

Learning-to-Context Slope: Evaluating In-Context Learning Effectiveness Beyond Performance Illusions

Dingzirui Wang¹ Xuanliang Zhang¹ Keyan Xu¹
Qingfu Zhu¹ Wanxiang Che^{1*} Yang Deng²

¹Harbin Institute of Technology ²Singapore Management University
{dzrwang, xuanliangzhang, kyxu, qfzhu, car}@ir.hit.edu.cn
ydeng@smu.edu.sg

Abstract

The effectiveness of in-context learning (ICL) varies significantly across models and tasks, posing challenges for practitioners to determine when ICL reliably improves performance. Current evaluation approaches, reliant on performance changes 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 loss changes even when outputs are incorrect; (ii) its formulation attributes ICL failures to weak contextual alignment or strong output calibration; (iii) it minimizes reliance on labeled data via synthetic evaluation. Experiments show that LCS can better reflect ICL effectiveness when performance change fails, demonstrating its effectiveness.

1 Introduction

In-context learning (ICL) has emerged as a popular and effective paradigm for enhancing large language models (LLMs) 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 existing efforts have been made to unveil the underlying mechanisms of ICL (Zhou et al., 2024; Edelman et al., 2024a; Park et al., 2025) and to explore methods to further boost ICL performance (Wang et al., 2023b; Rubin et al., 2022; Agarwal et al., 2024).

However, as illustrated in Figure 1a, even with models that have strong ICL capability, such as Llama3.1 (Grattafiori et al., 2024), ICL fails to enhance performance and, in some cases, even harms it (DeepSeek-AI et al., 2025; Huang and Wang, 2025; Zheng et al., 2025), indicating varying 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 during inference. However, this approach suffers from two critical limitations. (i) *Low Reliability*: Performance fluctuations may stem from various factors, such as 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 of ICL effectiveness.
- For tasks *without labeled data*, there is no direct way to assess whether adding demonstrations for ICL actually improves outcomes during inference, leaving practitioners without clues for evaluation and improvements.

In light of these challenges, we propose a novel metric, named **Learning-to-Context Slope (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

*Corresponding author.

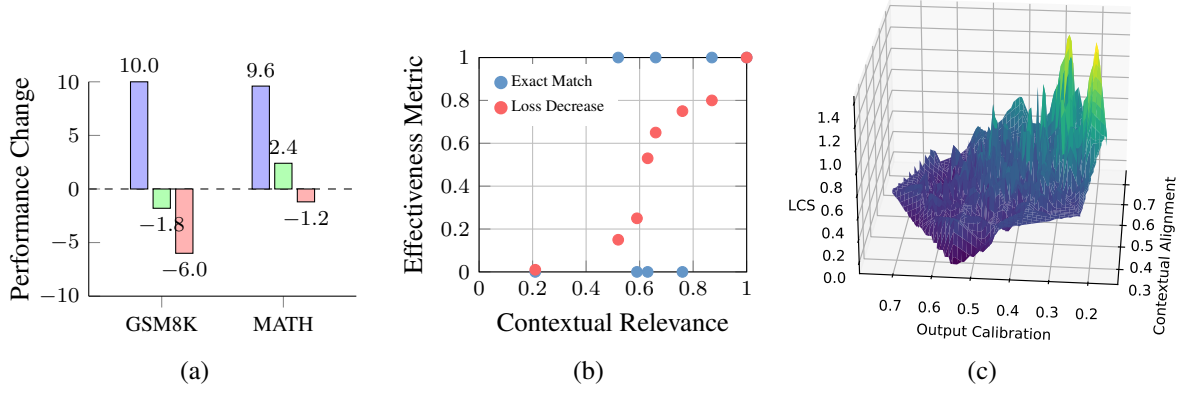


Figure 1: (a) Performance change of different models (From left to right: Llama2-7b, Llama3.1-8b, Llama-R1-8b) before and after applying ICL, where ICL exhibits varying effectiveness across different models on the same dataset. (b) Comparisons between metrics based on **exact match** and **loss decrease**. Each dot denotes an example instance 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 using Llama3.1-8b with MATH, which are two key factors that show positive relation with the ICL effectiveness proved in this paper.

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 in learning gain with demonstration relevance is correspondingly marginal during inference.

Compared to existing 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 a 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 still offers insights into ICL effectiveness.

Our contributions can be summarized as follows:

- We propose a novel metric, namely Learning-to-Context Slope (LCS), to measure ICL effectiveness by capturing the two most important elements in ICL: 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 LCS: Learning-to-Context Slope

We introduce a novel metric named Learning-to-Context Slope (LCS) to measure the ICL effectiveness. First, we interpret the ICL effectiveness by measuring the decrease in loss brought about by using demonstrations based on the Bayesian model (§ 2.1). Then, we present our LCS metric to measure the ICL effectiveness by connecting the loss decrease with the contextual relevance of demonstrations (§ 2.2). Next, we describe how to calculate LCS in practice (§ 2.3). Based on LCS, we discuss two main factors that influence the ICL effectiveness (§ 2.4). Finally, we discuss the relationship between the metric using synthetic data and real data, aiding the application under the data-insufficient scenario (§ 2.5). We discuss why the input order does not affect the conclusion of this paper in § 3.2.5 and Appendix C.2.

2.1 Interpreting ICL Effectiveness via Loss Decrease

Motivated by previous studies (Wang et al., 2024b; Yang et al., 2024), we interpret the ICL effectiveness of a given predictive distribution p with the parameter θ on a specific task $\mathcal{C} = (Q, X, D)$ through the loss decrease brought by demonstrations. Here, Q denotes the user input, X represents the labeled output corresponding to Q , and D denotes the demonstration, which are random variables in the sampling spaces $\mathcal{Q}$, $\mathcal{X}$, and $\mathcal{D}$, respectively. Given demonstrations, the generation loss decrease, *i.e.*, negative log-likelihood, is defined as:

$$\begin{aligned} \Delta \mathbb{L}_\theta(X|Q; D) &= \mathbb{L}_\theta(X|Q) - \mathbb{L}_\theta(X|Q; D) \\ &= \log p(X|Q; D) - \log p(X|Q). \end{aligned} \quad (1)$$

In this paper, the semicolon in $p(A|B; C)$ indicates that both B and C are used as conditions rather than specifying a particular input order in the prompt. We use the same notation for all conditional probabilities involved in the derivation and the empirical calculation in the following.

The effectiveness of ICL is reflected by how much the demonstrations reduce the loss of generating the labeled output:

$$\mathbb{L}_\theta(X|Q; D) = \mathbb{L}_\theta(X|Q) - \Delta \mathbb{L}_\theta(X|Q; D) \quad (2)$$

$\Delta \mathbb{L}_\theta(X|Q; D)$ measures the improvement obtained by providing D over the zero-shot setting.

Based on the Bayesian model (Zhang et al., 2025; Jesson et al., 2025), the loss can be decomposed as:

$$\begin{aligned} \mathbb{L}_\theta(X|Q; D) &= \\ \mathbb{L}_\theta(X|Q) - (\log p(D|Q; X) - \log p(D|Q)), \end{aligned} \quad (3)$$

where $\mathbb{L}_\theta(X|Q)$ represents the loss of zero-shot generation without demonstrations, which is fixed given the model and task. The proof of Equation 3 is presented in Appendix A.1. Equation 3 shows that the demonstration-dependent part of the loss decrease is $\log p(D|Q; X) - \log p(D|Q)$.

It is important to clarify the operational meaning of conditionals such as $p(D|Q)$ and $p(D|Q; X)$. They do not indicate that the model will generate a demonstration during deployment. Instead, they are model-based compatibility scores computed by likelihood scoring: $p(D|Q)$ measures how compatible the demonstration D is with the user input Q , while $p(D|Q; X)$ measures how compatible the same demonstration becomes when the labeled output X is also provided as a condition. Therefore, $\log p(D|Q; X) - \log p(D|Q)$ measures how much the labeled output helps the model identify the given demonstration as relevant to the input. Through Equation 3, this reverse likelihood difference is equivalent to the loss reduction obtained when the demonstration is provided.

2.2 Metric of ICL Effectiveness: LCS

To connect the loss decrease in Equation 3 with the effect of contextual evidence, we define two probability-scale quantities. We emphasize that the following names describe the *operational roles* of these quantities; the arrow notation only indicates which variable is added to the condition and whose likelihood is evaluated, rather than a causal or generation direction.

