Uncertainty Unveiled: Can Exposure to More In-context Examples Mitigate Uncertainty for Large Language Models?

Yifei Wang[1,2], **Yu Sheng**[1,2], **Linjing Li**[1,2*], **Daniel Zeng**[1,2]
[1] State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
{wangyifei2022, shengyu2021, linjing.li, dajun.zeng}@ia.ac.cn

Abstract

Recent advances in handling long sequences have facilitated the exploration of long-context in-context learning (ICL). While much of the existing research emphasizes performance improvements driven by additional in-context examples, the influence on the trustworthiness of generated responses remains underexplored. This paper addresses this gap by investigating how increased examples influence predictive uncertainty—an essential aspect in trustworthiness. We begin by systematically quantifying the uncertainty of ICL with varying shot counts, analyzing the impact of example quantity. Through uncertainty decomposition, we introduce a novel perspective on performance enhancement, with a focus on epistemic uncertainty (EU). Our results reveal that additional examples reduce total uncertainty in both simple and complex tasks by injecting task-specific knowledge, thereby diminishing EU and enhancing performance. For complex tasks, these advantages emerge only after addressing the increased noise and uncertainty associated with longer inputs. Finally, we explore the evolution of internal confidence across layers, unveiling the mechanisms driving the reduction in uncertainty.

1 Introduction

In-context learning has emerged as a pivotal paradigm for modern large language models (LLMs) in addressing real-world challenges (Brown et al., 2020; Dong et al., 2024). By presenting a few learning examples through carefully crafted prompts, LLMs achieve remarkable performance without requiring weight updates. The latest techniques of equipping LLMs with long-context capabilities have made strides (Jin et al., 2024), including continued fine-tuning(Rozière et al., 2024), position extrapolation (Su et al., 2024) and innovative architectures (Peng et al., 2023; Gu and Dao, 2024), open new avenues for areas previously constrained by context length.

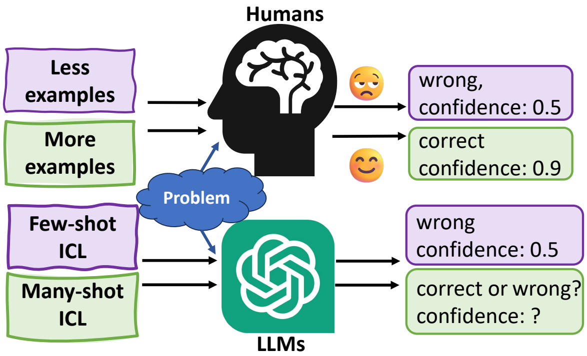


Figure 1: Humans tend to gain task-specific knowledge and confidence as they are exposed to more examples. This raises a natural question: can additional examples similarly reduce uncertainty in LLMs?

One such area is long-context ICL, also known as many-shot ICL, which involves feeding LLMs with hundreds or even thousands of input-output pairs. This regime of ICL allows LLMs to learn from large quantities of data once and could be deemed as a comparative alternative to fine-tuning methods. Despite its potential, the properties of many-shot ICL remain largely unexplored. While several studies have initiated preliminary investigations in this area, which mainly focus on performance gains from extra examples (Agarwal et al., 2024; Jiang et al., 2024), critical aspects such as trustworthiness and reliability of generations by LLMs (Wang et al., 2024) remain unexamined. Systematic investigation of these aspects is essential for advancing our understanding of long-context ICL and paves the way for its wider adoption in high-stake applications.

To fill this blank, we quantitatively examine the impact of increasing scales of in-context examples on LLMs' confidence through faithful uncertainty quantification (UQ) approaches. By incorporating model parameters, configurations, and various demonstration sets, we approximate the predictive

*Corresponding Authors.

Findings of the Association for Computational Linguistics: ACL 2025, pages 20659–20678

distribution in the output space. Then we compute entropy to measure total uncertainty (TU). Building on the framework proposed by (Ling et al., 2024), we employ a Bayesian framework to disentangle two core components from TU for many-shot ICL: epistemic uncertainty (EU) and aleatoric uncertainty (AU). EU arises from insufficient evidence or knowledge during model training, while AU stems from the inherent randomness and variability of the data (He et al., 2023) in Fig. 2. Our analysis reveals that the reduction in LLMs' uncertainty with more examples is primarily driven by a main decrease in EU. These examples enrich task-specific knowledge, thereby lowering EU, which in turn reduces TU and enhances performance. Furthermore, we demonstrate that the performance gains are attributed to increased informational content rather than extended context length. To explore the mechanisms behind reduced uncertainty, we project the residuals from all model layers into the vocabulary space, visualizing the evolution of internal confidence. The results reveal that long-context ICL enables LLMs to concentrate more logit mass on the correct answer and amplify the disparity between the correct response and distractors, effectively reducing uncertainty in predictions.

This study represents one of the earliest efforts to examine long-context ICL through the lens of uncertainty. The core research questions addressed are as follows:

- **RQ1:** Could more in-context examples mitigate uncertainty for LLMs? (§ 4.2)
- **RQ2:** Where do performance gains stem from, from the perspective of uncertainty decomposition? (§ 4.3)
- **RQ3:** What mechanisms underlie uncertainty reduction? (§ 5.2)

2 Related work

Long-context ICL The significant advancements in equipping LLMs with long context capabilities have expanded the potential for research in previously constrained areas, such as repository-level code understanding and multi-document QA. For ICL, an important emergent ability for LLMs (Brown et al., 2020), the extrapolation of context length enables the investigation into its performance limits and learning dynamics as the number of demonstrations scales.

Several studies have initiated preliminary investigations in this area. Agarwal et al. (2024), for instance, demonstrates notable performance gains with many-shot prompting across various generative and discriminative tasks using Gemini 1.5 Pro (Team et al., 2024). In parallel, Bertsch et al. (2024) offers valuable insights into the properties of many-shot ICL, particularly examining the influence of example retrieval and demonstration order. On a more optimistic note, Jiang et al. (2024) concludes that many-shot ICL can facilitate efficient adaptation of multimodal foundation models to new applications and domains. However, the benefits of long-context ICL are not universally positive. Li et al. (2024) argues that long-context models encounter difficulties with extreme-label classification tasks, especially when large label spaces are involved.

Uncertainty Quantification UQ has been extensively studied in traditional machine learning (Lakshminarayanan et al., 2017; Gawlikowski et al., 2022; Kong et al., 2023), which predominantly concentrates on estimating models' confidence and uncertainty in its prediction, called total uncertainty. Total uncertainty can be decomposed into two key components: epistemic (model) uncertainty and aleatoric (data) uncertainty (Hou et al., 2024; Valdenegro-Toro and Mori, 2022). The advent of LLMs has introduced new challenges in quantifying uncertainty, particularly due to the sequential and context-dependent nature of generative processes. Recent advances in UQ research can be categorized into two main approaches: black-box and white-box methods. Black-box UQ quantifies uncertainty by measuring the agreement across multiple generation samples (Zhang et al., 2024a), whereas white-box approaches assess internal model states or logits to capture intrinsic uncertainty (Liu et al., 2024; Bakman et al., 2024).

3 Uncertainty Quantification Framework for Long-context ICL

3.1 Formulation of ICL

Consider an LM $\mathcal{M}$ and a query $\boldsymbol{x}$, where $\mathcal{M}$ generates a response $\widehat{\boldsymbol{y}}$ by maximizing the joint probability $\mathcal{P}_\Theta(\widehat{\boldsymbol{y}} \mid \boldsymbol{x}) = \prod_{i\geq 1} \mathcal{P}_\Theta(\widehat{\boldsymbol{y}}_i \mid \widehat{\boldsymbol{y}}_{<i}, \boldsymbol{x})$. In the ICL regime, $\mathcal{M}$ would condition its output on a constructed prompt Ω, which typically includes an optional task-specific instruction $\mathcal{I}$, a series of N input-output demonstrations ("shots") $\boldsymbol{z}_{1:N}=\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$, and a test query $\boldsymbol{x}_{N+1}$. Consequently, the generation process of ICL can be

formalized as $\widehat{\boldsymbol{y}} := \mathcal{P}_\Theta(\widehat{\boldsymbol{y}} \mid \boldsymbol{x}_{N+1}, \boldsymbol{z}_{1:N}, \mathcal{I})$, enabling $\mathcal{M}$ to address diverse complex tasks (Gatt and Krahmer, 2018).

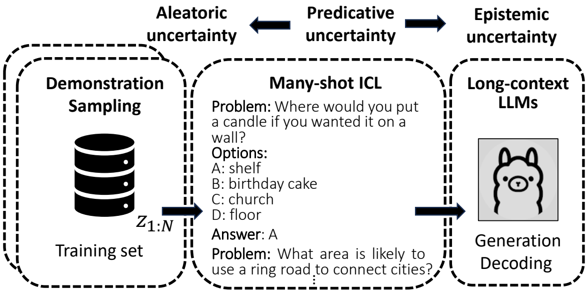


Figure 2: The sources of AU and EU in many-shot ICL. AU comes from the prompt Ω *e.g.*, vast examples and the process of demonstration selection. EU originates from the model's end, encompassing the generation and decoding processes.

