.79x - 0.18$
Qwen2.5-7b	$+2.2$	$0.81x - 0.16$	$+4.9$	$0.29x - 0.09$	$+27.0$	$0.42x + 0.00$
Qwen3-8b	-1.4	$-0.21x - 0.08$	$+6.7$	$0.53x + 0.21$	-1.0	$0.03x - 0.13$
Ministral-8b	-1.8	$0.04x - 0.02$	$+4.8$	$0.71x - 0.04$	$+4.0$	$0.69x - 0.11$

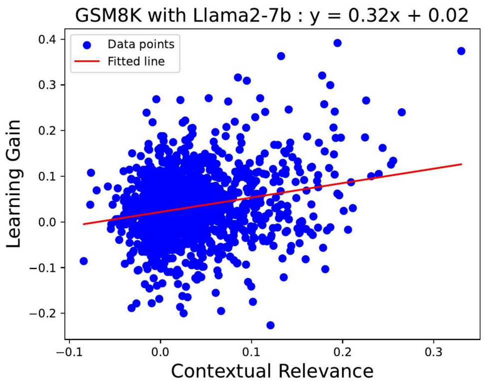
To evaluate the effectiveness of LCS on the models with different scales and series, we adapt the experiments on Qwen2.5-7b (Qwen et al., 2025), Qwen3-8b (Yang et al., 2025), Ministral-8B-Instruct-2410 (Ministral-8b) (Jiang et al., 2023), and Llama3.1-70b (Grattafiori et al., 2024). The experimental results are shown in Table 8. It can be seen that LCS still reflects the ICL effectiveness on the models with different scales and series, proving the generalization of our metric.

F.4 Intercept under Different Model Scales

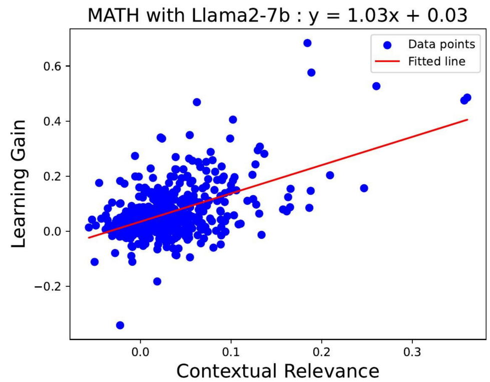
To more thoroughly compare the differences in the effective information learned from demonstrations by LLMs of varying capabilities, we conduct experiments on LLMs of different scales within the same series. The experimental results are shown in Figure 13. From the figure, it can be observed that the intercept of Llama3.1-70b is generally smaller than that of Llama3.1-8b, as discussed in Section §3.3.2, indicating that Llama3.1-70b learns less effective information.

F.5 Performance of LCS with Mismatch Label

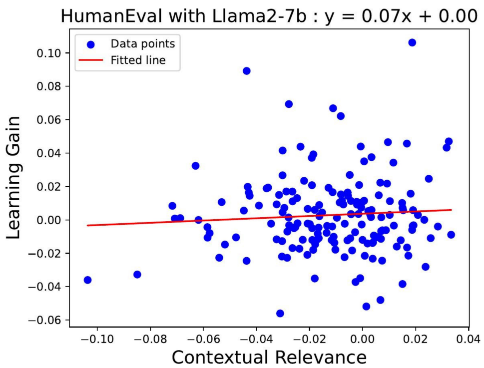
In this section, we investigate the effectiveness of LCS under adversarial labels. We mainly conduct experiments on the MATH and MMLU-Pro datasets. For MATH, following Madaan et al. (2023), we replace the values in the examples with placeholders. For MMLU-Pro, we randomly replace the choice corresponding to each question with another choice. The experimental results are shown in Table 9. From the table, we can see that when using adversarial labels, LCS can still faithfully reflect the effectiveness of ICL, which is consistent with prior work that adversarial labels can also lead to



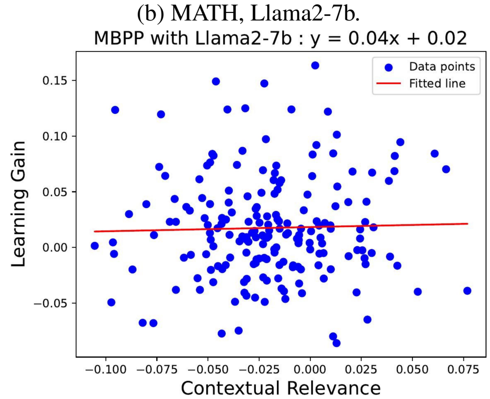
(a) GSM8K, Llama2-7b.