First, we define the **Learning Gain** as

$$I_p(X \rightarrow D|Q) = p(D|Q; X) - p(D|Q).$$

Although this quantity is expressed through a reverse conditional, it serves as a probability-scale proxy for the demonstration-induced loss reduction: its logarithmic counterpart, $\log p(D|Q; X) - \log p(D|Q)$, is exactly the decrease in the generation loss of X caused by introducing D , as shown in Equation 3. Thus, ‘‘Learning Gain’’ refers to its role as a proxy for loss reduction, rather than to the direction of generation.

Second, we define the **Contextual Relevance** as

$$I_p(D \rightarrow X|Q) = p(X|Q; D) - p(X|Q).$$

This quantity measures how much conditioning on the demonstration D increases the model’s support for the labeled output X . Here, “relevance” refers to the usefulness of D as contextual evidence for inferring X , rather than merely lexical or semantic similarity between D and Q .

The goal of LCS is to measure the ICL effectiveness of a given model on a given task, namely the extent to which providing demonstrations can improve the model beyond the zero-shot setting. This goal is different from evaluating the effectiveness of a particular selected demonstration set. A performance metric such as accuracy or exact match under one prompt can only show whether the model answers correctly with those specific demonstrations. It cannot distinguish whether poor performance is caused by the model-task pair having little potential to benefit from demonstrations or by the selected demonstrations failing to provide sufficiently relevant context. In contrast, LCS evaluates whether the learning gain systematically increases as demonstrations become more contextually relevant. Thus, LCS is designed to guide whether demonstrations are worth providing for the given model and task during inference, while performance metrics evaluate the realized effect of a particular demonstration configuration.

Based on this motivation, if ICL is effective for a given model and task, demonstrations that provide more contextual relevance should induce larger learning gains. Conversely, when the ICL effectiveness is low, increasing the contextual relevance of the demonstrations brings only marginal changes in learning gain. We have that the learning gain $I_p(X \rightarrow D|Q)$ and the contextual relevance $I_p(D \rightarrow X|Q)$ satisfy the following theorem:

Theorem 1 (Linear Relation between Learning Gain and Contextual Relevance).

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

The proof of Theorem 1 is presented in Appendix A.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. We

use this slope to measure ICL effectiveness and call it the **Learning-to-Context Slope (LCS)**.

Based on Theorem 1, it can be observed that there are two main factors influencing the ICL effectiveness during inference: *the contextual alignment capability* that learns the question-relevant information from the demonstrations ($\hat{p}(D|Q)$) and *the output calibration capability* that verifies the correctness of the output for 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* $\hat{p}(D|Q)$: The model fails to adequately comprehend the task-relevant information in the provided demonstrations. (ii) *High Output Calibration Capability* $\hat{p}(X|Q)$: The model possesses a strong inherent ability to verify the input consistency with the given output, leaving less room for demonstrations to further reduce the generation loss. We further discuss the meaning of the contextual alignment capability and the output calibration ability in detail in § 2.4.

2.3 Calculation of LCS

In this part, we discuss the calculation of LCS. In practice, let $\hat{p}$ represent the empirical probability distribution and $C = (q_i, x_i, d_i)_n$ be the sampling on $\mathcal{C}$. The calculation of LCS consists of two steps: reasoning process paraphrasing and likelihood calculation. The detailed prompts used for paraphrasing and synthesis of LCS are provided in Appendix B.2, and additional implementation details are provided in Appendix B.1.

Reasoning process paraphrasing For tasks that contain human-labeled reasoning processes, directly scoring the original reasoning process may introduce format-induced bias, because the human-written reasoning format may differ from the format preferred by the model. This issue is especially relevant for models with distinctive reasoning styles, such as the “<think>” tag of Llama-R1 (DeepSeek-AI et al., 2025). To reduce this confounding factor, for each data instance and demonstration, we input the question, answer, and human-labeled reasoning process if available, and ask the model to rephrase the output using its preferred reasoning style. The paraphrased result is then used for the following likelihood calculation.

Likelihood calculation For a conditional probability $\hat{p}(A|B)$, we treat B as the input context and A as the target output. We format B and A

with the corresponding model chat template, feed the formatted string into the model, and compute the likelihood of the tokens belonging to A under teacher forcing. Specifically, if $A = (a_1, \dots, a_{|A|})$, we calculate the length-normalized likelihood as:

$$\hat{p}(A|B) = \exp \left(\frac{1}{|A|} \sum_{j=1}^{|A|} \log p_{\theta}(a_j|B; a_{<j}) \right), \quad (4)$$

where length normalization is applied to reduce the influence of sequence length (Dai et al., 2019). This procedure is used to compute all conditional probabilities involved in LCS, including $\hat{p}(d_i|q_i; x_i)$, $\hat{p}(d_i|q_i)$, $\hat{p}(x_i|q_i; d_i)$, and $\hat{p}(x_i|q_i)$.

Slope fitting For each sampled triple (q_i, x_i, d_i) , we compute the empirical learning gain and contextual relevance as:

$$s_i = I_{\hat{p}}(x_i \rightarrow d_i|q_i), \quad t_i = I_{\hat{p}}(d_i \rightarrow x_i|q_i).$$

Based on these empirical quantities, we calculate the slope of Theorem 1 ($r_{\hat{p}}$) on C using the least squares method (Wang et al., 2018):

$$r_{\hat{p}} = \frac{\sum_{i=1}^n (t_i - \bar{t})(s_i - \bar{s})}{\sum_{i=1}^n (t_i - \bar{t})^2}, \text{ where} \quad (5)$$

$$\bar{s} = \frac{1}{n} \sum_{i=1}^n s_i, \quad \bar{t} = \frac{1}{n} \sum_{i=1}^n t_i.$$

We use $r_{\hat{p}}$ as the value of LCS. It should be noticed that although $r_p = \frac{p(D|Q)}{p(X|Q)}$ in Theorem 1, 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 A.3, we discuss the impact of error and prove that the impact of the error on $r_{\hat{p}}$ is less than $\frac{\hat{p}(D|Q)}{\hat{p}(X|Q)}$.

2.4 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. Here, X denotes the labeled output provided by the dataset, and D denotes the demonstration selected by the evaluation protocol. Thus, the following discussion does not assume that X must be the model’s most likely output or that D must be generated by the model. Instead, we examine whether the model assigns high probability to the given task-relevant output and demonstration.

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 the given demonstration to the question compared to $\hat{p}_2$, showing that $\hat{p}_1$ is a better demonstration scorer. 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 contextual alignment capability than $\hat{p}_2$.

It is worth noting that while both $\hat{p}(D|Q)$ and $I_p(D \rightarrow X|Q)$ measure the consistency between the demonstration and the user input to some extent, their fundamental perspectives differ. $I_p(D \rightarrow X|Q)$ primarily focuses on the data perspective, measuring whether the demonstration provides useful contextual evidence for inferring the labeled output. In contrast, $\hat{p}(D|Q)$ primarily focuses on the model perspective, observing whether the model can recognize the relationship between the input and the demonstration during likelihood scoring.

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 noted 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 one 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 given question. When $\hat{p}(X|Q)$ is already high, the model may have strong output calibration in the zero-shot setting, and the additional room for demonstrations to reduce the loss is limited, leading to low ICL effectiveness.

2.5 ICL Effectiveness without Labeling

Since the calculation of LCS in § 2.3 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 between LCS and synthetic data and real data when only unlabeled task inputs are available.

Theorem 2. Let $\hat{D}, D^*$ denote two demonstrations

satisfying that, for all $X \sim \mathcal{X}$ and $Q \sim \mathcal{Q}$:

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

The above condition means that demonstration D^* is better than $\hat{D}$ in helping to generate the correct answer. Let $\hat{X}$ denote the model-predicted output used with $\hat{D}$, and X^* denote the labeled output used with D^* . 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 about ICL in Theorem 2 suggests that the more demonstrations that can help the model make correct predictions, the larger the corresponding derived LCS we present. Considering that previous work has shown that the quality of synthesized data is generally lower than that of annotated data (Ashok and 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 ICL effectiveness to some extent, albeit conservatively, the resulting effectiveness estimate is lower than that for the real data.