3.2 Faithful Uncertainty Quantification

Predictive Distribution To quantify uncertainty stemming from both the demonstration sets $z_{1:N}$, and the model parameters or configurations Θ, we derive the predictive distribution by sampling generations across various configurations $\Theta \sim q(\Theta)$ and demonstration sets $z_{1:N} \sim Z$. This work focuses on classification and multiple-choice question-answering (MCQA) tasks. The selection of task types is discussed in Appendix B.

The advantage of UQ in these tasks lies in the categorical nature of their outputs: each numerical or symbolic label $\boldsymbol{y} \in \mathcal{Y}$ binds a predefined category or candidate answer. Thus, the probability of $\boldsymbol{y}$, denoted as $\mathcal{P}_{\boldsymbol{y}}$, is derived from the model's predicted logits and acts as a proxy for its confidence in their responses. Assume that for each demonstration set, we sample m decoded generations and repeat this process across L distinct sets $z^L_{1:N}$. This yields a probability set of size $L \times m$, capturing the uncertainty distribution over both demonstration sets and model configurations. Unlike classification or MCQA, where uncertainty can be assessed through well-defined probability distributions over discrete outputs, open-ended tasks involve variable-length outputs and lack clear ground truth, with no principled method existed for reliable UQ. Therefore, we hope we could probe the uncertain property of long-context ICL systems through MCQA tasks to provide a preliminary invertigation.

Entropy. By aggregating the probabilities from m decoded generations for each demonstration set into a distribution over the output space, we obtain $L \times |\mathcal{Y}|$ probability matrix $A_{L\times|\mathcal{Y}|}$, from which we compute the entropy as follows:

$$TU = -\mathcal{H}\left[\sigma\left(\left[\sum_{l=1}^{L} \mathcal{P}(\boldsymbol{y} \mid \boldsymbol{x}, z^l_{1:N})\right]_{\boldsymbol{y}\in|\mathcal{Y}|}\right)\right]$$

where σ is a normalization function that ensures the sum of probabilities equals one, and $\mathcal{H} = \sum_i p(\boldsymbol{x}) log(p(\boldsymbol{x}))$. Some studies indicate that logits may be uncalibrated (Liu et al., 2024; Agarwal et al., 2024). Aggregating the probability distributions from all decoded sequences can also help mitigate the errors and inaccuracies arising from uncalibrated logits, leading to a more reliable and robust output distribution.

3.3 Uncertainty Disentanglement

According to (Ling et al., 2024), from the Bayesian view, ICL maps demonstrations $z_{1:N}$ into a pre-existing latent concept β, which defines task-specific knowledge and enables LLMs to tackle a new in-domain task $\boldsymbol{x}_{N+1}$. The predictive distribution of ICL is formulated as follows:

$$\begin{aligned} p(\mathbf{y}|z_{1:N}) := \int & p(\mathbf{y}|x_{N+1}, z_{1:N}, \Theta, \beta) \\ & \cdot p(\beta|z_{1:N}) q(\Theta)\, d\beta\, d\Theta \end{aligned}$$

If Θ is specific, yielding $p(\mathbf{y}|z_{1:N}, \Theta) = \int p(\mathbf{y}|z_{1:N}, \beta, \Theta) p(\beta|z_{1:N}) d\beta$ with an associated entropy $H(\mathbf{y}|z_{1:T}, \beta, \Theta)$. The expected value of this entropy under different demonstration sets can be expressed as $\mathbb{E}_\beta\left[H(\mathbf{y}_T|\mathbf{x}_{1:T}, \beta, \Theta)\right]$, which serves as a metric to quantify the EU. AU is estimated as mutual information between $\mathbf{y}$ and the latent concept β as $I(\mathbf{y}, \beta|\Theta)$, which is the difference between TU and EU as follows:

$$\begin{aligned} & I(\mathbf{y}, \beta|\Theta) = \\ & \qquad H(\mathbf{y}|z_{1:N}, \Theta) - \mathbb{E}_\beta[H(\mathbf{y}|z_{1:N}, \beta, \Theta)] \end{aligned}$$

The latent concept β distribution could be obtained by sampling from different demonstrations. Beam search effectively approximates the posterior of Θ, which draws hypotheses from the most probable regions in the hypotheses space. Utilizing the probability matrix $A_{L\times m}$ obtained in Sec. 3.2, TU, EU and AU can be approximated as follows:

$$\begin{aligned} TU &= H(\sigma(\textstyle\sum[A_{j,:}])) \\ EU &= \frac{1}{L} H(\sigma(A_{j,:})) \\ AU &= H(\sigma(\textstyle\sum[A_{j,:}])) - \frac{1}{L} H(\sigma(A_{j,:})) \end{aligned}$$

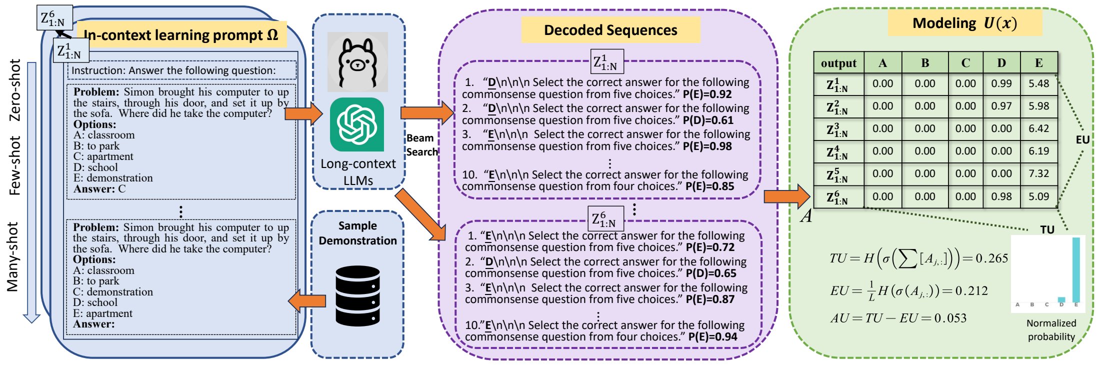


Figure 3: A workflow for uncertainty quantification and decomposition under many-shot ICL settings, involves the following components: a LLM $\mathcal{M}$ supporting long context windows, demonstration set selection, generation sampling, and the UQ modules detailed in Sec. 3.2 and 3.3.

Model	Size	Strategy	Support
Llama-3.1-8B	8B	Fine-tuning	128K
Mistral-7B-v0.2	7B	NTK-Aware Interpolation	32K
Qwen1.5-7B	7B	Fine-tuning	32K

Table 1: Long-Context LLMs Overview

4 Experiments

4.1 Experimental Settings

Models. We evaluate three widely used base models prior to instruction-tuning (Wei et al., 2022a): Llama-3.1-8B (Touvron et al., 2023), Mistral-7B-v0.2 (Jiang et al., 2023), and Qwen1.5-7B (Bai et al., 2023). The supported maximum context length, along with their respective strategies for long-context training, are summarized in Table 1.

Datasets and tasks. We define two modes for classification tasks and MCQA: easy and hard. The hard mode consists of three increasingly complex logical deduction tasks, including determining the order of a sequence of objects ranging from three to seven, from a suite of challenging algorithmic reasoning tasks known as BIG-Bench Hard (BBH) (Suzgun et al., 2023). In contrast, the easy mode encompasses traditional natural language understanding (NLU) tasks such as AGNews (Zhang et al., 2015) and SST2 (Socher et al., 2013), along with the commonsense reasoning task, CommonsenseQA (Talmor et al., 2019).

Long-context ICL settings. To investigate how uncertainty evolves with increasing exposure to examples, we apply UQ and uncertainty decomposition methods across different k-shot ICL. For demonstration selection, we randomly sample k shots from the training set for each test example. In all tasks, we employ beam search to generate 10 candidate outputs and set the temperature parameter as 0.7. For decomposing TU, we iterate six different demonstration sets to disentangle EU and AU. All open-source models are sourced from Hugging Face[1] and experimented on eight 80GB NVIDIA RTX A100 GPUs.

4.2 RQ1: Could more in-context examples mitigate uncertainty for LLMs?

Quality of Uncertainty Measures In the context of UQ, a key consideration is its ability to reflect the correctness and reliability of LLM outputs. High uncertainty most likely leads to incorrect predictions while low uncertainty indicates a higher likelihood of correct responses. To this end, we examine how the quality of uncertainty measures varies from few-shot to long-context ICL settings. Following prior works (Kuhn et al., 2023; Lin et al., 2024), we adopt Exact match as the metric for correctness and use uncertainty estimates to predict the correctness of response. We then compute AUROC [2] to evaluate whether the UQ measures employed are good indicators. The AUROC and accuracy results for Llama-3.1-8B are presented in Tab.8. As the number of demonstrations increases, AUROC values remain high with minimal fluctuations, suggesting that the UQ measures serve as high-quality indicators and generalize effectively to long-context ICL, which reinforces the validity of our experimental results and the conclusions drawn.

