# Why and How LLMs Benefit from Knowledge Introspection in Commonsense Reasoning

Anonymous ACL submission

## Abstract

Large Language Models (LLMs) can improve commonsense reasoning through generating intermediate knowledge. However, the effectiveness of this knowledge introspection is not always guaranteed. This paper first systematically investigates and reveals an **introspection paradox**: while simple introspection tends to benefit weaker models, it often degrades the performance of stronger ones, particularly on simpler tasks. Our deep analysis indicates that this paradox arises from a complex interplay among model capability, task difficulty and the quality of generated knowledge. Further interpretability analysis reveals the origins of low-quality knowledge generation. To better employ introspected knowledge in LLM, this paper proposes a training-free **Adaptive Introspection Strategy** that operates in two stages using only the model's internal states: **Knowledge Detection**, which dynamically identifies and discards potentially low-quality knowledge, and **Knowledge Regeneration**, which employs attention smoothing to guide the model away from harmful failure modes during knowledge generation. Extensive experiments on five Llama models with different sizes and eight commonsense reasoning benchmarks demonstrate that our approach effectively mitigates the limitations of standard introspection and has consistent performance gains across almost all settings.

## 1 Introduction

Large Language Models (LLMs) have achieved remarkable progress across a wide range of tasks. Techniques such as Chain-of-Thought (CoT) (Wei et al., 2022; Kojima et al., 2022) and Long CoT (OpenAI, 2024; DeepSeek-AI, 2025), which prompt the model to generate intermediate reasoning steps before producing a final answer, have been proven to be particularly effective in complex reasoning tasks such as mathematical problem solving and code generation (Qwen, 2025; Liao et al., 2025a). However, recent research suggests that beyond mathematical or logical domains, CoT offers limited benefits and even impairs performance for knowledge-intensive tasks such as commonsense reasoning (Kambhampati et al., 2024; Liu et al., 2024; Zheng et al., 2025).

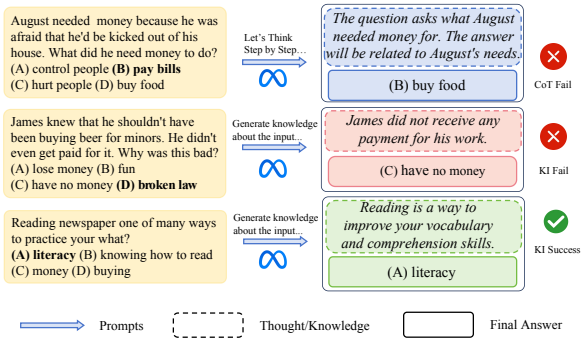



Figure 1: Unlike Chain-of-Thought (CoT) reasoning (top), Knowledge Introspection (KI) outputs relevant knowledge rather than the thinking process before generating the final answer. This makes KI more suitable for knowledge-intensive tasks. However, its effectiveness is not guaranteed, as the quality of the generated knowledge can significantly impact performance.

In fact, numerous studies (Xu et al., 2024; Yao et al., 2023; Tang et al., 2023; Liao et al., 2025b) suggest that the failure of such LLMs in knowledge-intensive tasks is primarily due to the improper activation of relevant internal knowledge during inference, rather than a lack of the required knowledge. Therefore, inspired by CoT-style prompting, a growing line of work explores knowledge introspection (KI, as shown in Figure 1) (Liu et al., 2022b,a, 2023; Molfese et al., 2024)—a process in which the model is guided to generate relevant supporting knowledge before providing a final answer. Different from previous superficially generating intermediate texts like CoT, introspection focuses on eliciting implicit relevant knowledge, including facts, concepts, etc., for explicit grounding. These approaches aim to bridge the gap between knowledge storage and utilization in LLMs, offering a

promising direction for improving LLM reasoning in knowledge-intensive tasks such as commonsense reasoning.

Nevertheless, this paper raises important questions: *Is knowledge introspection truly effective? For which models does it help? Under what task conditions does it succeed or fail?* To explore these questions, the paper conducts a preliminary analysis on commonsense reasoning tasks and reveals an introspection paradox. Surprisingly, the analysis results show that knowledge introspection does not always improve the reasoning performance. In fact, it is often beneficial for weaker models, which is consistent with the conclusion of prior work (Liu et al., 2022b). However, it also degrades the performance of stronger models, particularly on relatively simple tasks. This contradicts the intuitive assumption that more capable models, equipped with richer internal knowledge, should perform better on reasoning tasks (Liu et al., 2023; Berti et al., 2025).

To better understand this counterintuitive phenomenon, this paper carries out more comprehensive experiments and analyses to uncover when and why knowledge introspection (KI) helps or hurts the reasoning performance. Our analysis reveals a nuanced interplay among model capability, task difficulty, and the quality of generated knowledge, and has several important observations. **First**, as model capability increases, the gains from KI diminish, and the risk of performance drops due to low-quality knowledge grows. **Second**, harmful knowledge notably increases prediction uncertainty for stronger models and leads to performance degradation. **Third**, introspection becomes more useful on harder tasks, especially when the model cannot answer directly. It is because the proportion of helpful knowledge will increase when the task becomes difficult or complex. **Finally**, harmful knowledge generation is linked to an over-reliance on localized context. Attribution analysis shows higher focus and more concentrated attention during such cases, shedding light on the roots of these failures.

To address the problem of the aforementioned introspection paradox, the paper proposes a training-free adaptive knowledge introspection framework that dynamically adapts and refines the use of introspective knowledge according to the characteristics of both the task and the model. Specifically, our method consists of two stages, both leveraging the model's internal states without requiring additional training: (1) **Knowledge Detection**: Identifies and discards low-quality knowledge based on the interplay between model capability and task difficulty. (2) **Knowledge Regeneration**: Replaces discarded knowledge through refining the attention distributions. In this way, LLMs are encouraged to integrate broader contextual information for improved knowledge generation. The experiments on 5 LLMs with different sizes and 8 commonsense reasoning tasks show the consistent improvements and demonstrate the robustness of the proposed adaptive strategy.