3 Experiment

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

3.1 Experiment Setup

Dataset We conduct experiments on four mainstream tasks spanning diverse capabilities and domains across representative benchmark evaluation settings: math (GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021)), code (HumanEval (Chen et al., 2021a), MBPP (Austin et al., 2021)), reasoning (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)). For datasets of math, reasoning, and domain-specific knowledge, we use Exact Match (EM) (Cobbe et al., 2021) as the evaluation metric. For the code datasets, we use Pass@1 (Chen et al., 2021a) as the metric.

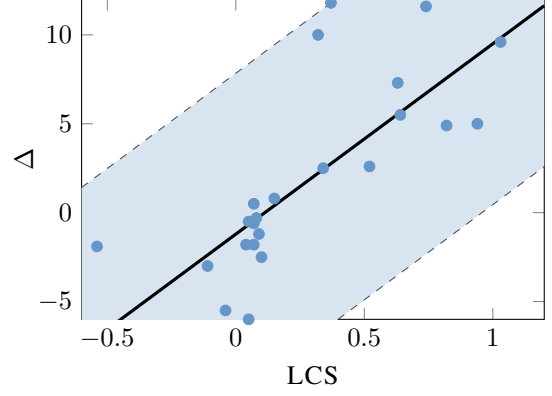


Figure 2: The performance improvement Δ brought by ICL (y-axis) with different LCS (x-axis) on different models and datasets. The solid line in the graph represents the fitted line for all data points. The Pearson coefficient is 0.737.

Model We conduct our experiments on 8 mainstream LLMs, which cover different ICL capabilities to fully evaluate whether our metric can reflect the ICL effectiveness. We discuss how to adapt LCS to black-box LLMs in Appendix D.2.

Implementation Details We evaluate the ICL performance on all datasets under 0-shot and 1-shot settings with a CoT-style prompt, 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 sample number to 1 and 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 for each dataset and model. All our experiments are based on Transformers (Wolf et al., 2020) and vLLM (Kwon et al., 2023). These settings ensure a fair comparison across models while preserving practical efficiency and reproducibility.

3.2 RQ1. How to Reliably Evaluate ICL Effectiveness?

To prove that LCS is a better metric than performance change for reflecting the ICL effectiveness, our main experiment includes two parts from the perspective of whether performance change can indicate effectiveness: (i) In §3.2.1, we evaluate that when performance changes can reflect the effectiveness of ICL, where LCS can also reflect the ICL effectiveness. (ii) In §3.2.2, we present that LCS can still reflect the ICL effectiveness, even when

Dataset	Llama2-7b		Llama3.1-8b		Llama-R1-8b		Llama3.1-70b		Ministral-8b		Qwen2.5-7b		Qwen3-8b		Qwen3.5-35b-A3B	
	Δ	LCS	Δ	LCS	Δ	LCS	Δ	LCS	Δ	LCS	Δ	LCS	Δ	LCS	Δ	LCS
GSM8K	+10.0±0.7	0.32±0.04	-1.8±0.6	0.07±0.03	-6.0±0.8	0.05±0.03	-0.9±0.5	0.05±0.03	+0.6±0.5	0.14±0.04	-0.4±0.5	0.08±0.03	-1.1±0.5	0.03±0.03	-0.6±0.4	-0.02±0.02
MATH	+9.6±1.1	1.03±0.07	+2.4±0.8	0.34±0.06	-1.2±0.8	0.09±0.06	-3.6±0.9	-0.13±0.06	-1.8±0.9	0.04±0.06	+2.2±0.9	0.81±0.06	-1.4±0.9	-0.21±0.06	-0.8±0.8	-0.05±0.05
HumanEval	-0.6±2.1	0.07±0.08	-2.5±2.3	0.10±0.08	-3.0±2.4	-0.11±0.08	-1.1±2.0	0.04±0.07	-0.8±2.1	0.06±0.07	-1.0±2.1	0.05±0.08	-1.3±2.0	0.02±0.07	-0.7±1.8	-0.03±0.06
MBPP	-0.5±1.3	0.05±0.05	+0.8±1.2	0.15±0.05	-6.4±1.4	0.07±0.06	+0.3±1.0	0.09±0.05	+0.4±1.1	0.11±0.05	+0.2±1.1	0.12±0.05	-0.6±1.0	0.05±0.05	-0.4±0.9	0.01±0.04
ARC-C	+11.6±0.7	0.74±0.04	-1.9±0.7	-0.54±0.04	-0.3±0.7	0.08±0.04	-0.7±0.6	0.06±0.03	+0.6±0.7	0.14±0.04	+0.5±0.7	0.17±0.04	-0.3±0.6	0.09±0.04	-0.4±0.5	0.02±0.03
MMLU-Pro	+5.5±0.7	0.64±0.05	+2.6±0.7	0.52±0.05	-5.5±0.8	-0.04±0.05	+0.6±0.6	0.16±0.04	+0.7±0.6	0.16±0.04	+0.8±0.6	0.18±0.04	+0.4±0.6	0.12±0.04	-0.5±0.5	-0.01±0.03
FinQA	+7.3±0.6	0.63±0.04	+4.9±0.6	0.82±0.04	-1.8±0.7	0.04±0.04	+7.6±0.6	0.77±0.04	+4.8±0.7	0.71±0.04	+4.9±0.6	0.29±0.04	+6.7±0.6	0.53±0.04	+0.4±0.5	0.11±0.04
Amazon	+0.5±2.4	0.07±0.09	+5.0±2.5	0.94±0.10	+11.8±2.7	0.37±0.08	+16.0±2.5	0.79±0.09	+4.0±2.2	0.69±0.09	+27.0±2.4	0.42±0.10	-1.0±2.4	0.03±0.08	+0.2±1.8	0.08±0.08

Table 1: Result across different models and datasets averaged on 5 runs. Δ denotes the performance change of 1-shot using ICL compared to 0-shot. ARC-C denotes ARC-Challenge and Amazon denotes Amazon Review. Results in green indicate a significant improvement with ICL, while those in red indicate improvement less than 1.0 or $LCS < 0.2$. Detailed performance is in Appendix D.1.

the demonstrations do not lead to improvements.

First, we discuss that LCS can accurately reflect the effectiveness of ICL. Subsequently, we provide experimental evidence demonstrating that performance improvement is insufficient to accurately reflect the ICL effectiveness. In addition, we present that LCS can reflect the performance improvement brought by ICL to a certain extent. We further examine the feasibility of applying LCS to black-box LLMs using a sampling-based pseudo-likelihood estimator, which is shown in Appendix D.2.

3.2.1 LCS Reliably Reflects the ICL Effectiveness

According to the main experimental results shown in Table 1, settings with significant performance improvements after using ICL ($\Delta \geq 1.0$) are highly consistent with settings where LCS is large, especially those where $LCS > 0.2$. In contrast, settings with insignificant or negative performance changes generally correspond to small LCS values. **This demonstrates that, in cases where performance improvement can serve as a reliable indicator of ICL effectiveness, LCS also accurately reflects whether providing demonstrations is beneficial for the given model and task.** Notably, performance metrics such as EM or accuracy only measure the realized effect of a particular demonstration configuration, whereas LCS aims to measure the extent to which the model can benefit from demonstrations on the given task. Therefore, the consistency between Δ and LCS in Table 1 supports the reliability of LCS as a task-level indicator of ICL effectiveness. Besides, there are several observations that follow:

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-R1-8b, 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, LLMs with high capability like 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. Consequently, the final trajectory of LCS is determined by the specific interplay and relative growth rates of these two competing factors, rather than following a strict linear correlation with model scale.

The performance improvement brought by ICL is positively related to LCS. To better 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 also discuss the empirical threshold of the ICL effectiveness using LCS in Appendix C.1 to guide the real application.

3.2.2 The Performance Change Cannot Reflect the ICL Effectiveness

In §3.1, we assume that whether performance improves or not genuinely reflects 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 that: (i) On subsets where EM is uninformative by construction, LCS still detects whether demonstrations induce systematic likelihood changes, which does not imply that the model is deployment-ready but identifies a latent learning signal that accuracy cannot distinguish. (ii) LCS has higher reliability, unlike performance, which is susceptible to factors like the instructions, as it directly evaluates $p(X|Y)$ by using Y as input and X as output without relying on instructions (Appendix B.1), thus providing a more faithful reflection of the ICL effectiveness.