Average View *Overall, many-shot ICL effectively reduces LLMs' uncertainty across models and datasets.* As shown in Figs. 4 and 5, the results indicate a simultaneous rise in accuracy and con-

[1]Model weights are loaded at float16 precision.
[2]the Area Under the Receiver Operating Characteristic

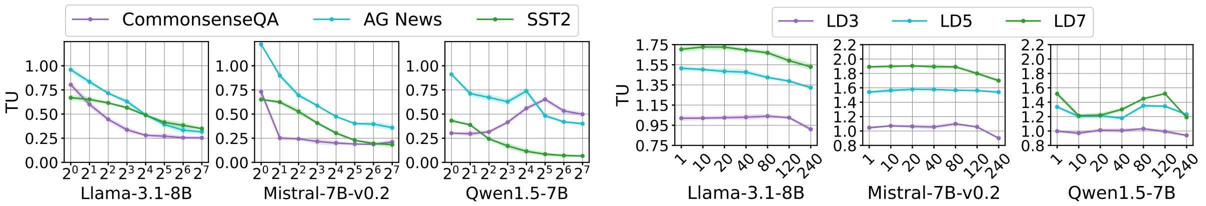


Figure 4: The average TU under k-shot ICL with error bands for three runs.

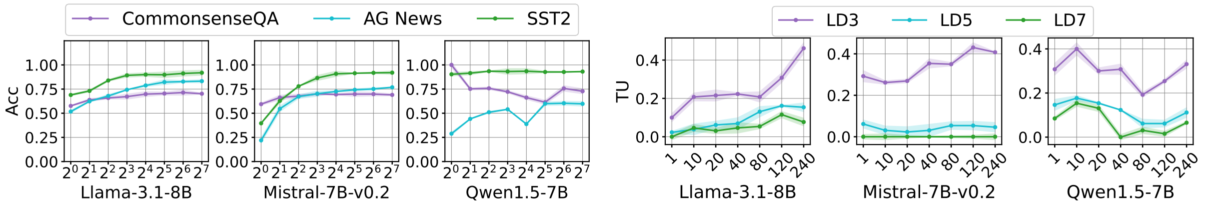


Figure 5: The average accuracy under k-shot ICL with error bands for three runs.

fidence as more in-context examples are provided, highlighting the correlation between improved confidence and performance gains for LLMs.

For *easy mode*, the inclusion of initial examples rapidly drives predictive entropy to a relatively low-uncertainty state, with further increases in examples yielding only marginal reductions in entropy (see Fig. 6 for a detailed view). In contrast, *hard mode* exhibits a distinct pattern. Predictive entropy remains higher in hard mode compared to easy mode due to the intrinsic complexity of the tasks, particularly those involving logical deduction with increasing object complexity ($TU_{LD3} < TU_{LD5} < TU_{LD7}$). Here, adding initial examples has minimal impact on entropy reduction until the number exceeds several hundred, at which point substantial performance gains emerge.

When demonstrations are incorporated, both Llama-3.1-8B and Mistral-7B-v0.2 exhibit consistent improvements in performance (↑) and reductions in uncertainty (↓). In contrast, Qwen1.5-7B demonstrates pronounced variability on datasets under hard mode, where fewer-shot ICL (e.g., 10-shot) achieves levels of confidence and accuracy comparable to certain many-shot settings (e.g., 240-shot). We term this phenomenon the "ICL sink", drawing analogies to sink patterns observed in attention mechanisms (Xiao et al., 2024). Notably, for Mistral models, even at the context limit in the 240-shot ICL setting on the LD7 dataset, Mistral sustains robust instruction-following and achieves performance comparable to Llama-3.1-8B, despite the latter's fourfold context-length capacity. This underscores the architectural strengths

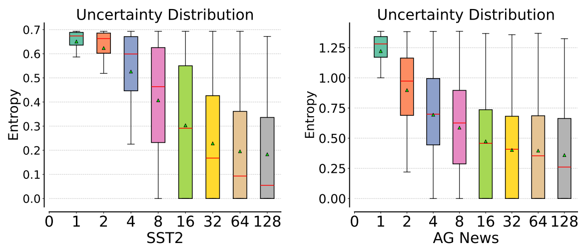


Figure 6: TU distribution of 2000 examples under certain k-shot ICL on AG News and SST2 datasets for Mistral-7B-v0.2.

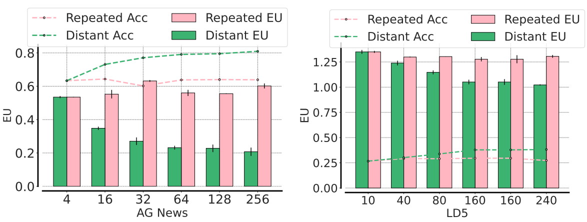


Figure 7: EU of Llama-3.1-8B on AG News and logical_deduction_five_objects datasets for distant examples vs. repeating 4/10 examples N times.

of Mistral, which leverages a sparse Mixture of Experts (MoEs) (Shazeer et al., 2017) and sliding window attention. *Thus, the influence of additional in-context examples on uncertainty fundamentally depends on the intrinsic long-context understanding capabilities.*

Micro View *An increasing number of examples effectively mitigates uncertainty for most questions.* Tables 2, 11, and 12 detail the percentage of questions exhibiting decreased or increased uncertainty under k-shot ICL. Despite 8.65% of cases experiencing heightened uncertainty with longer inputs in 128-shot learning, this effect minimally impacts overall model performance, as reflected by the small absolute values of ΔAcc. Crucially, the transition from few-shot (e.g., 4-shot) to many-shot ICL demonstrates a marked reduction in uncertainty for a larger proportion of questions, driving consistent performance improvements. These findings suggest that enhanced performance stems from increased confidence in the majority of questions.

Choices of k For practical applications, we recommend opting for a relatively larger k in in-context learning, as it simultaneously enhances performance and bolsters reliability.

Ablations with Model Size To further strengthen our analysis, we conducted additional experiments on the instruction-tuned versions of the more capable Qwen-2.5-14B and Qwen-2.5-32B models. The complete results are presented in Appendix C.

Across all uncertainty measures (TU, EU, and AU), larger models consistently exhibit substantially lower uncertainty values. For easy-mode, large LLMs follow similar uncertainty trends as smaller models; On more challenging tasks (hard mode), LLMs display distinct uncertainty patterns. Specifically, for Qwen-2.5-14B, EU steadily decreases as more demonstrations are provided, indicating more rapid task adaptation and improved performance, whereas AU remains relatively stable. Notably, a detailed analysis reveals that AU for Qwen-2.5-14B decreases slightly when initial examples are added but begins to rise beyond 80-shot, likely due to long-context effects introducing noise. In contrast, Qwen-2.5-32B does not exhibit this trend; instead, its AU continues to decrease as the number of examples increases.

Takeaways Large-scale LLMs exhibit greater confidence (i.e., lower uncertainty) and superior performance under many-shot settings, compared to smaller counterparts. The benefits of many-shot ICL remain evident, as additional demonstrations continue to enhance task-specific adaptation while maintaining low EU. Thus, the advantages of long-context IC, both in terms of performance and confidence, persist even at a larger scale.

4.3 RQ2: where do performance gains stem from?

In Sec. 4.2, we establish that reduced uncertainty improves performance. We hypothesize that additional examples in ICL foster a more refined task-specific conceptual framework, denoted as β, which empowers LLMs to approach novel problems x_{T+1} within the domain with increased confidence and efficacy. To validate this, we decompose total uncertainty into EU and AU, checking how these context helps LLMs to improve confidence by utilizing the definition and property of two special forms of uncertainty (Fig. 8).

Lower EU as the Primary Driver of TU Reduction. The decrease in TU is predominantly attributed to a decline in EU. Initially, EU accounts for the majority of TU, indicating that uncertainty primarily arises from the LLMs' insufficient in-domain knowledge, while their robust natural language understanding keeps AU relatively low. In simpler task settings, LLMs swiftly acquire task-specific knowledge, leading to a rapid decline in EU and sustaining consistently low AU. In contrast, for challenging tasks involving intricate logical structures, additional demonstrations may elevate AU (e.g., Llama-3.1-8B on the LD7 dataset), partially counteracting the reduction in EU and impeding significant decreases in total entropy. This underscores the persistent difficulty for current large models in effectively comprehending long texts with complex structures.

Additional Information Reduces EU. To validate that additional examples enhance the informational content and yield a clearer β for models (as shown in Fig. 7), we observe that only diverse examples effectively reduce EU under k-shot learning, whereas repetitive examples fail to achieve the same effect. This highlights that the true driver of uncertainty reduction lies in the increased informational richness of the examples provided.

5 Interpretability View for Uncertainty in K-shot ICL

To investigate the mechanisms by which increased in-context demonstrations reduce uncertainty in LLMs, we aim to delve into the models' internal states, unraveling the underlying processes governing answer selection and generation in in-context learning, thereby offering a comprehensive and interpretable analysis of this phenomenon.