The main contributions are as follows:

- The paper empirically identifies the introspection paradox of LLMs on the commonsense reasoning tasks. Our findings demonstrate that the effectiveness of introspection is not always guaranteed and varies across different scenarios.
- The paper systematically analyzes the underlying reasons. It is attributed to the critical interplay among task difficulty, model capability and the quality of generated knowledge.
- The paper proposes an adaptive introspection exploitation strategy, including Knowledge Detection and Knowledge Regeneration, which are solely based on model internal states to dynamically modulate the use of introspection.

## 2 Related Work

LLMs often benefit from relevant knowledge in knowledge-intensive reasoning tasks, whether from external or internal sources. To this end, various strategies have been proposed for integrating knowledge into LLMs.

**LLMs Augmented with External Knowledge** A significant body of work focused on augmenting LLMs with structured or unstructured external knowledge (Kaur et al., 2022; Wang et al., 2025). This includes integrating structured knowledge graphs like ConceptNet (Speer et al., 2017) and ATOMIC (Sap et al., 2019). In this context, graph neural network (GNN) based methods such as KagNet (Lin et al., 2019), QA-GNN (Yasunaga et al., 2021), and GreaseLM (Zhang et al., 2022) aimed to guide and augment reasoning by utilizing the encoded relations among knowledge units (e.g., entities). Another prominent line of work is retrieval-augmented generation (RAG) (Lewis et al., 2020), which retrieved relevant textual passages or facts to support generation for commonsense reasoning tasks (Yu et al., 2022). Despite

their effectiveness, the performance of such approaches depends heavily on the quality and coverage of external knowledge resources (Ma et al., 2019; Mitra et al., 2020; Talmor et al., 2021). Moreover, they often necessitate extra infrastructure, training, or fine-tuning, without directly leveraging the extensive internal knowledge embedded within LLM parameters. In contrast to these methods that depend on external modules, our work investigates a complementary approach—one that seeks to harness the internal knowledge inherently encoded in LLMs through pretraining, thereby offering a lightweight yet effective alternative.

**LLMs Augmented with Internal Knowledge** Recognizing that LLMs implicitly store substantial world knowledge (Davison et al., 2019; Jiang et al., 2020), another line of research focused on eliciting and leveraging the internal knowledge for reasoning tasks (Tang et al., 2023; Liao et al., 2025b). Prior efforts have largely focused on explicitly supervising LMs to generate commonsense knowledge (Bosselut et al., 2019; Zhou et al., 2021). These methods relied on curated external knowledge sources and structured generation objectives. By contrast, later introspective approaches sought to activate and refine the model's internal knowledge through self-supervised learning and reinforcement mechanisms (Liu et al., 2022a, 2023). Other lines of work have explored the generation of intermediate textual representations during inference via prompting strategies. These include template-based approaches such as Self-Talk (Shwartz et al., 2020), few-shot prompting methods like Generated Knowledge Prompting (GKP) (Liu et al., 2022b), and techniques leveraging auxiliary pretrained models (Bosselut et al., 2021). Alternatively, ZEBRA (Molfese et al., 2024) retrieved relevant examples to effectively augment the generation of knowledge.

While promising, existing approaches typically lack mechanisms to dynamically validate the generated knowledge or control the introspection process. In contrast, our work systematically analyzes when and why introspection improves or impairs performance. We further propose an adaptive framework that can dynamically control this process for off-the-shelf LLMs, which is distinguished from existing introspection strategies.

## 3 Why Introspection Succeeds or Fails

As mentioned above, knowledge introspection (KI) has been widely proposed as a general-purpose technique for enhancing LLM reasoning—particularly in commonsense reasoning tasks. However, we still wonder whether it has universal efficacy in all scenarios. In this section, we conduct a systematic investigation into the following research issues: (1) quantify the impact of KI across diverse model capabilities and task difficulties, and (2) analyze the underlying mechanisms governing knowledge generation. Through carefully controlled experiments, the paper challenges the prevailing assumption of the universal effectiveness of knowledge introspection. And it also reveals that the effectiveness of KI is highly contingent on an intricate interplay between model capability and task difficulty.

### 3.1 Introspection Paradox

In this subsection, we perform a comprehensive evaluation to verify the effectiveness of knowledge introspection (KI) on different scenarios, including (1) LLMs of varying capabilities (from 7B to 70B parameters), and (2) commonsense reasoning tasks spanning different complexity levels (from fact retrieval to advanced reasoning).

#### 3.1.1 Experimental setup

We evaluate Llama family models (Touvron et al., 2023; Grattafiori et al., 2024) on eight diverse commonsense reasoning datasets (see details in Appendix A and D). Specifically, two primary prompting conditions are used:

- **Direct Answer**: Standard zero-shot prompting where the model directly outputs the final answer.
- **Answer with KI**: First generate relevant knowledge and then provide the final answer based on the explicit knowledge.

#### 3.1.2 Introspection is not always beneficial

Table 1 shows the comparison results of **Answer with KI** and **Direct Answer**. From the table, we have the following important observations:

- **Observation 1: Benefit for Weaker Models.** Introspection consistently yields substantial performance gains for weaker models, exemplified by average improvements of +3.48, +3.50, and +1.79 for Llama-2-7b-Chat, Llama-2-13b-Chat, and Llama-2-70b-Chat respectively across most tasks.
- **Observation 2: Detriment for Stronger Models.** However, the effect is markedly different for more capable models like Llama-3-70B-Instruct (-2.37) and Llama-3-8B-Instruct (-0.07),