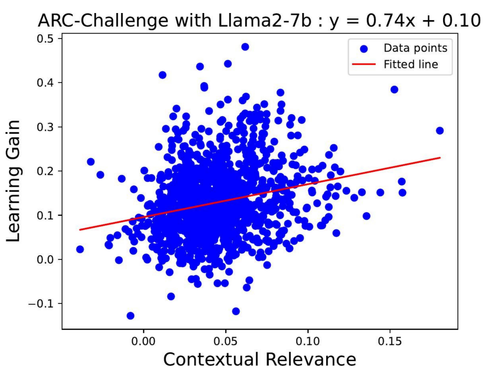
(b) MATH, Llama2-7b.



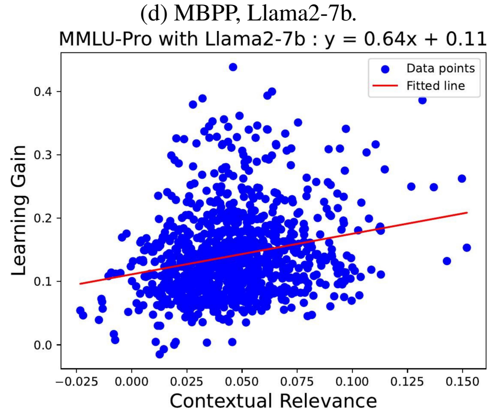
(c) HumanEval, Llama2-7b.



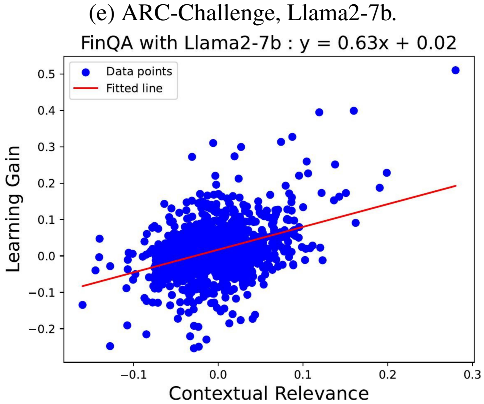
(d) MBPP, Llama2-7b.



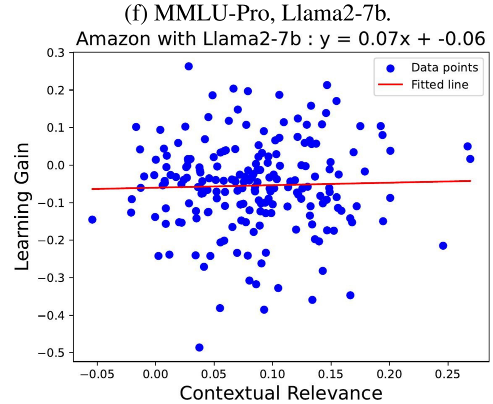
(e) ARC-Challenge, Llama2-7b.



(f) MMLU-Pro, Llama2-7b.

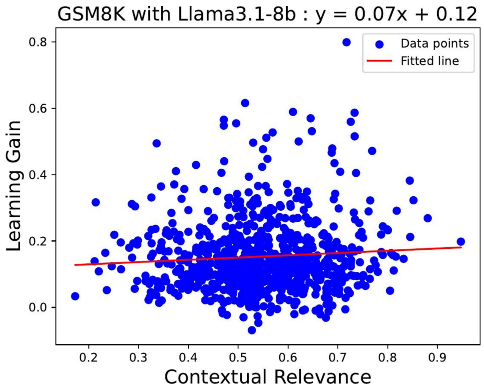


(g) FinQA, Llama2-7b.

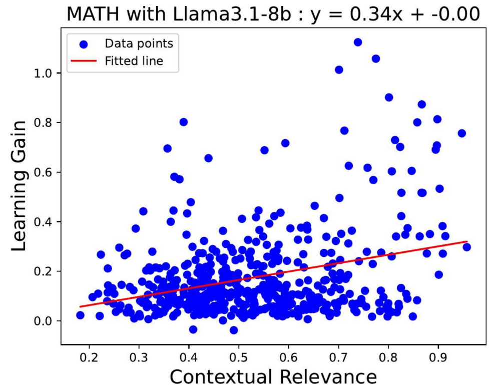


(h) Amazon, Llama2-7b.