This micro-level distinction is illustrated particularly clearly in Figure 3. The bad-case subsets of both GSM8K and Amazon have the same observed performance change ($\Delta = 0$), and are therefore indistinguishable using performance alone. However, their LCS values differ substantially (0.06 versus 1.06), revealing qualitatively different responses to contextual evidence. This illustrates the main diagnostic advantage of LCS: it preserves graded likelihood-level information even when discrete performance metrics collapse different cases to the same outcome.

3.2.3 LCS Outperforms Other Likelihood Baselines

Since LCS is derived from the decrease in loss brought by demonstrations, a natural question is whether it provides additional information beyond standard likelihood-based metrics. To answer this question, we compare LCS with three representative baselines: ΔNLL , which measures the reduction in the negative log-likelihood of the gold answer after adding demonstrations; ΔPPL , which measures the corresponding reduction in perplex-

Metric	$\rho_{\Delta\text{Perf}} \uparrow$	CV $\downarrow$
ΔNLL	0.61	0.34
ΔPPL	0.53	0.47
ΔCE	0.66	0.29
Influence	0.60	0.36
LCS	0.87	0.18

Table 2: Comparison between LCS and likelihood-based baselines. ΔNLL , ΔPPL , and ΔCE denote the reductions in negative log-likelihood, perplexity, and cross-entropy of the gold answer after adding demonstrations, respectively. Influence corresponds to the baseline in Table 3. Spearman $\rho_{\Delta\text{Perf}}$ measures correlation with actual ICL performance improvement, and CV measures variation across prompt templates or demonstration seeds.

ity; and ΔCE , which measures the reduction in token-level cross-entropy. These baselines directly quantify whether a demonstration improves the likelihood of the target answer, while LCS further measures whether such a learning gain changes systematically with the contextual relevance of demonstrations across varied tasks and settings.

As shown in Table 2, LCS achieves the highest correlation with actual ICL performance improvement, with a Spearman correlation of 0.87, substantially outperforming ΔNLL (0.61), ΔPPL (0.53), and ΔCE (0.66). This result indicates that LCS is more predictive of when ICL leads to observable performance gains. Moreover, LCS also obtains the lowest CV (0.18), compared with 0.34, 0.47, and 0.29 for ΔNLL , ΔPPL , and ΔCE , respectively. This suggests that LCS is less sensitive to prompt templates and demonstration seeds. Overall, these results show that LCS is not merely a re-parameterization of likelihood reduction. Instead, by explicitly modeling the relationship between contextual relevance and learning gain, LCS provides a more predictive and robust diagnostic signal for evaluating ICL effectiveness.

3.2.4 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

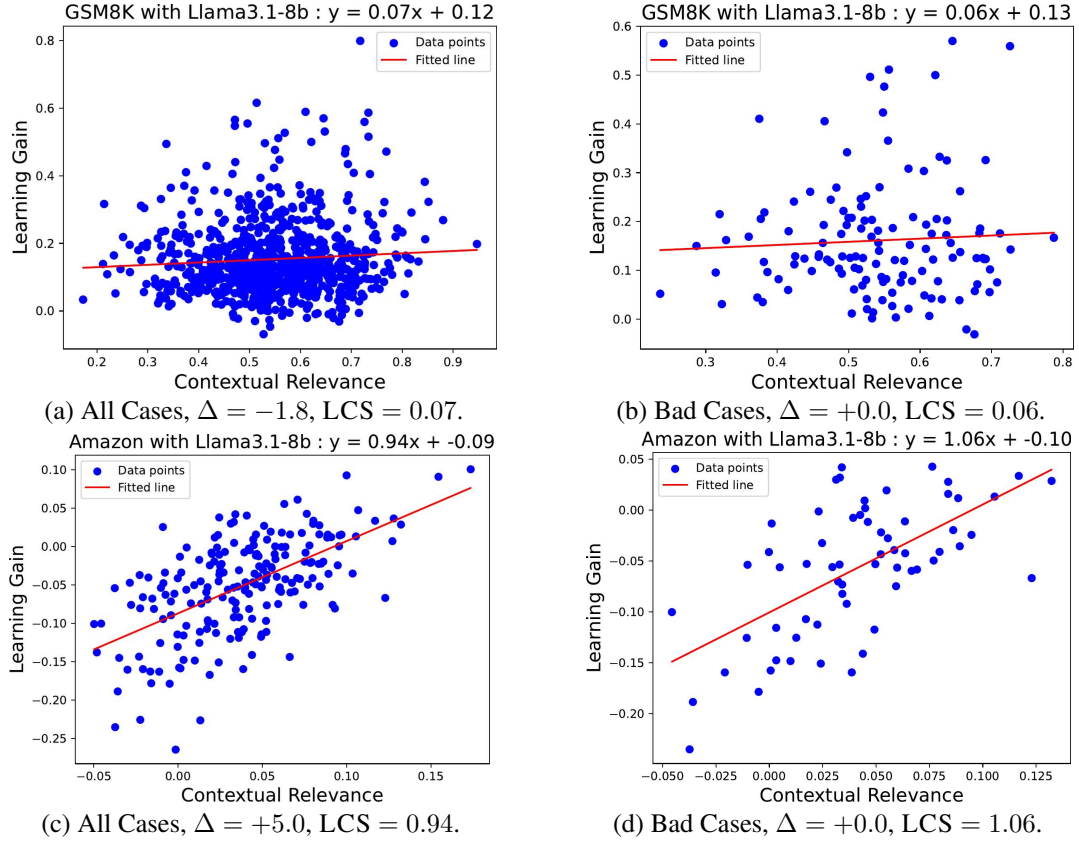


Figure 3: The change between contextual relevance and learning gain of using Llama3.1-8b on GSM8K and Amazon under the full set and the bad cases of ICL. Each point denotes the result of one instance. Δ denotes the performance change of ICL compared with zero-shot.

Method	GSM8K	MATH	HumanEval	MBPP	ARC-C	MMLU-Pro	FinQA	Amazon
Zero-Shot	86.4	48.4	65.9	54.8	82.1	50.4	49.7	63.5
BM25	84.2	50.8	63.4	55.6	80.2	53.0	54.6	68.5
GTR	82.1	50.8	63.4	55.8	80.5	53.5	55.0	68.5
Influence	83.9	51.0	64.6	55.3	81.3	52.4	54.6	69.5
IDS	85.3	50.4	64.0	55.6	82.4	52.4	54.6	68.0
Ours	86.4	51.2	65.9	56.1	82.5	54.3	55.1	70.0

Table 3: Performance of different demonstration selection methods using Llama3.1-8b. Baselines include BM25 (Robertson and Zaragoza, 2009), GTR (Luo et al., 2023), Influence (Nguyen and Wong, 2023), and IDS (Qin et al., 2024). The best performance of each setting is marked in **bold**.

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 3, our method outperforms other baselines, demonstrating the effectiveness of the LCS-based method. This indicates that, compared to surface-level lexical matching, selecting demonstrations based on estimated information gain aligns

more closely with the model’s underlying inference mechanism. In addition, on the datasets with low ICL effectiveness (e.g., GSM8K, ARC-Challenge), no method brings significant improvement, which is consistent with our conclusion in §3.2.1. It is worth noting that the method proposed in this section is mainly intended to validate the correctness of our theoretical analysis and to provide guidance for practical applications, rather than to introduce a new state-of-the-art method. Therefore, we leave more significant performance improvements and efficiency enhancements to future work.

3.2.5 Sensitivity of LCS to Prompt Order

Considering the discussion in Appendix C.2, in this section, we evaluate whether LCS is robust to prompt ordering by comparing the original prompt format with a swapped format that exchanges the positions of the demonstration and the user question. We keep the model, dataset, retrieved demonstration, and decoding setup unchanged, and only modify the prompt order. The goal is to test whether the practical usefulness of LCS depends

Dataset	Original		Swapped	
	Δ	LCS	Δ	LCS
GSM8K	-1.8	0.07	-1.4	0.10
Amazon	+5.0	0.94	+4.3	0.92
ARC-C	-1.9	-0.54	-1.0	-0.52

Table 4: The performance of different prompt orders with Llama3.1-8b. We compare the original prompt format (Demo $\rightarrow$ Question) and the swapped format (Question $\rightarrow$ Demo).