5.1 Residual Stream Projection

Residual Streams Residual streams function as iterative refinements of feature representations in deep neural networks (He et al.; Li and Papyan, 2023), encapsulating the process of hierarchical in-

Dataset	8-shot		16-shot		32-shot		64-shot		128-shot	
	ΔU	ΔAcc	ΔU	ΔAcc	ΔU	ΔAcc	ΔU	ΔAcc	ΔU	ΔAcc
	Easy Mode									
AG News	66.8	**+7.3**	83.6	**+11.5**	88.6	**+13.9**	91.2	**+15.2**	90.8	**+15.8**
	30.45	**-1.0**	15.00	**-0.7**	10.75	**-0.2**	8.4	**-0.35**	8.65	**-0.4**
SST-2	71.7	**+5.7**	82.9	**+6.1**	86.6	**+6.6**	88.5	**+7.1**	92.1	**+7.9**
	20.3	**-0.5**	12.5	**-0.4**	9.4	**-0.4**	8.2	**-0.3**	5.6	**-0.3**
Commonsense QA	62.2	**+1.8**	69.8	**+4.2**	69.0	**+4.8**	78.6	**+6.6**	81.2	**+5.2**
	26.2	**-0.4**	18.8	**-0.2**	17.8	**-0.2**	16.8	**-1.0**	16.6	**-0.8**
	Hard Mode									
	20-shot		40-shot		80-shot		120-shot		240-shot	
	ΔU	ΔAcc	ΔU	ΔAcc	ΔU	ΔAcc	ΔU	ΔAcc	ΔU	ΔAcc
Logical Deduction3	45.6	**+6.0**	38.1	**+ 4.0**	40.4	**+2.0**	53.2	**+7.60**	62.3	**+15.91**
	44.8	**-5.6**	51.2	**-8.0**	46.8	**-10.8**	40.4	**-6.8**	34.5	**-5.9**
Logical Deduction5	58.4	**+2.8**	64.4	**+2.8**	73.6	**+4.8**	79.6	**+8.0**	83.8	**+10.8**
	33.2	**-0.4**	30.0	**-1.2**	24.4	**-0.8**	18.0	**-1.6**	13.1	**-1.5**
Logical Deduction7	48.4	**+2.0**	54.4	**+3.6**	59.2	**+4.4**	75.0	**+12.0**	83.3	**+12.3**
	48.8	**-1.2**	42.0	**-0.8**	38.0	**-0.8**	24.5	**-0.0**	15.3	**-0.7**

Table 2: ΔU refers to the proportion of datasets displaying either a decrease or increase in uncertainty relative to the 4-shot baseline, with $|\Delta U| > \tau$ indicating significant uncertainty changes. For each dataset, the first row presents the proportion of questions exhibiting reduced uncertainty, while the second row reflects those with increased uncertainty. ΔAcc quantifies the performance shift associated with the corresponding subset. Model: Llama-3.1-8B.

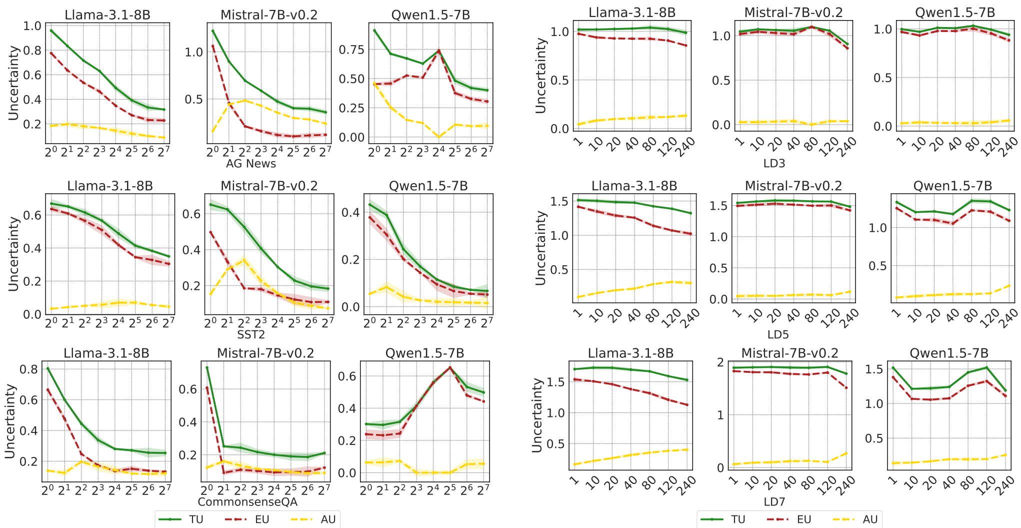


Figure 8: Uncertainty decomposition results for both easy mode (left) and hard mode (right).

formation aggregation. By leveraging residual connections, models reveal their mechanisms for constructing and iteratively refining outputs, thereby improving interpretability. Formally, in decoder-only LLMs, the hidden state of the i-th token at the l-th layer, denoted as $\mathbf{h}_i^{(l)}$, is computed as:

$$\begin{aligned}
\mathbf{h}_i^{(l)} &= \mathbf{h}_i^{(l-1)} + \mathbf{a}_i^{(l)} + \mathbf{m}_i^{(l)}, \\
\mathbf{a}_i^{(l)} &= \mathcal{MSHA}(\mathbf{h}_i^{(l-1)}), \\
\mathbf{m}_i^{(l)} &= \mathcal{MLP}\Big(\mathbf{h}_i^{(l-1)} + \mathbf{a}_i^{(l)}\Big),
\end{aligned}$$

where $\mathcal{MSHA}(\cdot)$ represents the multi-head self-attention mechanism (Vaswani, 2017), and $\mathcal{MLP}(\cdot)$ denotes the feed-forward neural network. For simplicity, detailed computations within the MHSA sublayer, such as the projection matrices $\mathbf{W}_{Q,K,V,O}$, and the splitting-merging operations across attention heads, are omitted here. Each decoder block, therefore, maintains two distinct residual pathways: one emerging from the MHSA, $\mathbf{h}_i^{(l)}$, and the other from MLP sublayer, $\mathbf{h}_i^{(l)} + \mathbf{a}_i^{(l)}$.

Projection into Vocabulary To uncover the latent information encoded within residual streams, projecting intermediate states onto a probability distribution over the vocabulary space V provides critical insights into the temporal and spatial dynamics of how these networks construct and refine their outputs (Geva et al., 2021; Belrose et al., 2023; Dar et al., 2023). Analogous to token generation, each residual stream $\mathbf{r}_i \in \mathbb{R}^d$ at the fi-

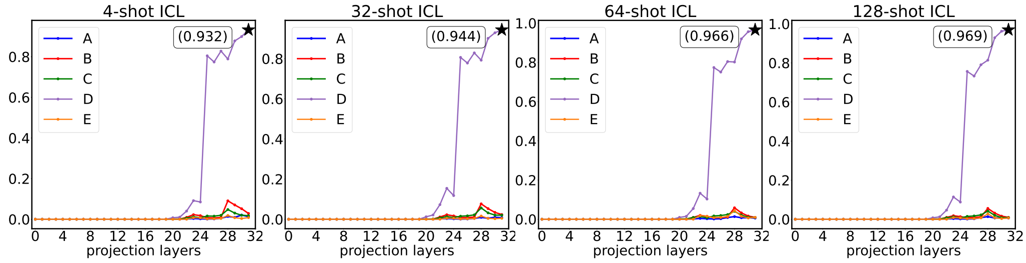


Figure 9: Average probabilities of Mistral-7B-v0.2 on the Commonsense QA dataset for MCQA items where the correct answer is "D". A 32-layer LM gets 64 residual streams, excluding the output hidden states.

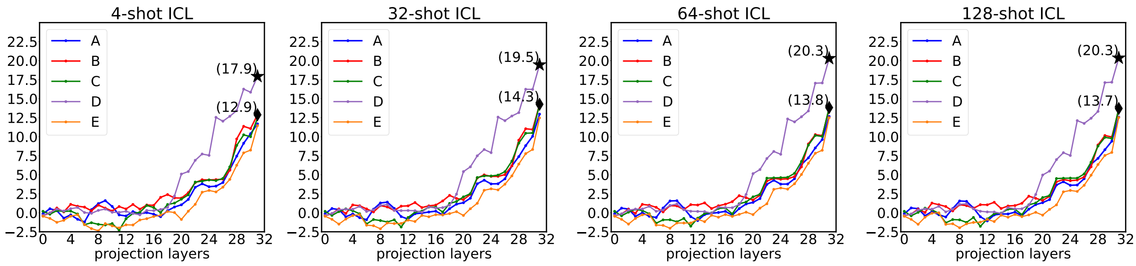


Figure 10: Average logits of Mistral-7B-v0.2 on the Commonsense QA dataset for MCQA items with the correct answer "D". Increasing in-context examples amplifies the logit of the correct option, thereby magnifying the difference between the logits of correct and incorrect options. ⋆ represent the highest logit and ⧫ the second highest logit. Refer to Appendix F.2 for additional results.

nal position—where i indexes the i-th residual in the model—undergoes transformation via an unembedding matrix $W_U \in \mathbb{R}^{d\times|V|}$ post layer normalization. This process yields calibrated logits $\mathbf{l}_i = W_U\mathbf{LayerNormalization}(\mathbf{r}_i)$ and the corresponding probabilities $\mathbf{p}_i = \mathbf{Softmax}(\mathbf{l}_i)$.