| Model | ARC.E | ARC.C | CSQA | CSQA2 | OBQA | PIQA | QASC | WG | Avg. All |
|---|---|---|---|---|---|---|---|---|---|
| | 97.69 | 92.83 | 81.82 | 75.91 | 91.60 | 89.28 | 90.10 | 69.46 | 86.09 |
| Llama-3-70B-Instruct | **92.47** (-5.22) | **89.59** (-3.24) | **79.85** (-1.97) | **76.78** (+0.87) | **84.20** (-7.40) | **86.18** (-3.10) | **82.83** (-7.27) | **77.90** (+8.44) | **83.72** (-2.37) |
| | 92.09 | 78.58 | 75.02 | 63.64 | 77.00 | 80.47 | 80.13 | 57.85 | 75.60 |
| Llama-3-8B-Instruct | **91.41** (-0.68) | **78.50** (-0.08) | **74.17** (-0.85) | **66.12** (+2.48) | **76.20** (-0.80) | **75.57** (-4.90) | **80.56** (+0.43) | **61.72** (+3.87) | **75.53** (-0.07) |
| | 85.82 | 73.98 | 72.65 | 60.02 | 74.40 | 79.43 | 71.60 | 52.01 | 71.24 |
| Llama-2-70b-Chat | **89.52** (+3.70) | **76.88** (+2.90) | **74.69** (+2.04) | **63.75** (+3.73) | **74.20** (-0.20) | **78.24** (-1.19) | **73.00** (+1.40) | **53.99** (+1.98) | **73.03** (+1.79) |
| | 79.29 | 61.95 | 63.55 | 56.51 | 59.80 | 77.15 | 60.58 | 52.64 | 63.94 |
| Llama-2-13b-Chat | **82.28** (+2.99) | **65.96** (+4.10) | **65.77** (+2.22) | **62.81** (+6.30) | **67.60** (+7.80) | **76.22** (-0.93) | **63.82** (+3.24) | **55.09** (+2.45) | **67.44** (+3.50) |
| | 71.42 | 53.58 | 52.25 | 52.03 | 51.40 | 60.66 | 43.41 | 50.83 | 54.45 |
| Llama-2-7b-Chat | **74.75** (+3.33) | **57.85** (+4.27) | **59.21** (+6.96) | **54.35** (+2.32) | **55.60** (+4.20) | **60.72** (+0.06) | **48.81** (+5.40) | **52.17** (+1.34) | **57.93** (+3.48) |

Table 1: Accuracy comparison: **Direct Answer** vs. **Answer with KI**. Each cell shows: Direct Answer Score (top line); Answer with KI Score and Accuracy Change (bottom line). Positive values (green) indicate improvement with introspection, while negative values (red) indicate performance degradation.

which frequently experience performance degradation. This phenomenon is particularly pronounced on simpler tasks such as ARC-Easy, ARC-Challenge, and PIQA.

- **Observation 3: Benefit for Harder Tasks.** Finally, these stronger models can still benefit from introspection, particularly on more complex tasks (e.g., CSQA2, WG).

These findings demonstrate that the benefits of knowledge introspection are highly context-dependent rather than universally applicable. Crucially, its effectiveness emerges from a nuanced interaction between model capability and task difficulty—a relationship that demands systematic examination. This insight compels us to investigate the fundamental mechanisms governing when and why introspection succeeds or fails. In the following section, we will conduct a systematic investigation through two complementary approaches: (1) *Quantitative statistical analysis* of the success and failure of KI, and (2) *Interpretability analysis* of intermediate states, aiming to comprehensively uncover the underlying mechanisms.

### 3.2 Quantitative Statistical Analysis of the Success and Failure of KI

To deeply understand the introspection paradox observed in Section 3.1, we analyze the quality of the generated knowledge and verify its subsequent impact on reasoning. In specific, according to the factual correctness and the relevance to the problem, the generated knowledge is identified as *Useful* and *Harmful* types (details in Appendix B).

Formally, let $C_I$ and $C_D$ be the events of correct answers for **Answer with KI** and **Direct Answer**, respectively. And $K_{Useful}$, $K_{Harmful}$ are the events of useful and harmful knowledge generation, respectively. Two metrics are used to quantify the potential influence associated with $K_{Useful}$, $K_{Harmful}$, respectively:

- **Gain Rate (GR):** The proportion of samples where useful knowledge helps correct a direct answering error: GR $= P(\neg C_D \wedge C_I | K_{Useful})$
- **Risk Rate (RR):** The proportion of samples where harmful knowledge causes a correct direct answer to become incorrect: RR $= P(C_D \wedge \neg C_I | K_{Harmful})$

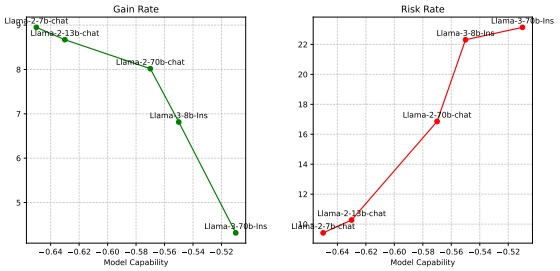



Figure 2: Gain Rate and Risk Rate trends across different Model Capabilities.

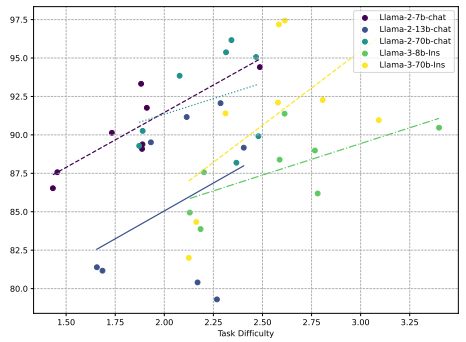



Figure 3: Percentage of generated useful knowledge when direct answering fails, with a regression line for each model.

This analysis aims to reveal the distinct trends in knowledge generation quality relative to model capability. Figure 2 shows that as model capability increases, GR has a generally downward trend while RR displays an upward trend. Stronger models exhibit reduced gains but increased risks relative to weaker ones. This result aligns with the aforementioned **Observation 1** and **Observation 2** and suggests that stronger models gain less from useful knowledge and are more vulnerable to harmful knowledge generated through introspection.

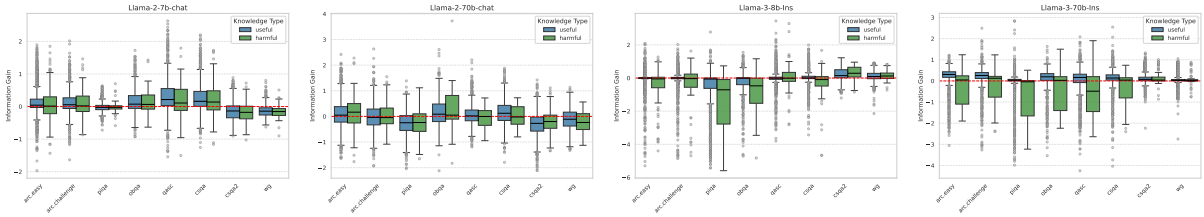



Figure 4: Information Gain across four models on different tasks.