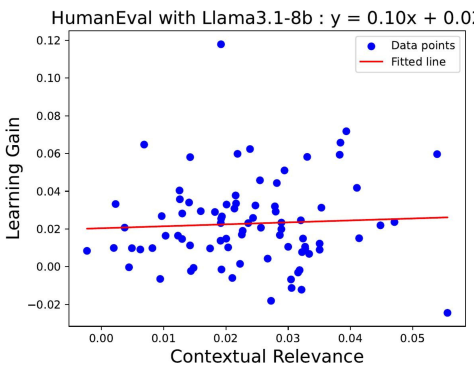
Figure 7: The variation of $I_{\hat{p}}(X \rightarrow D|Q)$ (y-axis) with $I_{\hat{p}}(D \rightarrow X|Q)$ (x-axis) on different datasets using Llama2-7b. The title of each plot displays the corresponding dataset and fitted line. Each blue dot in the plot represents a data point, and the red line indicates the fitted line of the data points.



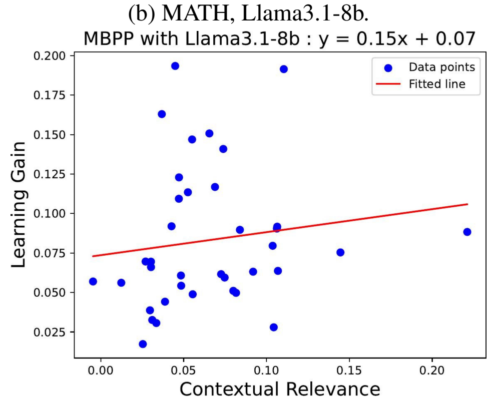
(a) GSM8K, Llama3.1-8b.



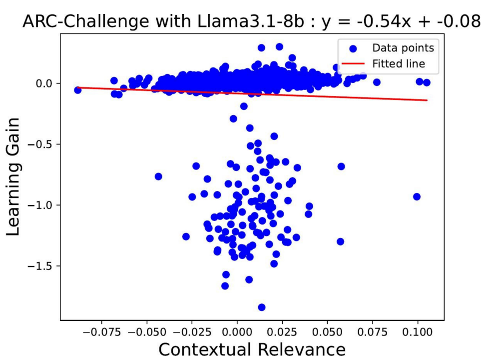
(b) MATH, Llama3.1-8b.



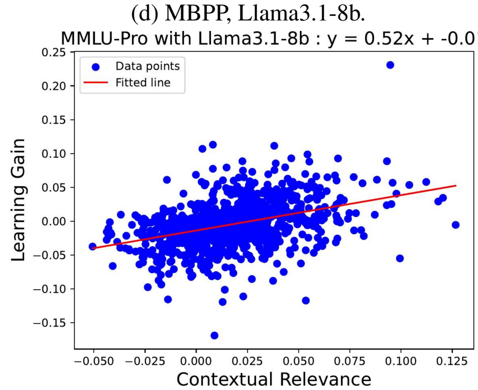
(c) HumanEval, Llama3.1-8b.



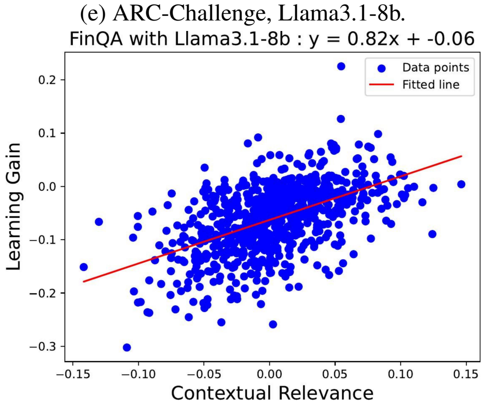
(d) MBPP, Llama3.1-8b.



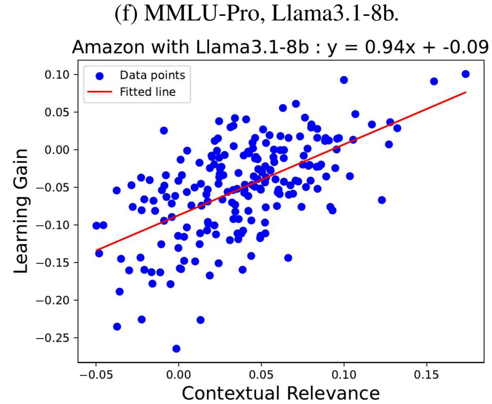
(e) ARC-Challenge, Llama3.1-8b.