Correlation with Uncertainty in ICL For k-shot in-context learning, consider projecting the residual representations at the answer position into the probability simplex $\Delta^{|V|}$ over the vocabulary V. Denote the resulting logits and probabilities of candidate symbols (e.g., "A", "B", "C") as ℓ_i and p_i, respectively. These logits ℓ_i or probabilities p_i serve as proxies for confidence levels associated with each candidate. Analyzing the evolution of ℓ_i across model layers reveals the hierarchical development of inner confidence throughout k-shot learning, offering a profound understanding of the underlying uncertainty dynamics.

5.2 RQ3: What mechanisms underlie uncertainty reduction?

Qualitative Analysis To begin with, we present a case study in Fig. 12, offering an intuitive and qualitative demonstration of how the number of shots influences uncertainty. In this case, the Mistral-7B model struggles to distinguish the correct answer, option "E", under a 4-shot ICL setting, as the other options continuously mislead the model throughout the process. This is evidenced by the fluctuating confidence levels, which rise and fall erratically. In contrast, as the number of shots increases (32-, 64-, and 128-shot settings), many-shot ICL consistently boosts the probability of selecting "E" as the correct answer from about 22nd layer onward, maintaining this highest probability thereafter. Simultaneously, it demonstrates robustness by maintaining near-zero probabilities for incorrect options, effectively eliminating the influence of distractors on the model's final prediction.

CMQA	4-shot	32-shot	64-shot	128-shot
Llama-3.1	2.86 / 24.98	2.75 / 27.03	2.55 / 27.66	2.53 / 28.01
Mistral-v0.2	2.78 / 17.14	2.24 / 19.60	2.57 / 20.38	2.75 / 20.84
Qwen1.5	3.51 / 29.11	3.62 / 30.49	3.73 / 30.97	3.76 / 30.94
LD3	4-shot	40-shot	120-shot	240-shot
Llama-3.1	0.51 / 15.93	0.77 / 17.15	0.65 / 16.6	0.77 / 16.87
Mistral-7B-v0.2	0.26 / 11.07	0.48 / 11.92	0.46 / 12.05	0.59 / 11.87
Qwen1.5-7B	0.45 / 15.98	0.46 / 16.49	0.43 / 16.54	0.49 / 16.72

Table 3: Average logit difference / the largest logit.

Extended Examples Amplify Logit Disparity. We compute the average logits ℓ_i and probabilities p_i (Figs. 10 and 9) across varying shot counts for groups sharing the same answer. The analysis reveals that extended ICL enhances the precision of LLMs, concentrating greater logit mass on the correct symbol while effectively suppressing alternatives. This dynamic, driven by the interplay between an amplified logit disparity and increased absolute logit values (Table 3), leverages the exponential sensitivity of the $\mathrm{Softmax}$ function to propel the probability of the correct symbol toward 1. Consequently, **the uncertainty in LLM predictions is significantly reduced.**

6 Further Discussion

Clarifications While our work builds upon the framework in (Ling et al., 2024), our research specifically investigates the evolution of uncertainty in long-context ICL, a topic that has not been examined to date. In contrast, Ling et al. primarily focus on introducing a framework for decomposing uncertainty in few-shot ICL. By shifting the focus to long-context scenarios, our study explores how uncertainty evolves as the number of in-context examples increases, thereby addressing an important yet understudied dimension of ICL.

7 Conclusion

This study investigates the impact of extra demonstrations on the confidence of LLMs in their responses. Experimental results demonstrate that additional examples significantly reduce TU across both simple and complex tasks by integrating task-specific knowledge. This reduction is primarily attributed to decreased model uncertainty, which enhances overall performance. However, in complex tasks, many-shot ICL faces challenges in reducing TU due to a concurrent increase in AU. Analysis of the internal representations of LLMs reveals that many-shot ICL not only reallocates greater logit mass toward correct responses but also enlarges the logit margin between correct answers and distractors, reflecting an increase in model confidence.

Limitation

Our study is the first systematic investigation into uncertainty evolution in long-context ICL, addressing a critical research gap. These foundational experiments hope to provide a basis for future UQ studies on open-ended tasks. However, several limitations must be acknowledged.

Exclusion of Open-Ended Tasks The scope of this work does not encompass the uncertainty analysis of open-ended tasks, such as abstractive summarization (Hasan et al., 2021) and machine translation (Costa-jussà et al., 2022), owing to the lack of robust UQ techniques for free-form generative scenarios. Nevertheless, applying ICL to rationale-intensive reasoning and generative contexts remains a promising direction. Future investigations should assess the reliability and trustworthiness of ICL in these domains, as advancements in this area could not only enhance task-solving performance but also broaden the applicability of UQ methodologies to more diverse and complex settings.

Limited Exploration of ICL Configurations This study also excludes several influential ICL paradigms, such as unsupervised ICL (Yu et al., 2024), reinforced ICL (Jiang et al., 2024), and CoT prompting (Wei et al., 2022b), the latter of which is widely adopted in reasoning tasks to elicit step-by-step rationales. Existing UQ methods fall short of capturing the logical complexity intrinsic to reasoning-intensive contexts. Furthermore, practical challenges, including the computational overhead and context-length constraints of current open-source LLMs, prevented us from investigating extreme-shot ICL scenarios involving thousands of demonstrations. These limitations underscore promising directions for future research, particularly in applying UQ methodologies to better accommodate the unique challenges posed by reasoning tasks. More discussion in Appendix A.

Broader Impact

Despite these limitations, this study marks a pivotal advancement in understanding the reliability of ICL by harnessing recent breakthroughs in uncertainty quantification and decomposition, an essential yet underexplored aspect of LLM research. The research on uncertainty in ICL enriches the field of uncertainty quantification, providing novel perspectives on the trustworthiness of many-shot ICL. These contributions lay a solid foundation for broadening ICL's applicability in high-stakes domains. Ultimately, these findings could catalyze the development of more dependable and interpretable AI systems, offering profound societal impact.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 72293575, and in part by the Strategic Priority Research Program of Chinese Academy of Sciences under Grant XDA0480301 and the Excellent Youth Program of State Key Laboratory of Multimodal Artificial Intelligence Systems MAIS2024310.

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*.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Yavuz Faruk Bakman, Duygu Nur Yaldiz, Baturalp Buyukates, Chenyang Tao, Dimitrios Dimitriadis, and Salman Avestimehr. 2024. MARS: Meaning-aware response scoring for uncertainty estimation in generative LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7752–7767, Bangkok, Thailand. Association for Computational Linguistics.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Amanda Bertsch, Maor Ivgi, Uri Alon, Jonathan Berant, Matthew R. Gormley, and Graham Neubig. 2024. In-context learning with long-context models: An in-depth exploration. In *First Workshop on Long-Context Foundation Models @ ICML 2024*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

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.

Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. 2022. Data distributional properties drive emergent in-context learning in transformers. *Advances in neural information processing systems*, 35:18878–18891.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv e-prints*, pages arXiv–2207.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. 2023. Analyzing transformers in embedding space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16124–16170, Toronto, Canada. Association for Computational Linguistics.

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.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: core tasks, applications and evaluation. *J. Artif. Int. Res.*, 61(1):65–170.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. 2022. A survey of uncertainty in deep neural networks.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition.

Wenchong He, Zhe Jiang, Tingsong Xiao, Zelin Xu, and Yukun Li. 2023. A survey on uncertainty quantification methods for deep learning. *arXiv preprint arXiv:2302.13425*.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2024. Decomposing uncertainty for large language models through input clarification ensembling. In *Forty-first International Conference on Machine Learning*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Yixing Jiang, Jeremy Andrew Irvin, Ji Hun Wang, Muhammad Ahmed Chaudhry, Jonathan H Chen, and Andrew Y. Ng. 2024. Many-shot in-context learning in multimodal foundation models. In *ICML 2024 Workshop on In-Context Learning*.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. 2024. Llm maybe longlm: Self-extend llm context window without tuning.

Lingkai Kong, Harshavardhan Kamarthi, Peng Chen, B. Aditya Prakash, and Chao Zhang. 2023. Uncertainty quantification in deep learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 5809–5810, New York, NY, USA. Association for Computing Machinery.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Jianing Li and Vardan Papyan. 2023. Residual alignment: Uncovering the mechanisms of residual networks. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024. Long-context llms struggle with long in-context learning.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, Guangji Bai, Liang Zhao, and Haifeng Chen. 2024. Uncertainty quantification for in-context learning of large language models. 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 3357–3370, Mexico City, Mexico. Association for Computational Linguistics.

Xin Liu, Muhammad Khalifa, and Lu Wang. 2024. Litcab: Lightweight language model calibration over short- and long-form responses. In *The Twelfth International Conference on Learning Representations*.

Sewon Min, Xinxi 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.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Gv, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartłomiej Koptyra, Hayden Lau, Jiaju Lin, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Johan Wind, Stanisław Woźniak, Zhenyuan Zhang, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. 2023. RWKV: Reinventing RNNs for the transformer era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14048–14077, Singapore. Association for Computational Linguistics.

Joshua Robinson and David Wingate. 2023. Leveraging large language models for multiple choice question answering. In *The Eleventh International Conference on Learning Representations*.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2024. Code llama: Open foundation models for code.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks:

The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568(C).