Furthermore, we investigate how task difficulty influences the quality of generated knowledge, particularly in scenarios where introspection is most needed (i.e., when direct answering fails). Specifically, we measure the percentage of generated useful knowledge given that the direct answer is incorrect, formally defined as $P(K_{Useful}|\neg C_D)$. Figure 3 shows that this percentage tends to increase as task difficulty increases across different models. Their regression lines exhibit a positive slope, respectively. This phenomenon supports **Observation 3** that when tasks are harder, LLMs are more likely to generate useful knowledge through introspection. Such a finding is more obvious in difficult instances.

To further quantify the impact on the model's prediction confidence, we measure the Information Gain (IG) derived from introspection and offer another perspective on **Observation 1** and **Observation 2**: $IG = H(A|Q) - H(A|Q,K)$. Here, $H(\cdot)$ represents Shannon entropy. $A$, $Q$ and $K$ denote the predicted answers, question and knowledge. Figure 4 shows that introspection tends to decrease uncertainty for weaker models but increase uncertainty for stronger models, especially when generating harmful knowledge. It means that introspection often introduces conflicting signals for stronger models, rather than providing clear grounding.

### 3.3 Interpretability Analysis of the Success and Failure of KI

To understand the mechanisms of knowledge introspection, particularly the origins of generating harmful knowledge, we employ attribution tracing (Hao et al., 2021; Dai et al., 2022; Li et al., 2024) to quantify the influence of the input question context $q$ on the generation of the knowledge $k$.

Since the attention module involves interactions between different tokens, we compute the attribution score matrix for the $h$-th attention head in layer $l$, i.e. $\text{Attr}(A_h^{(l)})$, via Riemann approximation of the integration. Here, $m$ is the number of approximation steps (Sundararajan et al., 2017):

$$\begin{aligned}\text{Attr}(A_h^{(l)}) &= A_h^{(l)} \odot \int_{\alpha=0}^{1} \frac{\partial F(\alpha A_h^{(l)})}{\partial A_h^{(l)}} d\alpha \\ &\approx A_h^{(l)} \odot \left( \frac{1}{m} \sum_{s=1}^{m} \frac{\partial F(\frac{s}{m} A_h^{(l)})}{\partial A_h^{(l)}} \right)\end{aligned} \tag{1}$$

where $\odot$ denotes element-wise multiplication and $F(\cdot)$ represents the model's output. Each element $[\text{Attr}(A_h^{(l)})]_{i,j}$ represents the attribution of the $i \to j$ token interaction for head $h$ in layer $l$.

To obtain the total information flow from the question $q$ to the knowledge $k$ within layer $l$, we aggregate scores across all $H$ heads and relevant token pairs following (Hao et al., 2021; Li et al., 2024):

$$\text{Attr}^{(l)}(q \to k) = \sum_{(i,j) \in C_{qk}} \left( \sum_{h=1}^{H} \left| [\text{Attr}(A_h^{(l)})]_{i,j} \right| \right) \tag{2}$$

where $C_{qk} = \{(i,j) | q_s \le i \le q_e, k_s \le j \le k_e\}$ includes pairs with token $i$ in the question and token $j$ in the knowledge statement. We sum the absolute values across all $H$ attention heads to get the final score. By comparing $\text{Attr}^{(l)}(q \to k)$ values for useful and harmful knowledge, we assess how question information is leveraged during introspection.

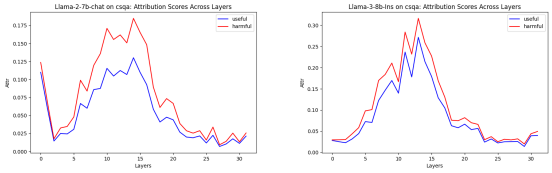



Figure 5: Layer-wise attribution scores from question context to generated knowledge on the CSQA.

Figure 5 illustrates the experimental results of Llama-2-7b-Chat and Llama-3-8B-Instruct on CSQA (full results shown in Appendix C.1). Both models exhibit consistently higher attribution scores from the question context to generated knowledge when producing harmful knowledge (red line) compared to the useful one (blue line). This difference is particularly pronounced in the intermediate layers, which are crucial for the model

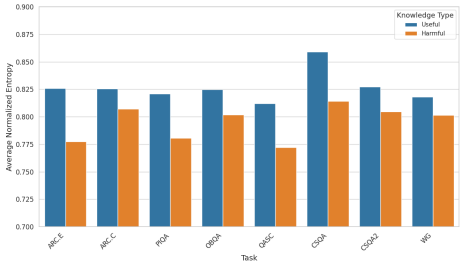



Figure 6: Average normalized entropy of attribution scores for Llama-3-8B-Instruct across different tasks.

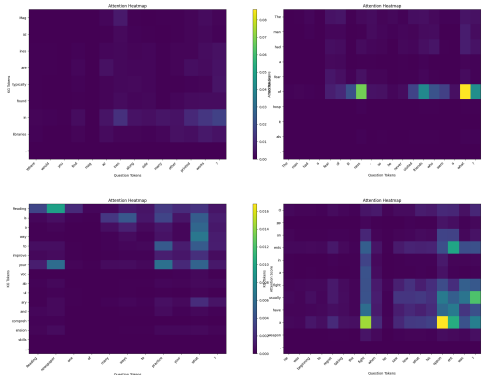


Figure 7: Attention attribution heatmaps in intermediate layer from question context tokens (X-axis) to generated knowledge tokens (Y-axis) for Llama-2-7b-chat on CSQA examples. The left column illustrates examples of useful knowledge generation, and the right column illustrates harmful ones.

to extract contextual information. This observation reveals a failure mode where the generation of harmful knowledge might be linked to an intensified, yet misguided focus on the context.