(f) MMLU-Pro, Llama3.1-8b.

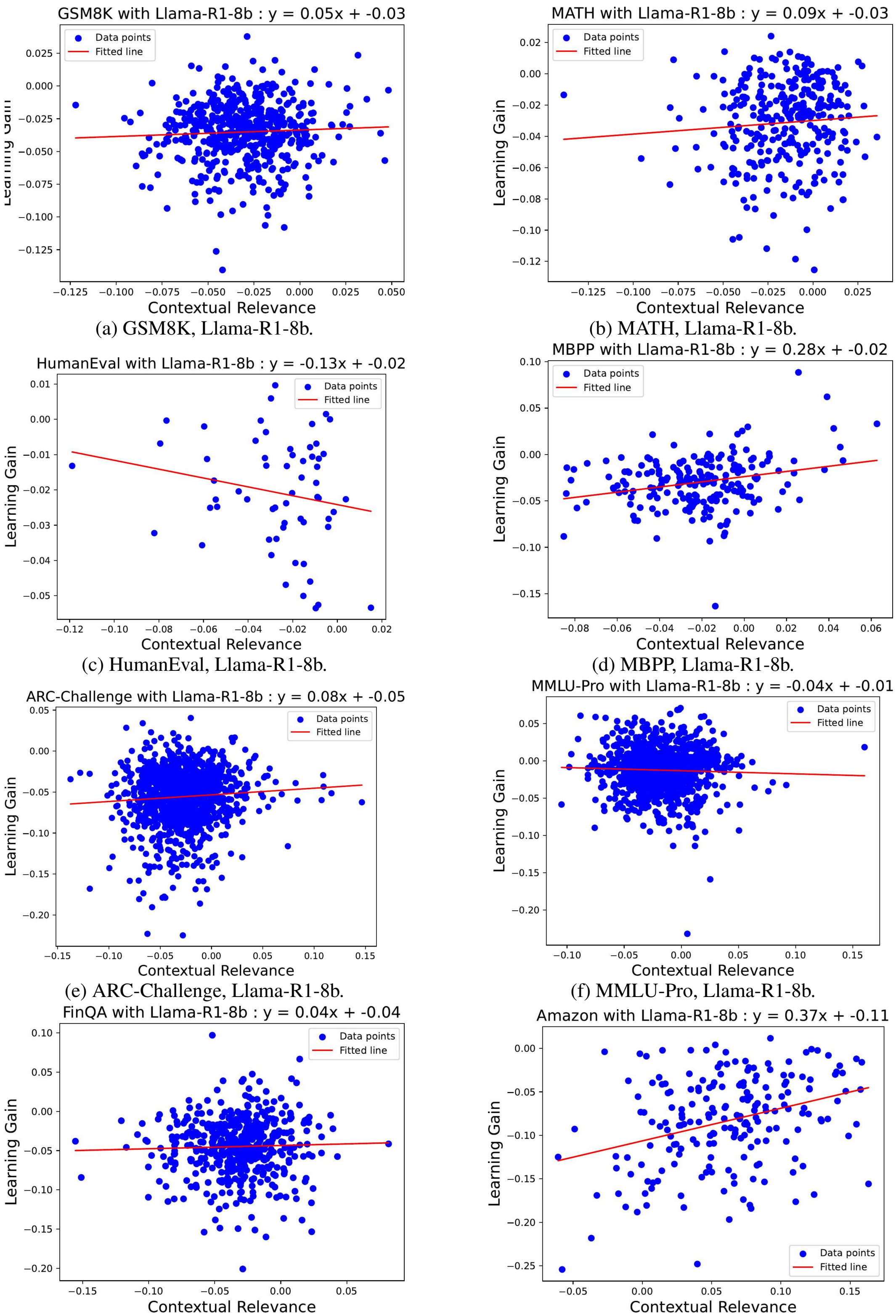


(g) FinQA, Llama3.1-8b.



(h) Amazon, Llama3.1-8b.

Figure 8: The variation of $I_{\hat{p}}(X \to D|Q)$ (y-axis) with $I_{\hat{p}}(D \to X|Q)$ (x-axis) on different datasets using Llama3.1-8b. The legend is the same as Figure 7.