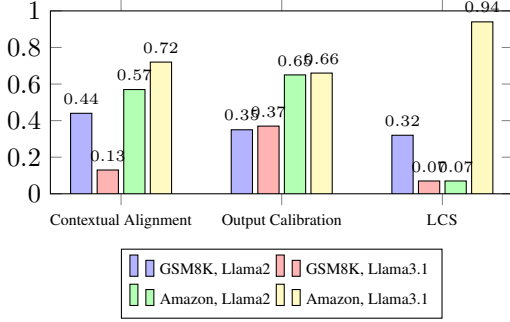


Figure 4: The average of the contextual alignment capability ($\hat{p}(D|Q)$), the output calibration capability ($\hat{p}(X|Q)$) and LCS of Llama2-7b and Llama3.1-8b on GSM8K and Amazon Review.

critically on the idealized order-invariance assumption. Table 4 shows that swapping the prompt order changes the absolute values of Δ and LCS to some extent, but the overall qualitative pattern is largely preserved. In particular, strong-positive, near-zero, and negative-LCS settings remain in the same regime after the swap. This suggests that, although auto-regressive LLMs are not strictly order-invariant, LCS is relatively stable under this prompt perturbation. Therefore, its practical value does not rely on a single prompt template, providing empirical support for using LCS as a robust effectiveness signal of ICL.

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

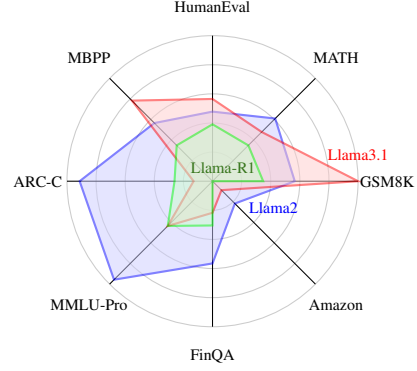


Figure 5: The intercept of the fitted line obtained with Equation 5 on each dataset and each model.

GSM8K and Amazon Review, as shown in Figure 4. 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 $\hat{p}$, where we discuss how errors affect LCS in detail in Appendix A.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 obtained in Equation 5 also reflects the effectiveness of ICL under different settings. Therefore, we present 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 A.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. This inspires future work to further

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	0.26	-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

Table 5: Performance change (Δ) and LCS using human-labeled and synthetic demonstrations. Results in green indicate a significant improvement with ICL, while those in red indicate improvement less than 1.0 or $LCS < 0.2$.

investigate the mechanisms behind this diminishing marginal gain, thereby providing theoretical guidance for designing efficient ICL for LLMs.

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 A.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. Consequently, these findings suggest that future work utilizing LCS should prioritize multi-shot settings to mitigate estimation variance, thereby ensuring a more robust and accurate assessment of intrinsic ICL effectiveness.

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

To verify the conclusions regarding the synthetic data presented in §2.5, we conduct experiments to calculate LCS using synthetic data. During synthe-

sis, we follow the procedure in Wang et al. (2025b) by inputting the task definition to generate corresponding demonstrations. In each iteration, we provide the model with the task definition and the synthetic results from the previous iteration (empty in the first iteration), asking the model to generate demonstrations. The prompt we used is shown in Appendix B.2. 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 conduct experiments on four datasets, which are shown in Table 5. From the table, we observe the following: (i) The trend of LCS using synthetic data is consistent with that derived from labeled data, showing that synthetic data tracks the same qualitative trend of ICL effectiveness in the evaluated settings. (ii) Compared to labeled data, the values of LCS obtained from synthetic data are smaller, which supports the conclusion of Theorem 2 and further indicates that our theoretical analysis remains broadly applicable under practical synthetic demonstration scenarios.

4 Related Work

In-Context Learning In-context learning (ICL) guides LLM reasoning by providing task-relevant demonstrations in the input, thereby improving performance (Brown et al., 2020; Dong et al., 2024; Zhao et al., 2025). Existing ICL research mainly focuses on demonstration construction and selection. For demonstration construction, many studies improve demonstration quality offline, such as increasing diversity through synthetic data generation or diverse example selection to provide broader task coverage and improve model generalization (Wang et al., 2024a, 2025a; Chen et al., 2023a; Su et al., 2024; Levy et al., 2023), and synthesizing or augmenting reasoning steps to better guide inference (Li et al., 2024; Zelikman et al., 2022; ZHAO et al., 2023). Other work dynamically generates or rewrites demonstrations based on the user input, sometimes leveraging the LLM itself to better adapt examples to individual queries (He et al., 2024; Chang and Fosler-Lussier, 2023; Kim et al., 2022). For demonstration selection, existing methods aim to retrieve examples most relevant or informative to the user query (Luo et al., 2024; Vu et al., 2023), using signals such as n-grams (Li et al., 2023), embedding-based semantic similar-

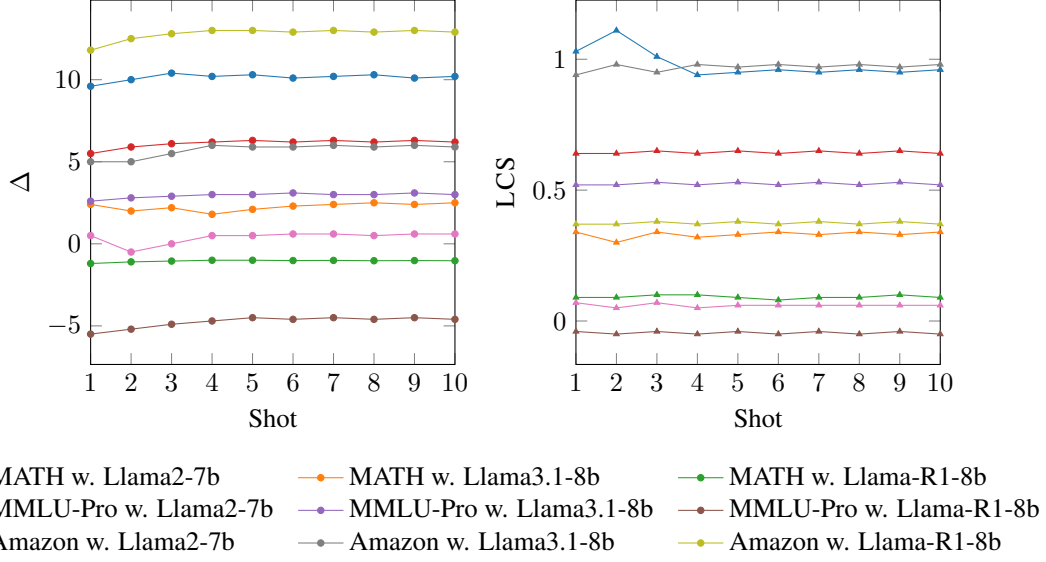


Figure 6: The performance change (Y-axis, left figure) and LCS (Y-axis, right figure) on MATH, MMLU-Pro, and Amazon Review with different shots (X-axis) using various models.

ity (Yang et al., 2023; Luo et al., 2023), or hybrid retrieval and ranking strategies (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 to better understand and improve its reasoning capability (Zhou et al., 2024; Dong et al., 2024). One line of research studies how ICL ability emerges by controlling the types of tasks used during pretraining (Edelman et al., 2024b; Han et al., 2023). Existing findings suggest that ICL ability depends more on task diversity than data scale, with models gradually generalizing from in-domain to out-of-domain tasks (Raventos et al., 2023). Related work further suggests that knowledge acquisition and ICL ability may rely on functionally distinct components: increasing data volume mainly improves knowledge acquisition, whereas ICL benefits more from diverse training tasks (Nguyen and Reddy, 2025). Another line of work studies ICL reasoning by examining the relationship between demonstrations and the user query (Park et al., 2025; Li et al., 2025; Min et al., 2022; Wang et al., 2023a). Some studies argue that ICL works by learning input-label mappings from demonstrations (Kossen et al., 2024), while others suggest that models learn and imitate the reasoning processes embedded in demonstrations (Lampinen et al., 2022). These findings provide complementary explanations of how information contained in demonstrations is utilized during inference.

However, most prior work focuses on improving ICL performance or explaining why it works, often

implicitly assuming its effectiveness. Recent studies instead show that ICL may provide little benefit or even degrade performance for certain tasks and models (DeepSeek-AI et al., 2025; Huang and Wang, 2025; Zheng et al., 2025). In this work, we investigate the main factors that determine ICL effectiveness and propose a method for evaluating whether ICL is effective for a given model and task, with the goal of informing future ICL research and applications.