To further investigate this, we analyze the distribution of the attribution scores matrix. Specifically, we calculate the normalized Shannon entropy of the attribution matrix (details on the calculation and full results in Appendix C.2). Figure 6 shows that harmful knowledge exhibits lower normalized entropy than useful knowledge. This lower entropy suggests that the influence is concentrated on fewer specific (question token, knowledge token) pairs, reflecting a more peaked distribution. We also conduct case studies to visualize the attribution heatmaps (Figure 7). From the heatmaps of harmful knowledge, we observe exceptionally high attribution scores concentrated on specific question tokens when generating some knowledge tokens. It means that the model is overly dependent on localized cues from the question context. Conversely, the attribution heatmaps for useful knowledge tend to display a smoother pattern. It indicates that the influence from the question context is more evenly distributed across relevant semantic parts.

# 4 How to Modulate and Enhance Introspection

Based on our analysis in section 3, this paper proposes an **Adaptive Introspection** strategy (Figure 8) to exploit introspected knowledge well. It consists of two stages: knowledge detection and knowledge regeneration, after the original knowledge generation.

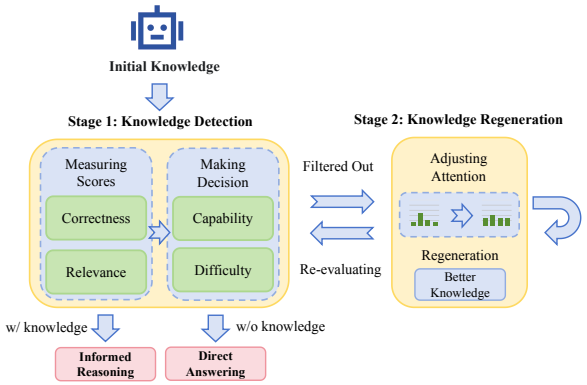



Figure 8: Workflow of the proposed Adaptive Introspection Strategy. It dynamically filters low-quality initial knowledge and optionally regenerates knowledge before providing the final answer.

## 4.1 Stage 1: Knowledge Detection

**Goal:** Selectively identify and remove potentially low-quality knowledge before it influences the final answer.

**Method:** We assess each generated knowledge statement along the two dimensions used in our annotation: *Correctness* and *Relevance*. Specifically, we employ the following solution to compute these two metrics (details in Appendix D):

- *Correctness:* We use the prediction entropy of the generated knowledge as a proxy for the model's confidence. Higher entropy suggests lower confidence, indicating uncertainty or factual incorrectness.
- *Relevance:* We measure the cosine similarity between the hidden states of the generated knowledge and the question context for contextual relevance.

We normalize these scores using z-score normalization. Knowledge pieces with scores under a dynamic threshold related to model capability and task difficulty are filtered out.[1]

**Rationale:** This stage leverages our findings in Section 3.2 that stronger models exhibit a higher

[1]In this work, we consider the linear relationships based on the observations in Section 3.2 and empirically calibrate the threshold by linearly combining scalar proxy values for task difficulty and model capability.

*Risk Rate* and lower *Gain Rate* from introspection, and are prone to be negatively impacted by low-quality knowledge. Meanwhile, introspection is more likely to yield useful knowledge on harder tasks when direct answering fails. Therefore, we aim to control the model's reliance on generated knowledge in different scenarios, from favoring direct answering to introspection.

### 4.2 Stage 2: Knowledge Regeneration

**Goal:** Regenerate higher-quality knowledge alternative when the initially generated knowledge was filtered out.

**Method:** Inspired by the use of temperature to shape a probability distribution (Ackley et al., 1985), we apply an analogous concept to the attention scores to guide the generation process away from failure modes as we discovered in Section 3.3. The attention temperature $\tau_i$ for head $i$ is computed using the following formula:

$$\tau_i = 1 + \alpha \left(1 - \frac{H_i}{H_i^{\max}}\right) \quad (3)$$

Here, $H_i$ represents the Shannon entropy of the current attention distribution for head $i$. This entropy value is normalized by the maximum possible entropy $H_i^{\max}$ (typically $\log l$ for a sequence of length $l$). $\alpha$ is a hyperparameter controlling the intensity of the temperature scaling ($\alpha \geq 0$). We apply this temperature to the original attention scores $\mathbf{S}$ for head $i$ to obtain the final attention distribution $\mathbf{A}^{\text{final}}$:

$$\mathbf{A}^{\text{final}} = \text{Softmax}\left(\frac{\mathbf{S}}{\tau_i}\right) \quad (4)$$

Its quality will also be evaluated using the criteria from Stage 1 before final answer generation.

**Rationale:** This stage leverages the analysis results in Section 3.3 that harmful knowledge generation may correlate with overly sharp attention distributions. By smoothing these distributions during the regeneration attempt, we encourage the model to integrate broader contextual information and avoid the pitfalls that led to the initial low-quality knowledge generation.

### 4.3 Experimental Setup

To evaluate the effectiveness of our proposed two-stage method, we conducted experiments across the 8 commonsense reasoning benchmarks and 5 LLMs described in Section 3.1.1. More implementation details are reported in Appendix D. We compare against several baselines:

- **Direct Answer**: The model answers directly without any intermediate knowledge.
- **CoT**: Standard zero-shot Chain-of-Thought approach (Kojima et al., 2022)
- **Supervised Introspection**: Approaches involve training specialized models to act as introspectors, including Rainier-large (Liu et al., 2022a), Crystal-3B and Crystal-11B (Liu et al., 2023).
- **Unsupervised Introspection**: The model generates intermediate knowledge itself, including Self-Talk (Shwartz et al., 2020), GKP (Liu et al., 2022b) and ZEBRA (Molfese et al., 2024).

### 4.4 Main Results

Table 2 presents the main experimental results. We can get the following conclusions: **(1) Vanilla introspection and CoT often exhibit inconsistent performance and can degrade accuracy.** While beneficial for weaker models, introspection methods frequently fail with stronger models and simpler tasks compared to direct answer. Similarly, CoT also displays mixed results, with notable degradation in specific settings. **(2) Our proposed strategy effectively overcomes these limitations and yields performance improvements across diverse models and tasks.** By modulating the introspection process, our two-stage approach achieves positive accuracy gains (indicated by green backgrounds) in almost all evaluated scenarios. The strategy effectively mitigates the limitations of vanilla introspection and demonstrates superior performance compared to existing methods. This consistent improvement underscores the efficacy of our strategy in reliably leveraging knowledge introspection for commonsense reasoning.