(a) GSM8K, Llama-R1-8b.

(b) MATH, Llama-R1-8b.

(c) HumanEval, Llama-R1-8b.

(d) MBPP, Llama-R1-8b.

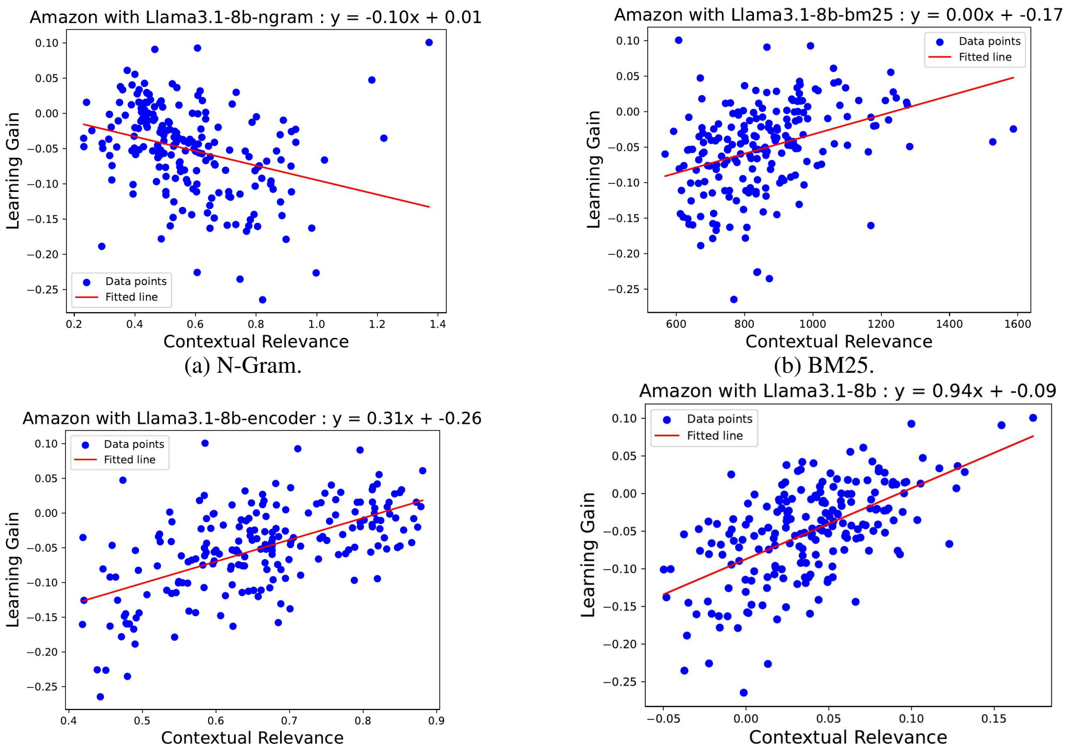
(e) ARC-Challenge, Llama-R1-8b.

(f) MMLU-Pro, Llama-R1-8b.

(g) FinQA, Llama-R1-8b.

(h) Amazon, Llama-R1-8b.

Figure 9: The variation of $I_{\hat{p}}(X \to D|Q)$ (y-axis) with $I_{\hat{p}}(D \to X|Q)$ (x-axis) on different datasets using Llama-R1-8b. The legend is the same as Figure 7.



(a) N-Gram. (b) BM25. (c) Cosine Similarity. (d) Contextual Relevance.

Figure 10: The variation of the learning gain (y-axis) with different similarity metrics (x-axis) on Amazon using Llama3.1-8b.

Dataset	**Type**	**Llama2-7b** Δ	LCS	**Llama3.1-8b** Δ	LCS	**Llama-R1-8b** Δ	LCS
MATH	Real	+9.6	1.03	+2.4	0.34	−1.2	0.09
	Mismatch	+4.2	0.72	+2.0	0.22	−1.4	0.03
MMLU-Pro	Real	+5.5	0.64	+2.6	0.52	−5.5	−0.04
	Mismatch	+4.3	0.52	+1.9	0.30	−5.0	−0.12

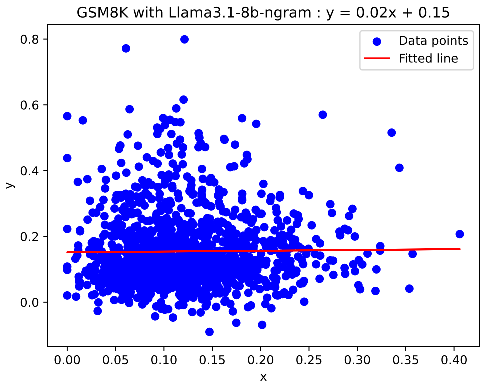
Table 9: The performance with origin and adversarial labels. Real denotes the origin label, and Mismatch denotes the adversarial label.

performance improvements (Madaan et al., 2023). Moreover, compared with original labels, LCS under adversarial labels are relatively lower, which is consistent with the conclusion of Theorem 2, since demonstrations with adversarial labels are of lower quality than those using the original labels.