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 observe 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.

Acknowledgment

We gratefully acknowledge the support of the National Natural Science Foundation of China (NSFC) via grant 62236004 and 62476073.

References

- Rishabh Agarwal, Avi Singh, Lei M Zhang, Bernd Bohnet, Luis Rosias, Stephanie C.Y. Chan, Biao Zhang, Aleksandra Faust, and Hugo Larochelle. 2024. [Many-shot in-context learning](#). In *ICML 2024 Workshop on In-Context Learning*.
- Dhananjay Ashok and Jonathan May. 2025. [A little human data goes a long way](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 381–413, Vienna, Austria. Association for Computational Linguistics.
- Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. [Program synthesis with large language models](#).
- Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(null):1137–1155.
- 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. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Shuaichen Chang and Eric Fosler-Lussier. 2023. [Selective demonstrations for cross-domain text-to-SQL](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14174–14189, Singapore. Association for Computational Linguistics.
- 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 Hebggen 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. 2021a. [Evaluating large language models trained on code](#).
- Wei-Lin Chen, Cheng-Kuang Wu, Yun-Nung Chen, and Hsin-Hsi Chen. 2023a. [Self-ICL: Zero-shot in-context learning with self-generated demonstrations](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15651–15662, Singapore. Association for Computational Linguistics.
- Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023b. [Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks](#). *Transactions on Machine Learning Research*.
- 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. 2021b. Finqa: A dataset of numerical reasoning over financial data. *Proceedings of EMNLP 2021*.
- 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. 2021. [Training verifiers to solve math word problems](#).
- Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov.

2019. [Transformer-xl: Attentive language models beyond a fixed-length context.](#)

- 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, Cheng-gang 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, Miao-jun 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, Shan-huang 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, Wanbiao 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. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.](#)
- 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. 2024. [A survey on in-context learning.](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.
- Ezra Edelman, Nikolaos Tsilivis, Benjamin L. Edelman, eran malach, and Surbhi Goel. 2024a. [The evolution of statistical induction heads: In-context learning markov chains.](#) In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Ezra Edelman, Nikolaos Tsilivis, Benjamin L. Edelman, eran malach, and Surbhi Goel. 2024b. [The evolution of statistical induction heads: In-context learning markov chains.](#) In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- 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 Esobu, 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-Yearly, 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 Tsimploukelli, 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 Rapparthi, 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, Wenxin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwon 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, Dingkan 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 Smithers, 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, Gregory 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 Damla, 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 Khabisa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Kenally, 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 Ritner, 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. 2024. [The llama 3 herd of models](#).
- Samaksh Gulati, Anshrit Verma, Manoj Parmar, and Palash Chaudhary. 2023. [Efficacy of machine-generated instructions](#).
- Qi Guo, Leiyu Wang, Yidong Wang, Wei Ye, and Shikun Zhang. 2024. [What makes a good order of examples in in-context learning](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14892–14904, Bangkok, Thailand. Association for Computational Linguistics.
- Xiaochuang Han, Daniel Simig, Todor Mihaylov, Yulia Tsvetkov, Asli Celikyilmaz, and Tianlu Wang. 2023. [Understanding in-context learning via supportive pretraining data](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12660–12673, Toronto, Canada. Association for Computational Linguistics.
- Yaru Hao, Yutao Sun, Li Dong, Zhixiong Han, Yuxian Gu, and Furu Wei. 2022. [Structured prompting: Scaling in-context learning to 1,000 examples](#).
- Wei He, Shichun Liu, Jun Zhao, Yiwen Ding, Yi Lu, Zhiheng Xi, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. [Self-demos: Eliciting out-of-demonstration generalizability in large language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3829–3845, Mexico City, Mexico. Association for Computational Linguistics.

- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. [Measuring mathematical problem solving with the MATH dataset](#). In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- Donghao Huang and Zhaoxia Wang. 2025. [Explorable sentiment analysis with deepseek-r1: Performance, efficiency, and few-shot learning](#).
- Andrew Jesson, Nicolas Beltran-Velez, and David Blei. 2025. [Can generative AI solve your in-context learning problem? a martingale perspective](#). In *The Thirteenth International Conference on Learning Representations*.
- Masahiro Kaneko, Youmi Ma, Yuki Wata, and Naoaki Okazaki. 2025. [Sampling-based pseudo-likelihood for membership inference attacks](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8894–8907, Vienna, Austria. Association for Computational Linguistics.
- Gyeonghary Kim, Won Ik Cho, Seok Min Shin, Sang-goo Lee, and Jamin Kim. 2022. Self-generated in-context learning: Leveraging autoregressive language models as a demonstration generator. *arXiv preprint arXiv:2206.08082*.
- Jannik Kossen, Yarin Gal, and Tom Rainforth. 2024. [In-context learning learns label relationships but is not conventional learning](#). In *The Twelfth International Conference on Learning Representations*.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. 2022. [Can language models learn from explanations in context?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 537–563, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Noah Lee, Na Min An, and James Thorne. 2023. [Can large language models capture dissenting human voices?](#) In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4585, Singapore. Association for Computational Linguistics.
- Itay Levy, Ben Bogin, and Jonathan Berant. 2023. [Diverse demonstrations improve in-context compositional generalization](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1401–1422, Toronto, Canada. Association for Computational Linguistics.
- Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Eric Tang, Sumanth Hegde, Kourosh Hakhmaneshi, Shishir G. Patil, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025. [Llms can easily learn to reason from demonstrations structure, not content, is what matters!](#)
- Junlong Li, Jinyuan Wang, Zhuosheng Zhang, and Hai Zhao. 2024. [Self-prompting large language models for zero-shot open-domain QA](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 296–310, Mexico City, Mexico. Association for Computational Linguistics.
- Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023. [Unified demonstration retriever for in-context learning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4644–4668, Toronto, Canada. Association for Computational Linguistics.
- Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. [Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.
- Man Luo, Xin Xu, Zhuyun Dai, Panupong Pasupat, Mehran Kazemi, Chitta Baral, Vaiva Im-

- brasaite, and Vincent Y Zhao. 2023. [Dr.icl: Demonstration-retrieved in-context learning](#).
- Man Luo, Xin Xu, Yue Liu, Panupong Pasupat, and Mehran Kazemi. 2024. [In-context learning with retrieved demonstrations for language models: A survey](#). *Transactions on Machine Learning Research*. Survey Certification.
- Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. [Rethinking the role of demonstrations: What makes in-context learning work?](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Alex Nguyen and Gautam Reddy. 2025. [Differential learning kinetics govern the transition from memorization to generalization during in-context learning](#). In *The Thirteenth International Conference on Learning Representations*.
- Tai Nguyen and Eric Wong. 2023. [In-context example selection with influences](#).
- Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. [Justifying recommendations using distantly-labeled reviews and fine-grained aspects](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.
- OpenAI. 2025. Introducing gpt-5. <https://openai.com/index/introducing-gpt-5/>.
- Core Francisco Park, Ekdeep Singh Lubana, and Hidenori Tanaka. 2025. [Competition dynamics shape algorithmic phases of in-context learning](#). In *The Thirteenth International Conference on Learning Representations*.
- Kha Pham, Hung Le, Man Ngo, and Truyen Tran. 2025. [Rapid selection and ordering of in-context demonstrations via prompt embedding clustering](#). In *The Thirteenth International Conference on Learning Representations*.
- Chengwei Qin, Aston Zhang, Chen Chen, Anirudh Dagar, and Wenming Ye. 2024. [In-context learning with iterative demonstration selection](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7441–7455, Miami, Florida, USA. Association for Computational Linguistics.
- Allan Raventos, Mansheej Paul, Feng Chen, and Surya Ganguli. 2023. [Pretraining task diversity and the emergence of non-bayesian in-context learning for regression](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: Bm25 and beyond](#). *Found. Trends Inf. Retr.*, 3(4):333–389.
- Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. [Learning to retrieve prompts for in-context learning](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.
- Yikang Shen, Shawn Tan, Alessandro Sordoni, and Aaron Courville. 2019. [Ordered neurons: Integrating tree structures into recurrent neural networks](#). In *International Conference on Learning Representations*.
- Yi Su, Yunpeng Tai, Yixin Ji, Juntao Li, Yan Bowen, and Min Zhang. 2024. [Demonstration augmentation for zero-shot in-context learning](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14232–14244, Bangkok, Thailand. Association for Computational Linguistics.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, NIPS’14*, page 3104–3112, Cambridge, MA, USA. MIT Press.
- Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansoor Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. [Large language models for](#)