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Matias Valdenegro-Toro and Daniel Saromo Mori. 2022. A deeper look into aleatoric and epistemic uncertainty disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1509–1517.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Yifei Wang, Yuheng Chen, Wanting Wen, Yu Sheng, Linjing Li, and Daniel Dajun Zeng. 2024. Unveiling factual recall behaviors of large language models through knowledge neurons. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7402, Miami, Florida, USA. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hannaneh Hajishirzi, and Ashish Sabharwal. 2024. Answer, assemble, ace: Understanding how transformers answer multiple choice questions.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Guoxin Yu, Lemao Liu, Mo Yu, Yue Yu, and Xiang Ao. 2024. Rethinking the evaluation of in-context learning for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14068–14082, Miami, Florida, USA. Association for Computational Linguistics.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024a. LUQ: Long-text uncertainty quantification for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5244–5262, Miami, Florida, USA. Association for Computational Linguistics.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024b. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

A Related Work

Overview. Estimating uncertainty in generation tasks presents greater challenges (Kuhn et al., 2023) compared to tasks with a predefined candidate set like classification tasks (Zhang et al., 2024b) and multiple-choice question answering (MCQA) (Robinson and Wingate, 2023). This is primarily due to the vast, high-dimensional semantic space inherent in natural language, which results in an effectively infinite generation space (Lin et al., 2024; Ling et al., 2024; Liu et al., 2024). In contrast, classification tasks provide LLMs with a finite set of discrete candidates, where the model's task is limited to selecting the most probable answer from a predefined set (Wiegreffe et al., 2024).

B Further Discussion

B.1 Limitations of UQ in Open-ended Tasks

UQ in open-ended tasks primarily focuses on knowledge-intensive QA tasks, which differs fundamentally from the typical ICL paradigm. ICL primarily relies on: pattern matching (Min et al., 2022); distribution alignment (Chan et al., 2022) ;implicit fine-tuning (Akyürek et al.). In contrast, knowledge-intensive QA depends on retrieving from external knowledge and parametric knowledge, rather than adapting through in-context distribution learning. As a result, many-shot ICL is not well-suited for knowledge-intensive QA scenarios, making existing UQ methods for this domain inapplicable. Moreover, prior research on the performance of many-shot ICL has primarily focused on reasoning tasks and extreme-label classification (Li et al., 2024), rather than knowledge-intensive tasks.

Challenges in Extending UQ to Open-ended Tasks Open-ended tasks encompass summarization, intermediate reasoning, code generation, program synthesis, and planning. However, existing UQ methods struggle to generalize effectively to these tasks, particularly in long-context ICL settings. For instance, semantic entropy (Kuhn et al., 2023), a widely used UQ approach, measures uncertainty based on semantic dispersion. However, in summarization tasks, summary quality is judged primarily by its fidelity to the source content, rather than semantic variability alone. This presents key limitations: A summary may deviate semantically yet still provide a valid abstraction of the original text. Summarization evaluation involves coverage, conciseness, and coherence, which semantic entropy alone cannot quantify. Given these limitations, we focus on classification and multiple-choice tasks, which offer a robust evaluation framework for analyzing uncertainty evolution in long-context ICL.

B.2 Limitations of UQ for CoT

In CoT tasks, uncertainty accumulates throughout the reasoning process, influencing the final answer. This uncertainty propagation occurs in intermediate reasoning steps, and the final answer generation. Current UQ techniques primarily focus on single-step inference or static tasks, whereas CoT relies on multi-step reasoning. This multi-stage nature makes it difficult for existing methods to effectively capture uncertainty propagation across reasoning steps. While research on CoT uncertainty is still in its early stages, some prior works have explored possible approaches. For instance, some work proposed a stepwise scoring mechanism which assigns a confidence score to each intermediate explanation. However, this approach has notable limitations:(1) *Overconfidence*: LLMs tend to be overconfident in their predictions, making single-step confidence scores unreliable; (2). *Lack of global coherence*: stepwise scoring ignores dependencies across reasoning steps, failing to capture uncertainty propagation across the entire reasoning chain; (3). *Step mismatch*: the reasoning steps generated may not align with the logical steps required for complex reasoning tasks, limiting the effectiveness in capturing uncertainty flow.

Potential Strategies: A Topological Perspective To better model uncertainty propagation in CoT reasoning, we propose leveraging topological structures. CoT reasoning typically involves problem decomposition, backtracking and correction, evaluation and verification, and final integration. While current models generate reasoning in an autoregressive (linear) manner, actual human reasoning follows a more complex topological structure. Inspired by Tree-based CoT (Yao et al., 2023) and Graph-based CoT (Besta et al., 2024), we propose modeling CoT uncertainty using graph or tree structures. In this framework: each reasoning step is represented as a node; uncertainty from prior steps propagates through the topological structure to influence subsequent steps; the final answer (root node) aggregates the propagated uncertainties from all previous steps. By explicitly modeling uncertainty flow in a structured manner, this approach could overcome the limitations of stepwise scoring and offer a systematic framework for analyzing uncertainty evolution in multi-step reasoning. We believe this direction holds promise for improving uncertainty estimation in CoT-based tasks.

C Generalization Results on Larger LLMs

C.1 Qwen2.5-14B-Instruct

AG_News	2	4	8	16	32	64	128	256
TU	0.148	0.127	0.113	0.125	0.115	0.105	0.086	0.065
EU	0.057	0.029	0.030	0.033	0.038	0.040	0.028	0.026
AU	0.091	0.098	0.083	0.092	0.077	0.065	0.058	0.039
ACC	87.6	87.2	88.9	88.19	88.7	88.5	89.9	90.5

Table 4: Performance of **Qwen2.5-14B-Instruct** on AG_News with varying numbers of in-context examples

LD5	1	10	20	40	80	120	240
TU	0.345	0.307	0.302	0.257	0.279	0.272	0.245
EU	0.229	0.194	0.190	0.148	0.157	0.139	0.124
AU	0.116	0.112	0.112	0.109	0.122	0.133	0.121
ACC	62.4	63.6	64.4	68.4	67.2	72.1	72.8

Table 5: Performance of **Qwen2.5-14B-Instruct** on LD5 with varying numbers of in-context examples

C.2 Qwen2.5-32B-Instruct

AG_News	2	4	8	16	32	64	128
TU	0.220	0.171	0.151	0.099	0.076	0.060	0.049
EU	0.151	0.093	0.059	0.030	0.017	0.020	0.018
AU	0.069	0.078	0.091	0.068	0.059	0.040	0.030
ACC	88.9	86.8	87.1	89.5	89.5	92.4	92.8

Table 6: Performance of **Qwen2.5-32B-Instruct** on AG_News with varying numbers of in-context examples

LD5	1	10	20	40	80	120
TU	0.353	0.313	0.284	0.247	0.225	0.202
EU	0.121	0.102	0.102	0.099	0.091	0.082
AU	0.232	0.210	0.182	0.147	0.134	0.120
ACC	74.8	79.6	82.0	82.4	83.6	84.1

Table 7: Performance of **Qwen2.5-32B-Instruct** on LD5 with varying numbers of in-context examples

D Quality of UQ for other LLMs

Dataset	Llama-3.1-8B							
	1-shot	2-shot	4-shot	8-shot	16-shot	32-shot	64-shot	128-shot
	Easy Mode							
AGNews	0.686	0.704	0.725	0.735	0.780	0.804	0.822	0.837
SST-2	0.714	0.751	0.751	0.748	0.740	0.741	0.742	0.750
Commonsense QA	0.563	0.599	0.636	0.673	0.726	0.774	0.784	0.798
	Hard Mode							
	1-shot	4-shot	10-shot	20-shot	40-shot	80-shot	120-shot	240-shot
Logical Deduction 3	0.973	0.965	0.939	0.948	0.951	0.939	0.966	0.947
Logical Deduction 5	0.996	0.995	0.963	0.983	0.971	0.983	0.959	0.974
Logical Deduction 7	0.987	0.997	0.976	0.987	0.982	0.986	0.986	0.964

Table 8: **AUROC** of Llama-3.1-8B model. High AUROC indicates the good quality of UQ measures.

Dataset	Mistral-7B-v0.2							
	1-shot	2-shot	4-shot	8-shot	16-shot	32-shot	64-shot	128-shot
	Easy Mode							
AGNews	0.633	0.696	0.734	0.753	0.769	0.778	0.790	0.780
SST-2	0.714	0.723	0.685	0.772	0.813	0.849	0.846	0.871
Commonsense QA	0.739	0.728	0.731	0.733	0.743	0.711	0.710	0.728
	Hard Mode							
	1-shot	4-shot	10-shot	20-shot	40-shot	80-shot	120-shot	240-shot
Logical Deduction 3	0.956	0.987	0.951	0.976	0.951	0.986	0.966	0.947
Logical Deduction 5	0.938	0.929	0.918	0.922	0.934	0.913	0.918	0.912
Logical Deduction 7	0.923	0.939	0.928	0.919	0.925	0.925	0.936	0.946

Table 9: **AUROC** of Mistral-7B-v0.2 model. High AUROC indicates the good quality of UQ measures.