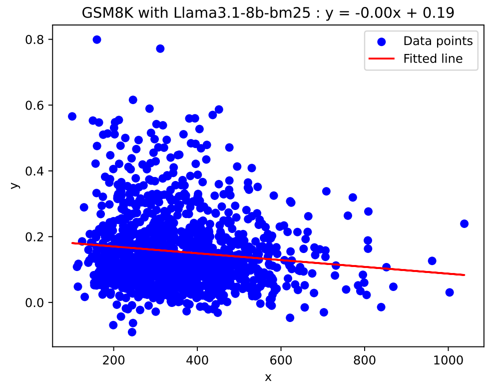
F.6 LCS WITH BLACK-BOX LLMS

Model	**GSM8K** Δ	$r_{\hat{p}}x+b$	**MATH** Δ	$r_{\hat{p}}x+b$	**FinQA** Δ	$r_{\hat{p}}x+b$
`gpt-5-nano`	−1.1	$0.01x+0.27$	−3.9	$-0.12x-0.03$	+7.6	$0.31x+0.05$

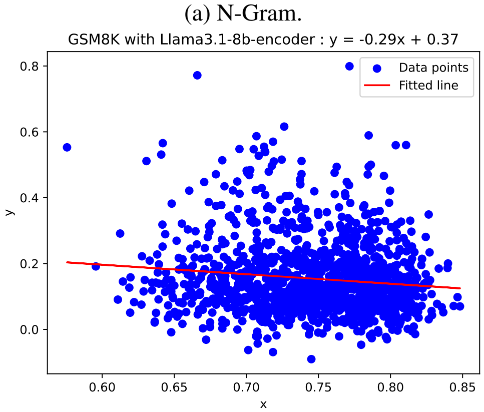
Table 10: LCS with `gpt-5-nano` on GSM8K and MATH. The probability is generated following Kaneko et al. (2025). Considering the API cost, we randomly sample 128 examples from each dataset.



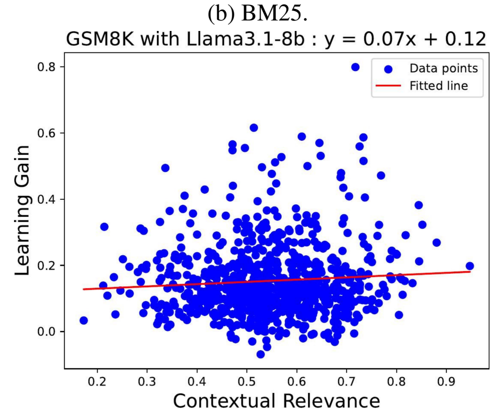
(a) N-Gram.



(b) BM25.



(c) Cosine Similarity.