- data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.
- Tu Vu, Heming Liu, David Dohan, and Denny Zhou. 2023. Active learning principles for in-context learning with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4891–4901. Association for Computational Linguistics.
- Xingchen Wan, Han Zhou, Ruoxi Sun, and Serkan O Arik. 2025. [From few to many: Self-improving many-shot reasoners through iterative optimization and generation](#). In *The Thirteenth International Conference on Learning Representations*.
- Dingzirui Wang, Longxu Dou, Xuanliang Zhang, Qingfu Zhu, and Wanxiang Che. 2024a. [Improving demonstration diversity by human-free fusing for text-to-SQL](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1193–1207, Miami, Florida, USA. Association for Computational Linguistics.
- 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. 2025a. [In-context transfer learning: Demonstration synthesis by transferring similar tasks](#).
- Dingzirui Wang, Xuanliang Zhang, Keyan Xu, Qingfu Zhu, Wanxiang Che, and Yang Deng. 2025b. [V-synthesis: Task-agnostic synthesis of consistent and diverse in-context demonstrations from scratch via v-entropy](#).
- George Wang, Matthew Farrugia-Roberts, Jesse Hoogland, Liam Carroll, Susan Wei, and Daniel Murfet. 2024b. [Loss landscape geometry reveals stagewise development of transformers](#). In *High-dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning*.
- Guorong Wang, Yimin Wei, and Sanzheng Qiao. 2018. *Generalized Inverses: Theory and Computations*, volume 53 of *Developments in Mathematics*. Springer, Singapore.
- Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023a. [Label words are anchors: An information flow perspective for understanding in-context learning](#). In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Liang Wang, Nan Yang, and Furu Wei. 2024c. [Learning to retrieve in-context examples for large language models](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1752–1767, St. Julian’s, Malta. Association for Computational Linguistics.
- Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. 2023b. [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*.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuan-sheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. 2024d. [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*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Yanzheng Xiang, Hanqi Yan, Lin Gui, and Yulan He. 2024. [Addressing order sensitivity of in-context demonstration examples in causal language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6467–6481, Bangkok, Thailand. Association for Computational Linguistics.

- Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2019. [Quick and \(not so\) dirty: Unsupervised selection of justification sentences for multi-hop question answering](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2578–2589, Hong Kong, China. Association for Computational Linguistics.
- Tong Yang, Yu Huang, Yingbin Liang, and Yuejie Chi. 2024. [In-context learning with representations: Contextual generalization of trained transformers](#). In *ICML 2024 Workshop on Theoretical Foundations of Foundation Models*.
- Zhao Yang, Yuanzhe Zhang, Dianbo Sui, Cao Liu, Jun Zhao, and Kang Liu. 2023. [Representative demonstration selection for in-context learning with two-stage determinantal point process](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5443–5456, Singapore. Association for Computational Linguistics.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. [STar: Bootstrapping reasoning with reasoning](#). In *Advances in Neural Information Processing Systems*.
- Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. 2025. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 1684–1692. PMLR.
- Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. 2025. [Is in-context learning sufficient for instruction following in LLMs?](#) In *The Thirteenth International Conference on Learning Representations*.
- Jiachen ZHAO, Zonghai Yao, zhichao Yang, and hong yu. 2023. [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*.
- Tianshi Zheng, Yixiang Chen, Chengxi Li, Chunyang Li, Qing Zong, Haochen Shi, Baixuan Xu, Yangqiu Song, Ginny Y. Wong, and Simon See. 2025. [The curse of cot: On the limitations of chain-of-thought in in-context learning](#).
- Yuxiang Zhou, Jiazheng Li, Yanzheng Xiang, Hanqi Yan, Lin Gui, and Yulan He. 2024. [The mystery of in-context learning: A comprehensive survey on interpretation and analysis](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14365–14378, Miami, Florida, USA. Association for Computational Linguistics.

A Proof

A.1 Equation 3

Proof. Suppose $X = (x_1, \dots, x_{|X|})$, where x_t is the t -th token of X and $x_{<t} = (x_1, \dots, x_{t-1})$. By the autoregressive factorization of the predictive distribution, we have:

$$\begin{aligned}\mathbb{L}_p(X|Q; D) &= -\log p(X|Q; D) \\ &= \sum_{t=1}^{|X|} (-\log p(x_t|Q; D; x_{<t})) .\end{aligned}$$

For each token x_t , using Bayes' rule, we can rewrite the conditional probability as:

$$\begin{aligned}p(x_t|Q; D; x_{<t}) &= \frac{p(x_t, D|Q; x_{<t})}{p(D|Q; x_{<t})} \\ &= \frac{p(x_t|Q; x_{<t})p(D|Q; x_{\leq t})}{p(D|Q; x_{<t})},\end{aligned}$$

where $x_{\leq t} = (x_1, \dots, x_t)$. Substituting this expression into the loss gives:

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

Therefore, we obtain:

$$\mathbb{L}_p(X|Q; D) = \mathbb{L}_p(X|Q) - (\log p(D|Q; X) - \log p(D|Q)) .$$

□

A.2 Theorem 1

Proof.

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

Therefore, we can conclude that:

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

□

A.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 $\frac{\hat{p}(D|Q)}{\hat{p}(X|Q)}$.

A.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:

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

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:

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

□

B Additional Information

B.1 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 B.2.

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

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 paraphrase.

Prompt of Synthesis
<pre>```md {task_definition} ```</pre> 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.

Table 7: The prompt of the demonstration synthesis.

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.

B.2 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 6 and Table 7.

C Additional Discussion

C.1 Empirical Threshold of LCS

Specifically, based on Figure 2, we can use $\text{LCS} = 0.2$ as an empirical threshold, since when $\text{LCS} \leq 0.2$, 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.

C.2 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 3 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 A.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 3 even when instantiated with standard autoregressive LLMs.

In this work, we define the predictive probability for an output sequence $\mathbf{X} = (x_1, \dots, x_T)$ 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.

C.3 Why Removing Logarithm in Learning Gain

In this section, we discuss why, when defining the learning gain $I_p(X \rightarrow D \mid Q)$, we remove the log based on Equation 3. 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 about by introducing the demonstration. Then we have:

$$\begin{aligned} p(D \mid Q; X) &= e^k p(D \mid Q) \\ \Rightarrow p(D \mid Q; X) - p(D \mid Q) &= (e^k - 1)p(D \mid Q) \end{aligned}$$

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, $I_p(X \rightarrow D \mid Q)$ is positively correlated with the change in loss.