Dataset	Qwen1.5-7B							
	1-shot	2-shot	4-shot	8-shot	16-shot	32-shot	64-shot	128-shot
	Easy Mode							
AGNews	0.634	0.716	0.743	0.744	0.688	0.739	0.731	0.741
SST-2	0.742	0.766	0.842	0.854	0.872	0.870	0.870	0.879
Commonsense QA	0.768	0.818	0.801	0.801	0.799	0.776	0.772	0.770
	Hard Mode							
	1-shot	4-shot	10-shot	20-shot	40-shot	80-shot	120-shot	240-shot
Logical Deduction 3	0.875	0.846	0.918	0.900	0.928	0.871	0.966	0.788
Logical Deduction 5	0.935	0.918	0.903	0.849	0.962	0.934	0.921	0.912
Logical Deduction 7	0.923	0.879	0.893	0.911	0.934	0.925	0.946	0.956

Table 10: **AUROC** of Qwen1.5-7B model. High AUROC indicates the good quality of UQ measures.

E Question-level Analysis

Dataset	Mistral-7B-v0.2									
	8-shot		16-shot		32-shot		64-shot		128-shot	
	ΔU	ΔAcc	ΔU	ΔAcc	ΔU	ΔAcc	ΔU	ΔAcc	ΔU	ΔAcc
	Easy Mode									
AG News	61.4	**+8.1**	70.9	**+9.5**	77.9	**+10.7**	76.6	**+11.4**	78.8	**+12.5**
	34.8	**-4.9**	25.1	**-4.2**	18.75	**-3.6**	19.15	**-3.4**	16.45	**-2.8**
SST-2	67.3	**+9.1**	79.2	**+12.8**	86.8	**+13.7**	87.9	**+13.8**	90.0	**+14.6**
	27.3	**-0.7**	16.3	**-0.0**	10.8	**-0.5**	9.7	**-0.3**	8.6	**-0.3**
Commonsense QA	49.8	**+1.8**	40.0	**+1.4**	37.2	**+3.4**	38	**+1.6**	36.0	**+2.0**
	21.8	**+0.4**	20.2	**+0.4**	18.4	**-1.4**	19.2	**+0.4**	23.6	**-0.8**
	Hard Mode									
	20-shot		40-shot		80-shot		120-shot		240-shot	
	ΔU	ΔAcc	ΔU	ΔAcc	ΔU	ΔAcc	ΔU	ΔAcc	ΔU	ΔAcc
Logical Deduction3	80.4	**+13.6**	78.4	**+20.4**	79.3	**+20.5**	76.8	**+18.4**	81.6	**+20.4**
	9.6	**-0.4**	9.2	**-0.8**	44.4	**-6.0**	12.4	**-0.4**	9.6	**-1.6**
Logical Deduction5	31.6	**+0.4**	39.6	**+0.8**	48.12	**+0.0**	57.9	**+0.0**	79.2	**+1.5**
	41.2	**-0.4**	36.8	**-0.8**	33.0	**-0.9**	24.0	**-0.8**	15.4	**-0.5**
Logical Deduction7	46.8	**+0.0**	60	**+0.0**	62.8	**+0.0**	47.2	**+0.0**	96.5	**+0.0**
	38.8	**-0.0**	25.2	**-0.0**	28.4	**-0.0**	34.8	**-0.0**	1.17	**-0.0**

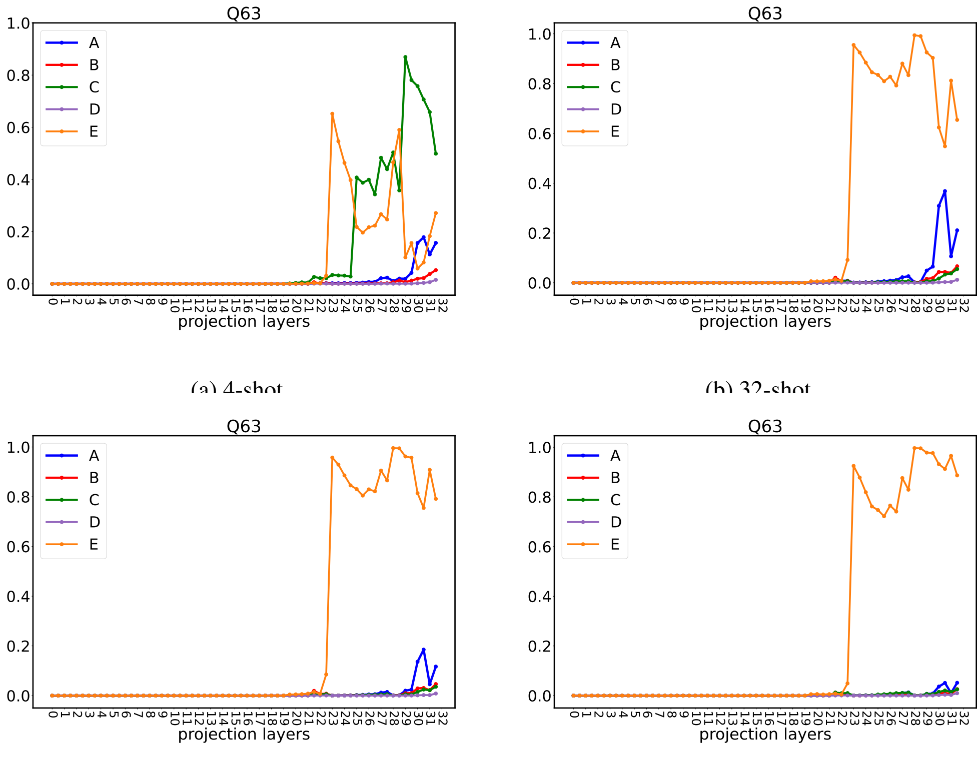
Table 11: Δ U denotes the proportion of datasets whose uncertainty decreases/increases compared to 4-shot settings, with the first line for each dataset giving the ratio of decreased uncertainty questions and the second line for each dataset giving the ratio of increased uncertainty questions. ΔAcc represents the performance changes caused by the corresponding part of examples.

Dataset	Qwen1.5-7B									
	8-shot		16-shot		32-shot		64-shot		128-shot	
	ΔU	ΔAcc	ΔU	ΔAcc	ΔU	ΔAcc	ΔU	ΔAcc	ΔU	ΔAcc
	Easy Mode									
AG News	62.2	**+3.7**	40.4	**-0.8**	79.5	**+9.3**	83.9	**+9.8**	83.8	**+10.0**
	34.4	**-0.8**	52.0	**-10.0**	17.25	**-0.6**	13.0	**-0.5**	12.6	**-1.4**
SST-2	78.0	**+0.2**	86.6	**+0.0**	82.9	**+0.1**	77.3	**-0.5**	71.8	**-0.3**
	13.0	**-0.5**	4.3	**-0.1**	2.9	**-0.5**	1.8	**-0.1**	2.4	**-0.3**
Commonsense QA	33.6	**+0.0**	12.2	**+0.4**	11.0	**+0.2**	8.9	**+0.81**	15.0	**+0.4**
	48.1	**-3.6**	68.6	**-10.4**	76.4	**-14.4**	79.7	**-1.6**	72.6	**-2.4**
	Hard Mode									
	20-shot		40-shot		80-shot		120-shot		240-shot	
	ΔU	ΔAcc	ΔU	ΔAcc	ΔU	ΔAcc	ΔU	ΔAcc	ΔU	ΔAcc
Logical Deduction3	13.6	**+2.0**	13.2	**+1.6**	9.2	**+0.4**	20.8	**+1.6**	36.0	**+6.0**
	84	**-13.6**	82.4	**-16.0**	88.4	**-22.8**	77.2	**-16.8**	60.8	**-10.4**
Logical Deduction5	73.6	**+7.19**	74.8	**+3.6**	50.8	**+0.0**	49.2	**+0.0**	69.2	**+6.4**
	24.0	**-2.8**	24.4	**-3.6**	47.2	**-12.0**	47.6	**-11.2**	15.4	**-5.6**
Logical Deduction7	81.2	**+4.8**	54.4	**+2.0**	44.0	**-0.4**	41.1	**-1.6**	52.5	**+10.3**
	17.6	**-0.8**	100	**-20.8**	54.8	**-11.6**	58.9	**-16.1**	45.3	**-9.2**

Table 12: Δ U denotes the proportion of datasets whose uncertainty decreases/increases compared to 4-shot settings, with the first line for each dataset giving the ratio of decreased uncertainty questions and the second line for each dataset giving the ratio of increased uncertainty questions. ΔAcc represents the performance changes caused by the corresponding part of examples.

F Interprebility for k-shot ICL

F.1 Case Study



(a) 4-shot (b) 32-shot

(a) 64-shot (b) 128-shot

Figure 12: The inner confidence changes (0-1 probability) of five options ["A", "B", "C", "D", "E"] for a specific question in Commonsense QA for Mistral-7B-v0.2 under 4-shot (a), 32-shot (b), 64-shot (c), and 128-shot(d) ICL. **The correct option is "E"** and LLMs only made a mistake under 4-shot ICL and got correct with more examples.