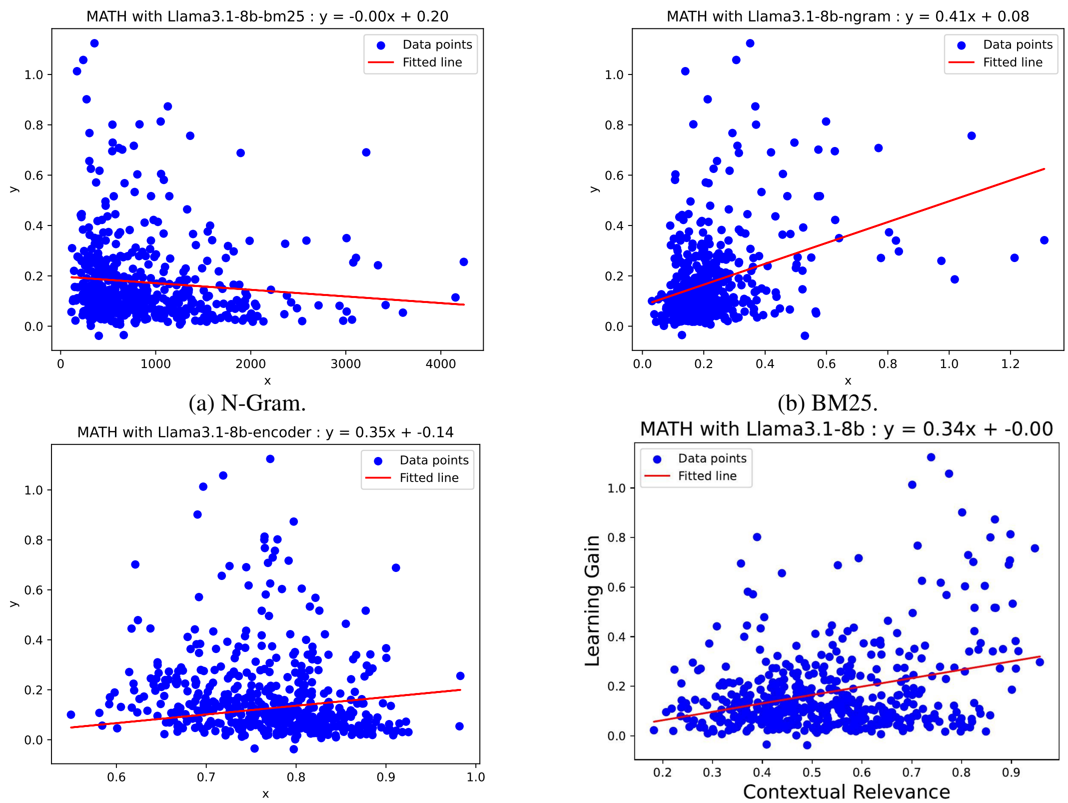


(d) Contextual Relevance.

Figure 11: The variation of the learning gain (y-axis) with different similarity metrics (x-axis) on GSM8K using Llama3.1-8b.

Based on Theorem 1, the main point to calculate LCS with black-box LLMs is how to obtain the logprob. Following Lee et al. (2023); Kaneko et al. (2025), we employ a sampling-based pseudo-likelihood estimator for recovering LLM output distributions from samples. Specifically, for each prompt–response pair, we first feed the prompt into the model and sample multiple outputs. Then, we compute the average ROUGE-N score between these outputs and the response. Kaneko et al. (2025) shows that when the number of samples is sufficiently large, this average ROUGE-N score converges to the real probability. Although the above computation procedure is not very efficient, here we only provide an idea of how to apply LCS to black-box models, and improving the efficiency of computing log probabilities is beyond the scope of this paper.

We conduct experiments using `gpt-5-nano` (OpenAI, 2025). Due to the high API cost, we only randomly sample 128 examples from GSM8K and MATH for our experiments. The experimental results are shown in Table 10, from which we can observe that LCS also accurately reflects the effectiveness of ICL, thereby demonstrating the feasibility of applying LCS to black-box models.



(a) N-Gram. (b) BM25.

(c) Cosine Similarity. (d) Contextual Relevance.

Figure 12: The variation of the learning gain (y-axis) with different similarity metrics (x-axis) on MATH using Llama3.1-8b.

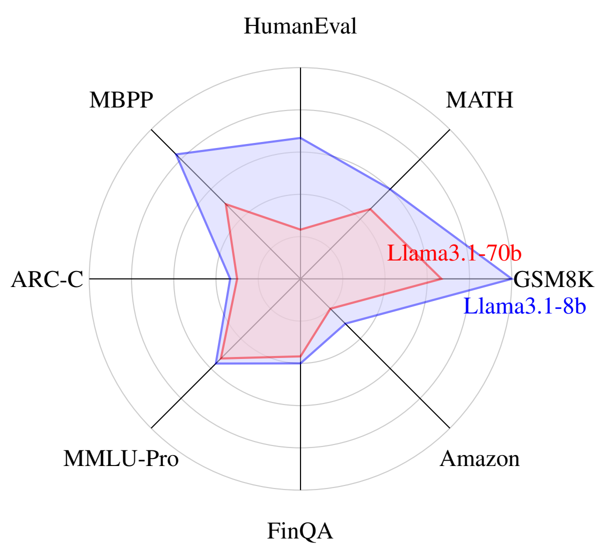


Figure 13: The intercepts of the fitted lines of Llama3.1-8b and Llama3.1-70b.