### 4.5 Ablation Study and Analysis

**Contribution of Knowledge Detection.** While vanilla introspection frequently degrades the performance, especially for stronger models and simpler tasks, Adaptive Intro (Stage 1) significantly mitigates this degradation by identifying and filtering potentially low-quality knowledge. For Llama-3-70B-Instruct, Stage 1 reverses the -2.4 average drop of GKP into a positive average gain of +0.4. While GKP improves weaker models significantly (e.g., Llama-2-7b: +3.5), Stage 1 also provides competitive gains (Llama-2-7b: +2.1). This demonstrates that detection is beneficial even when starting from a positive baseline.

**Contribution of Knowledge Regeneration.** As Stage 1 effectively prevents performance drops,

Table 2: Performance comparison across models and methods (%). ↑ and ↓ indicate the change compared to the respective model's Direct Answer baseline. **Green background indicates improvement, red background indicates degradation compared to baseline.** For brevity, the *Chat* and *Instruct* suffixes for the models are omitted.

| Category | Method (Model) | ARC-E | ARC-C | CSQA | CSQA2 | OBQA | PIQA | QASC | WG | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| **Baseline** | Direct Answer (Llama-2-7b) * | 71.4 | 53.6 | 52.3 | 52.0 | 51.4 | 60.7 | 43.4 | 50.8 | 54.4 |
| | Direct Answer (Llama-2-13b) * | 79.3 | 61.9 | 63.6 | 56.5 | 59.8 | 77.1 | 60.6 | 52.6 | 63.9 |
| | Direct Answer (Llama-2-70b) * | 85.8 | 74.0 | 72.6 | 60.0 | 74.4 | 79.4 | 71.6 | 52.0 | 71.2 |
| | Direct Answer (Llama-3-8B) * | 92.1 | 78.6 | 75.0 | 63.6 | 77.0 | 80.5 | 80.1 | 57.9 | 75.6 |
| | Direct Answer (Llama-3-70B) * | 97.7 | 92.8 | 81.8 | 75.9 | 91.6 | 89.3 | 90.1 | 69.5 | 86.1 |
| **CoT** | CoT (Llama-2-7b) | 65.8 (↓5.6) | 48.6 (↓5.0) | 53.8 (↑1.5) | 48.5 (↓3.5) | 49.2 (↓2.2) | 53.1 (↓7.6) | 44.9 (↑1.5) | 45.9 (↓4.9) | 51.2 (↓3.2) |
| | CoT (Llama-2-13b) | 74.8 (↓4.5) | 61.9 (↑0.0) | 56.8 (↓6.8) | 59.5 (↑3.0) | 57.6 (↓2.2) | 70.2 (↓6.9) | 56.8 (↓3.8) | 50.3 (↓2.3) | 61.0 (↓2.9) |
| | CoT (Llama-2-70b) | 85.1 (↓0.7) | 75.9 (↑1.9) | 72.6 (↑0.0) | 64.7 (↑4.7) | 73.6 (↓0.8) | 74.3 (↓5.1) | 70.4 (↓1.2) | 56.4 (↑4.4) | 71.6 (↑0.4) |
| | CoT (Llama-3-8B) | 92.4 (↑0.3) | 81.1 (↑2.5) | 75.4 (↑0.4) | 64.5 (↑0.9) | 76.8 (↓0.2) | 79.5 (↓1.0) | 79.8 (↓0.3) | 58.0 (↑0.1) | 75.9 (↑0.3) |
| | CoT (Llama-3-70B) | 97.3 (↓0.4) | 93.5 (↑0.7) | 80.9 (↓0.9) | 78.2 (↑2.3) | 89.6 (↓2.0) | 88.7 (↓0.6) | 89.2 (↓0.9) | 75.5 (↑6.0) | 86.6 (↑0.5) |
| **Supervised Introspection** | Rainier-L (Llama-2-7b) | 70.6 (↓0.8) | 52.0 (↓1.6) | 55.8 (↑3.5) | 52.2 (↑0.2) | 49.0 (↓2.4) | 61.4 (↑0.7) | 46.2 (↑2.8) | 51.4 (↑0.6) | 54.9 (↑0.5) |
| | Rainier-L (Llama-2-13b) | 79.4 (↑0.1) | 62.2 (↑0.3) | 65.4 (↑1.8) | 53.8 (↓2.7) | 61.0 (↑1.2) | 75.0 (↓2.1) | 63.3 (↑2.7) | 53.9 (↑1.3) | 64.2 (↑0.3) |
| | Rainier-L (Llama-2-70b) | 82.6 (↓3.2) | 68.6 (↓5.4) | 66.3 (↓6.3) | 53.7 (↓6.3) | 66.8 (↓7.6) | 74.4 (↓5.0) | 66.1 (↓5.5) | 52.4 (↑0.4) | 66.4 (↓4.8) |
| | Rainier-L (Llama-3-8B) | 87.2 (↓4.9) | 73.5 (↓5.1) | 71.3 (↓3.7) | 58.5 (↓5.1) | 71.8 (↓5.2) | 77.1 (↓3.4) | 72.9 (↓7.2) | 61.9 (↑4.0) | 71.8 (↓3.8) |