Model	Shot	Math		Code		Reasoning		Domain	
		GSM8K	MATH	HumanE	MBPP	A-C	M-P	FinQA	Amazon
Llama2-7b	0	12.7 $\pm$ 0.6	5.0 $\pm$ 0.8	14.0 $\pm$ 2.0	23.0 $\pm$ 1.2	34.6 $\pm$ 0.7	14.1 $\pm$ 0.6	10.5 $\pm$ 0.5	28.5 $\pm$ 2.2
	1	22.7 $\pm$ 0.7	14.6 $\pm$ 1.0	13.4 $\pm$ 2.1	22.5 $\pm$ 1.2	46.2 $\pm$ 0.8	19.6 $\pm$ 0.7	17.8 $\pm$ 0.6	29.0 $\pm$ 2.3
	Δ	+10.0 $\pm$ 0.7	+9.6 $\pm$ 1.1	-0.6 $\pm$ 2.1	-0.5 $\pm$ 1.3	+11.6 $\pm$ 0.7	+5.5 $\pm$ 0.7	+7.3 $\pm$ 0.6	+0.5 $\pm$ 2.4
Llama3.1-8b	0	86.4 $\pm$ 0.5	48.4 $\pm$ 0.7	65.9 $\pm$ 2.1	54.8 $\pm$ 1.2	82.1 $\pm$ 0.6	50.4 $\pm$ 0.6	49.7 $\pm$ 0.6	63.5 $\pm$ 2.3
	1	84.6 $\pm$ 0.5	50.8 $\pm$ 0.8	63.4 $\pm$ 2.2	55.6 $\pm$ 1.3	80.2 $\pm$ 0.7	53.0 $\pm$ 0.7	54.6 $\pm$ 0.6	68.5 $\pm$ 2.4
	Δ	-1.8 $\pm$ 0.6	+2.4 $\pm$ 0.8	-2.5 $\pm$ 2.3	+0.8 $\pm$ 1.2	-1.9 $\pm$ 0.7	+2.6 $\pm$ 0.7	+4.9 $\pm$ 0.6	+5.0 $\pm$ 2.5
Llama-R1-8b	0	86.1 $\pm$ 0.6	75.4 $\pm$ 0.7	70.7 $\pm$ 2.2	67.2 $\pm$ 1.3	84.8 $\pm$ 0.6	58.2 $\pm$ 0.7	45.2 $\pm$ 0.6	53.5 $\pm$ 2.5
	1	80.1 $\pm$ 0.7	74.2 $\pm$ 0.8	67.7 $\pm$ 2.3	60.8 $\pm$ 1.4	84.5 $\pm$ 0.7	52.7 $\pm$ 0.8	43.4 $\pm$ 0.7	65.3 $\pm$ 2.6
	Δ	-6.0 $\pm$ 0.8	-1.2 $\pm$ 0.8	-3.0 $\pm$ 2.4	-6.4 $\pm$ 1.4	-0.3 $\pm$ 0.7	-5.5 $\pm$ 0.8	-1.8 $\pm$ 0.7	+11.8 $\pm$ 2.7
Llama3.1-70b	0	94.6 $\pm$ 0.4	68.4 $\pm$ 0.8	80.4 $\pm$ 1.8	72.0 $\pm$ 1.0	89.6 $\pm$ 0.5	63.5 $\pm$ 0.5	58.4 $\pm$ 0.5	62.8 $\pm$ 2.2
	1	93.7 $\pm$ 0.5	64.8 $\pm$ 0.9	79.3 $\pm$ 1.9	72.3 $\pm$ 1.0	88.9 $\pm$ 0.6	64.1 $\pm$ 0.6	66.0 $\pm$ 0.6	78.8 $\pm$ 2.4
	Δ	-0.9 $\pm$ 0.5	-3.6 $\pm$ 0.9	-1.1 $\pm$ 2.0	+0.3 $\pm$ 1.0	-0.7 $\pm$ 0.6	+0.6 $\pm$ 0.6	+7.6 $\pm$ 0.6	+16.0 $\pm$ 2.5
Ministral-8b	0	83.5 $\pm$ 0.5	54.3 $\pm$ 0.8	66.8 $\pm$ 2.0	59.4 $\pm$ 1.1	79.8 $\pm$ 0.6	47.2 $\pm$ 0.6	44.6 $\pm$ 0.6	61.0 $\pm$ 2.0
	1	84.1 $\pm$ 0.5	52.5 $\pm$ 0.9	66.0 $\pm$ 2.0	59.8 $\pm$ 1.1	80.4 $\pm$ 0.7	47.9 $\pm$ 0.6	49.4 $\pm$ 0.7	65.0 $\pm$ 2.1
	Δ	+0.6 $\pm$ 0.5	-1.8 $\pm$ 0.9	-0.8 $\pm$ 2.1	+0.4 $\pm$ 1.1	+0.6 $\pm$ 0.7	+0.7 $\pm$ 0.6	+4.8 $\pm$ 0.7	+4.0 $\pm$ 2.2
Qwen2.5-7b	0	91.2 $\pm$ 0.5	75.0 $\pm$ 0.8	79.0 $\pm$ 2.0	66.8 $\pm$ 1.1	84.2 $\pm$ 0.6	56.3 $\pm$ 0.6	51.4 $\pm$ 0.5	48.0 $\pm$ 2.3
	1	90.8 $\pm$ 0.5	77.2 $\pm$ 0.9	78.0 $\pm$ 2.1	67.0 $\pm$ 1.1	84.7 $\pm$ 0.7	57.1 $\pm$ 0.6	56.3 $\pm$ 0.6	75.0 $\pm$ 2.4
	Δ	-0.4 $\pm$ 0.5	+2.2 $\pm$ 0.9	-1.0 $\pm$ 2.1	+0.2 $\pm$ 1.1	+0.5 $\pm$ 0.7	+0.8 $\pm$ 0.6	+4.9 $\pm$ 0.6	+27.0 $\pm$ 2.4
Qwen3-8b	0	92.4 $\pm$ 0.4	81.2 $\pm$ 0.8	76.5 $\pm$ 1.9	68.0 $\pm$ 1.0	86.4 $\pm$ 0.5	60.2 $\pm$ 0.5	53.8 $\pm$ 0.5	72.0 $\pm$ 2.1
	1	91.3 $\pm$ 0.5	79.8 $\pm$ 0.9	75.2 $\pm$ 2.0	67.4 $\pm$ 1.0	86.1 $\pm$ 0.6	60.6 $\pm$ 0.6	60.5 $\pm$ 0.6	71.0 $\pm$ 2.3
	Δ	-1.1 $\pm$ 0.5	-1.4 $\pm$ 0.9	-1.3 $\pm$ 2.0	-0.6 $\pm$ 1.0	-0.3 $\pm$ 0.6	+0.4 $\pm$ 0.6	+6.7 $\pm$ 0.6	-1.0 $\pm$ 2.4
Qwen3.5-35b-A3B	0	94.8 $\pm$ 0.4	84.4 $\pm$ 0.7	82.0 $\pm$ 1.7	74.0 $\pm$ 0.9	90.2 $\pm$ 0.4	66.8 $\pm$ 0.5	59.4 $\pm$ 0.4	78.0 $\pm$ 1.8
	1	94.2 $\pm$ 0.4	83.6 $\pm$ 0.8	81.3 $\pm$ 1.8	73.6 $\pm$ 0.9	89.8 $\pm$ 0.5	66.3 $\pm$ 0.5	59.8 $\pm$ 0.5	78.2 $\pm$ 1.9
	Δ	-0.6 $\pm$ 0.4	-0.8 $\pm$ 0.8	-0.7 $\pm$ 1.8	-0.4 $\pm$ 0.9	-0.4 $\pm$ 0.5	-0.5 $\pm$ 0.5	+0.4 $\pm$ 0.5	+0.2 $\pm$ 1.8

Table 8: The performance of different models on different datasets using 0-shot and 1-shot averaged on 5 runs. 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	GSM8K		MATH		FinQA	
	Δ	$r_{\hat{p}}x + b$	Δ	$r_{\hat{p}}x + b$	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 9: LCS with gpt-5-nano on GSM8K, MATH and FinQA. The probability is generated following (Kaneko et al., 2025). Considering the API cost, we randomly sample 128 examples from each dataset.

D Additional Experiments

D.1 Overall Performance

In this part, we present the performance of 0-shot and 1-shot on different models and datasets, as shown in Table 8.

D.2 LCS with Black-Box LLMs

Based on Theorem 1, the main point in calculating LCS with black-box LLMs is how to obtain the log probability. Following (Lee et al., 2023; Kaneko et al., 2025), we employ a sampling-based pseudo-likelihood estimator to recover 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 true probability. Although the above computational 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.

Metric	GSM8K	MATH	ARC-C	MMLU-Pro	FinQA	Amazon
Performance	86.4±0.6	51.2±0.8	82.5±0.7	54.3±0.7	55.1±0.6	70.0±2.5
LCS	0.07±0.03	0.34±0.06	-0.54±0.04	0.52±0.05	0.82±0.04	0.94±0.10

Table 10: The performance and LCS of Llama3.1-8b with 5 runs of demonstration permutation.

We conduct experiments using `gpt-5-nano` (OpenAI, 2025). Due to the high API cost, we only randomly sample 128 examples from GSM8K, MATH and FinQA for our experiments. The experimental results are shown in Table 9, from which we can observe that LCS accurately reflects the effectiveness of ICL, thereby demonstrating the feasibility of applying LCS to black-box models.

D.3 LCS with Demonstration Permutation

We analyze the performance and LCS of Llama3.1-8b with 5 runs of demonstration permutation. Results of Table 10 show the robustness of LCS to the demonstration permutation.