F.2 Additional results: Average logits and probabilities

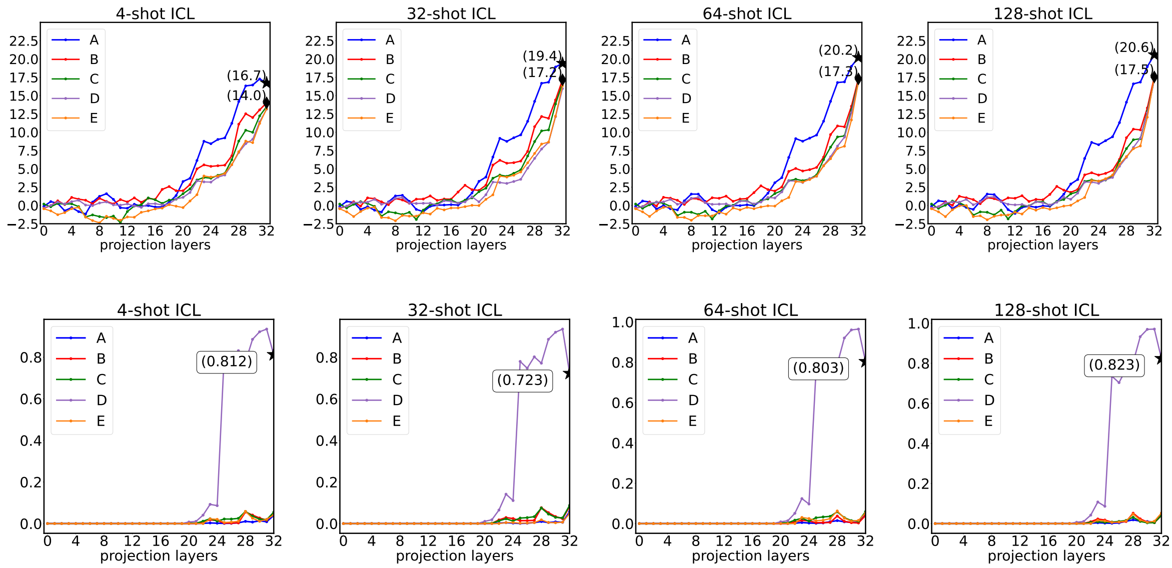


Figure 14: Average logits and probabilities of Mistral-7B-v0.2 on the Commonsense QA dataset for MCQA items where the correct answer is "A".

G AI Assistant Usage

We used *chatgpt* to assist with correcting spelling errors in writing .

H Experimental Details

H.1 Prompt templates

Classify the topic of the following sentence into four labels: [0: world, 1: sports, 2: business, 3: Sci/Tech]
Provide answer in a structured format WITHOUT additional comments, I just want the numerical label for each sentence.

Sentence: Fears for T N pension after talks Unions representing workers at Turner Newall say they are 'disappointed' after talks with stricken parent firm Federal Mogul.
Label: 2

Sentence: The Race is On: Second Private Team Sets Launch Date for Human Spaceflight (SPACE.com) SPACE.com - TORONTO, Canada -- A second\team of rocketeers competing for the #36;10 million Ansari X Prize, a contest for\privately funded suborbital space flight, has officially announced the first\launch date for its manned rocket.
Label: 3

Sentence: They've caught his eye In quote; helping themselves, quote; Ricky Bryant, Chas Gessner, Michael Jennings, and David Patten did nothing Friday night to make Bill Belichick's decision on what to do with his receivers any easier.
Label: 1

...

Sentence: Sister of man who died in Vancouver police custody slams chief (Canadian Press) Canadian Press - VANCOUVER (CP) - The sister of a man who died after a violent confrontation with police has demanded the city's chief constable resign for defending the officer involved.
Label:

Classify the topic of the following sentence into four labels: [0: world, 1: sports, 2: business, 3: Sci/Tech]
Provide answer in a structured format WITHOUT additional comments, I just want the numerical label for each sentence.

Figure 15: Prompt template with a test input for AG News dataset.

Classify the following sentence into two categories: [0: negative, 1: positive]
Provide answer in a structured format WITHOUT additional comments, I just want the numerical label for each sentence.

Sentence: that loves its characters and communicates something rather beautiful about human nature
Label: 1

Sentence: remains utterly satisfied to remain the same throughout
Label: 0

Sentence: on the worst revenge-of-the-nerds clichés the filmmakers could dredge up.
Label: 0

...

Sentence: that 's far too tragic to merit such superficial treatment
Label: 0

Sentence: very well-written and very well-acted .
Label: 1

Sentence: clumsy dialogue , heavy-handed phoney-feeling sentiment ,
Label: 0

Classify the following sentence into two categories: [0: negative, 1: positive]
Provide answer in a structured format WITHOUT additional comments, I just want the numerical label for each sentence.

Figure 16: Prompt template with a test input for SST-2 dataset.

Select the correct answer for the following commonsense question from five choices.
Provide answer in a structured format WITHOUT additional comments, I just want the option letter for each answer.

Question: The sanctions against the school were a punishing blow, and they seemed to what the efforts the school had made to change?
Choices
A. ignore
B. enforce
C. authoritarian
D. yell at
E. avoid
Answer: A

...

Question: To locate a choker not located in a jewelry box or boutique where would you go?
Choices
A. jewelry store
B. neck
C. jewelry box
D. jewelry box
E. boutique
Answer:

Select the correct answer for the following commonsense question from five choices.
Provide answer in a structured format WITHOUT additional comments, I just want the option letter for each answer.

Figure 17: Prompt template with a test input for Commonsense QA dataset.

Select the correct answer for the following logical deduction problem from three choices.
Provide answer in a structured format WITHOUT additional comments, I just want the option letter for each answer.

The following paragraphs each describe a set of three objects arranged in a fixed order. The statements are logically consistent within each paragraph. On a branch, there are three birds: a blue jay, a quail, and a falcon. The falcon is to the right of the blue jay. The blue jay is to the right of the quail.
Options:
(A) The blue jay is the second from the left
(B) The quail is the second from the left
(C) The falcon is the second from the left
Answer: (A)

...

The following paragraphs each describe a set of three objects arranged in a fixed order. The statements are logically consistent within each paragraph. On a shelf, there are three books: a blue book, an orange book, and a red book. The blue book is the rightmost. The orange book is the leftmost.
Options:
(A) The blue book is the second from the left
(B) The orange book is the second from the left
(C) The red book is the second from the left
Answer:

Select the correct answer for the following logical deduction problem from three choices.
Provide answer in a structured format WITHOUT additional comments, I just want the option letter for each answer.

Figure 18: Prompt template with a test input for logical deduction three objects dataset.

Select the correct answer for the following logical deduction problem from five choices.
Provide answer in a structured format WITHOUT additional comments, I just want the option letter for each answer.
Problem: The following paragraphs each describe a set of five objects arranged in a fixed order. The statements are logically consistent within each paragraph. On a branch, there are five birds: a quail, an owl, a raven, a falcon, and a robin. The owl is the leftmost. The robin is to the left of the raven. The quail is the rightmost. The raven is the third from the left.
Options:
(A) The quail is the rightmost
(B) The owl is the rightmost
(C) The raven is the rightmost
(D) The falcon is the rightmost
(E) The robin is the rightmost
Answer: (A)

...

Problem: The following paragraphs each describe a set of five objects arranged in a fixed order. The statements are logically consistent within each paragraph. In an antique car show, there are five vehicles: a hatchback, a bus, a convertible, a tractor, and a minivan. The tractor is older than the bus. The minivan is newer than the bus. The hatchback is the second-newest. The minivan is older than the convertible.
Options:
(A) The hatchback is the second-oldest
(B) The bus is the second-oldest
(C) The convertible is the second-oldest
(D) The tractor is the second-oldest
(E) The minivan is the second-oldest
Answer:

Select the correct answer for the following logical deduction problem from three choices.
Provide answer in a structured format WITHOUT additional comments, I just want the option letter for each answer.

Figure 19: Prompt template with a test input for logical deduction five objects dataset.

Select the correct answer for the following logical deduction problem from seven choices.
Provide answer in a structured format WITHOUT additional comments, I just want the option letter for each answer.
Problem: The following paragraphs each describe a set of seven objects arranged in a fixed order. The statements are logically consistent within each paragraph. In a golf tournament, there were seven golfers: Ana, Eve, Ada, Dan, Rob, Amy, and Joe. Dan finished third. Ana finished above Ada. Amy finished last. Dan finished below Rob. Eve finished below Ada. Rob finished below Joe.
Options:
(A) Ana finished third
(B) Eve finished third
(C) Ada finished third
(D) Dan finished third
(E) Rob finished third
(F) Amy finished third
(G) Joe finished third
Answer: (D)

...

Problem: The following paragraphs each describe a set of seven objects arranged in a fixed order. The statements are logically consistent within each paragraph. In an antique car show, there are seven vehicles: a bus, a motorcyle, a hatchback, a station wagon, a minivan, a truck, and a limousine. The station wagon is the fourth-newest. The motorcyle is newer than the truck. The station wagon is older than the hatchback. The minivan is newer than the hatchback. The bus is newer than the minivan. The truck is newer than the limousine.
Options:
(A) The bus is the third-oldest
(B) The motorcyle is the third-oldest
(C) ...
Answer:
Select the correct answer for the following logical deduction problem from seven choices.
Provide answer in a structured format WITHOUT additional comments, I just want the option letter for each answer.

Figure 20: Prompt template with a test input for logical deduction seven objects dataset.