| | Rainier-L (Llama-3-70B) | 94.1 (↓3.6) | 87.8 (↓5.0) | 76.9 (↓4.9) | 62.0 (↓13.9) | 83.8 (↓7.8) | 83.4 (↓5.9) | 76.7 (↓13.4) | 74.0 (↑4.5) | 79.8 (↓6.3) |
| | Crystal-3B (Llama-2-7b) | 72.2 (↑0.8) | 52.9 (↓0.7) | 58.2 (↑5.9) | 53.0 (↑1.0) | 53.0 (↑1.6) | 61.3 (↑0.6) | 47.7 (↑4.3) | 50.7 (↓0.1) | 56.1 (↑1.7) |
| | Crystal-3B (Llama-2-13b) | 81.0 (↑1.7) | 62.4 (↑0.5) | 66.0 (↑2.4) | 54.1 (↓2.4) | 65.4 (↑5.6) | 75.4 (↓1.7) | 64.1 (↑3.5) | 54.1 (↑1.5) | 65.3 (↑1.4) |
| | Crystal-3B (Llama-2-70b) | 82.3 (↓3.5) | 68.1 (↓5.9) | 69.7 (↓2.9) | 54.9 (↓5.1) | 73.6 (↓0.8) | 77.3 (↓2.1) | 70.1 (↓1.5) | 52.2 (↑0.2) | 68.5 (↓2.7) |
| | Crystal-3B (Llama-3-8B) | 87.5 (↓4.6) | 73.6 (↓5.0) | 73.9 (↓1.1) | 59.0 (↓4.6) | 75.8 (↓1.2) | 78.2 (↓2.3) | 75.2 (↓4.9) | 61.5 (↑3.6) | 73.1 (↓2.5) |
| | Crystal-3B (Llama-3-70B) | 93.3 (↓4.4) | 86.9 (↓5.9) | 79.0 (↓2.8) | 64.9 (↓11.0) | 86.6 (↓5.0) | 86.0 (↓3.3) | 80.2 (↓9.9) | 74.7 (↑5.2) | 81.4 (↓4.7) |
| | Crystal-11B (Llama-2-7b) | 75.2 (↑3.8) | 55.8 (↑2.2) | 56.7 (↑4.4) | 53.4 (↑1.4) | 54.8 (↑3.4) | 62.7 (↑2.0) | 49.2 (↑5.8) | 52.2 (↑1.4) | 57.5 (↑3.1) |
| | Crystal-11B (Llama-2-13b) | 82.7 (↑3.4) | 64.2 (↑2.3) | 67.0 (↑3.4) | 55.8 (↓0.7) | 67.4 (↑7.6) | 77.4 (↑0.3) | 66.4 (↑5.8) | 54.8 (↑2.2) | 66.9 (↑3.0) |
| | Crystal-11B (Llama-2-70b) | 86.1 (↑0.3) | 72.6 (↓1.4) | 70.5 (↓2.1) | 57.0 (↓3.0) | 74.2 (↓0.2) | 79.9 (↑0.5) | 72.4 (↑0.8) | 53.0 (↑1.0) | 70.7 (↓0.5) |
| | Crystal-11B (Llama-3-8B) | 89.6 (↓2.5) | 76.5 (↓2.1) | 74.1 (↓0.9) | 60.8 (↓2.8) | 76.4 (↓0.6) | 80.1 (↓0.4) | 78.9 (↓1.2) | 61.0 (↑3.1) | 74.7 (↓0.9) |
| | Crystal-11B (Llama-3-70B) | 94.9 (↓2.8) | 89.2 (↓3.6) | 78.6 (↓3.2) | 66.2 (↓9.7) | 89.0 (↓2.6) | 87.4 (↓1.9) | 82.3 (↓7.8) | 77.5 (↑8.0) | 83.1 (↓3.0) |
| **Unsupervised Introspection** | Self-Talk (Llama-2-7b) | 63.2 (↓8.2) | 48.2 (↓5.4) | 51.8 (↓0.5) | 51.2 (↓0.8) | 41.6 (↓9.8) | 59.3 (↓1.4) | 40.3 (↓3.1) | 50.5 (↓0.3) | 50.8 (↓3.6) |
| | Self-Talk (Llama-2-13b) | 78.1 (↓1.2) | 59.5 (↓2.4) | 61.7 (↓1.9) | 55.5 (↓1.0) | 62.0 (↑2.2) | 75.2 (↓1.9) | 60.2 (↓0.4) | 53.7 (↑1.1) | 63.2 (↓0.7) |
| | Self-Talk (Llama-2-70b) | 74.9 (↓10.9) | 62.8 (↓11.2) | 64.3 (↓8.3) | 55.4 (↓4.6) | 60.4 (↓14.0) | 76.0 (↓3.4) | 61.8 (↓9.8) | 51.7 (↓0.3) | 63.4 (↓7.8) |
| | Self-Talk (Llama-3-8B) | 89.0 (↓3.1) | 75.8 (↓2.8) | 70.6 (↓4.4) | 63.1 (↓0.5) | 71.0 (↓6.0) | 77.7 (↓2.8) | 75.2 (↓4.9) | 59.8 (↑1.9) | 72.8 (↓2.8) |
| | Self-Talk (Llama-3-70B) | 95.7 (↓2.0) | 89.5 (↓3.3) | 80.9 (↓0.9) | 70.1 (↓5.8) | 86.8 (↓4.8) | 88.1 (↓1.2) | 86.0 (↓4.1) | 75.5 (↑6.0) | 84.1 (↓2.0) |
| | GKP (Llama-2-7b) * | 74.7 (↑3.3) | 57.8 (↑4.2) | 59.2 (↑6.9) | 54.3 (↑2.3) | 55.6 (↑4.2) | 60.7 (↑0.0) | 48.8 (↑5.4) | 52.2 (↑1.4) | 57.9 (↑3.5) |
| | GKP (Llama-2-13b) * | 82.3 (↑3.0) | 66.0 (↑4.1) | 65.8 (↑2.2) | 62.8 (↑6.3) | 67.6 (↑7.8) | 76.2 (↓0.9) | 63.8 (↑3.2) | 55.1 (↑2.5) | 67.4 (↑3.5) |
| | GKP (Llama-2-70b) * | 89.5 (↑3.7) | 76.9 (↑2.9) | 74.7 (↑2.1) | 63.8 (↑3.8) | 74.2 (↓0.2) | 78.2 (↓1.2) | 73.0 (↑1.4) | 54.0 (↑2.0) | 73.0 (↑1.8) |
| | GKP (Llama-3-8B) * | 91.4 (↓0.7) | 78.5 (↓0.1) | 74.2 (↓0.8) | 66.1 (↑2.5) | 76.2 (↓0.8) | 75.6 (↓4.9) | 80.6 (↑0.5) | 61.7 (↑3.8) | 75.5 (↓0.1) |
| | GKP (Llama-3-70B) * | 92.5 (↓5.2) | 89.6 (↓3.2) | 79.9 (↓1.9) | 76.8 (↑0.9) | 84.2 (↓7.4) | 86.2 (↓3.1) | 82.8 (↓7.3) | 77.9 (↑8.4) | 83.7 (↓2.4) |
| | ZEBRA (k=5) (Llama-2-7b) | 75.0 (↑3.6) | 55.9 (↑2.3) | 60.8 (↑8.5) | 54.3 (↑2.3) | 53.6 (↑2.2) | 66.1 (↑5.4) | 46.8 (↑3.4) | 52.9 (↑2.1) | 58.2 (↑3.8) |
| | ZEBRA (k=5) (Llama-2-13b) | 82.8 (↑3.5) | 65.8 (↑3.9) | 67.4 (↑3.8) | 59.8 (↑3.3) | 65.2 (↑5.4) | 75.7 (↓1.4) | 63.5 (↑2.9) | 53.0 (↑0.4) | 66.7 (↑2.8) |
| | ZEBRA (k=5) (Llama-2-70b) | 87.9 (↑2.1) | 76.6 (↑2.6) | 75.6 (↑3.0) | 61.3 (↑1.3) | 73.4 (↓1.0) | 77.4 (↓2.0) | 72.2 (↑0.6) | 55.4 (↑3.4) | 72.5 (↑1.3) |
| | ZEBRA (k=5) (Llama-3-8B) | 92.3 (↑0.2) | 78.4 (↓0.2) | 77.5 (↑2.5) | 62.1 (↓1.5) | 74.4 (↓2.6) | 76.1 (↓4.4) | 77.4 (↓2.7) | 60.6 (↑2.7) | 74.8 (↓0.8) |
| | ZEBRA (k=5) (Llama-3-70B) | 94.1 (↓3.6) | 87.7 (↓5.1) | 80.8 (↓1.0) | 75.1 (↓0.8) | 86.6 (↓5.0) | 83.2 (↓6.1) | 79.6 (↓10.5) | 75.9 (↑6.4) | 82.9 (↓3.2) |
| **Proposed** | Adaptive Intro (Stage 1) (Llama-2-7b) | 74.8 (↑3.4) | 57.5 (↑3.9) | 56.6 (↑4.3) | 53.2 (↑1.2) | 53.8 (↑2.4) | 60.7 (↑0.0) | 44.7 (↑1.3) | 50.8 (↑0.0) | 56.5 (↑2.1) |
| | Adaptive Intro (Stage 1) (Llama-2-13b) | 82.2 (↑2.9) | 65.3 (↑3.4) | 65.7 (↑2.1) | 62.9 (↑6.4) | 67.8 (↑8.0) | 76.1 (↓1.0) | 63.3 (↑2.7) | 54.9 (↑2.3) | 67.3 (↑3.4) |
| | Adaptive Intro (Stage 1) (Llama-2-70b) | 88.2 (↑2.4) | 74.5 (↑0.5) | 73.4 (↑0.8) | 62.9 (↑2.9) | 76.2 (↑1.8) | 78.5 (↓0.9) | 73.0 (↑1.4) | 53.4 (↑1.4) | 72.5 (↑1.3) |
| | Adaptive Intro (Stage 1) (Llama-3-8B) | 91.8 (↓0.3) | 79.7 (↑1.1) | 76.3 (↑1.3) | 64.2 (↑0.6) | 77.8 (↑0.8) | 79.4 (↓1.1) | 79.5 (↓0.6) | 60.7 (↑2.8) | 76.2 (↑0.6) |
| | Adaptive Intro (Stage 1) (Llama-3-70B) | 97.9 (↑0.2) | 92.7 (↓0.1) | 82.4 (↑0.6) | 75.6 (↓0.3) | 92.4 (↑0.8) | 89.3 (↑0.0) | 89.8 (↓0.3) | 71.6 (↑2.1) | 86.5 (↑0.4) |
| | Adaptive Intro (Stage 1+2) (Llama-2-7b) | 75.3 (↑3.9) | 57.8 (↑4.2) | 57.0 (↑4.7) | 53.4 (↑1.4) | 54.0 (↑2.6) | 61.0 (↑0.3) | 45.2 (↑1.8) | 50.8 (↑0.0) | 56.8 (↑2.4) |
| | Adaptive Intro (Stage 1+2) (Llama-2-13b) | 82.4 (↑3.1) | 67.0 (↑5.1) | 66.3 (↑2.7) | 63.2 (↑6.7) | 69.2 (↑9.4) | 76.6 (↓0.5) | 63.8 (↑3.2) | 55.2 (↑2.6) | 67.9 (↑4.0) |
| | Adaptive Intro (Stage 1+2) (Llama-2-70b) | 88.6 (↑2.8) | 76.5 (↑2.5) | 74.3 (↑1.7) | 63.2 (↑3.2) | 78.4 (↑4.0) | 79.5 (↑0.1) | 73.3 (↑1.7) | 54.5 (↑2.5) | 73.5 (↑2.3) |
| | Adaptive Intro (Stage 1+2) (Llama-3-8B) | 92.3 (↑0.2) | 80.8 (↑2.2) | 77.6 (↑2.6) | 65.9 (↑2.3) | 79.6 (↑2.6) | 80.0 (↓0.5) | 79.8 (↓0.3) | 61.6 (↑3.7) | 77.2 (↑1.6) |
| | Adaptive Intro (Stage 1+2) (Llama-3-70B) | 97.9 (↑0.2) | 93.0 (↑0.2) | 82.6 (↑0.8) | 75.8 (↓0.1) | 92.4 (↑0.8) | 89.7 (↑0.4) | 89.7 (↓0.4) | 71.9 (↑2.4) | 86.6 (↑0.5) |

Stage 2 aims to provide further enhancement by regenerating higher-quality knowledge. Across all models, Stage 1+2 achieves higher average accuracy. However, we observe a slight decrease in performance for Llama-3-70B-Instruct on QASC when moving from Stage 1 (-0.3) to Stage 1+2 (-0.4). This suggests that while the regeneration techniques in Stage 2 prove effective in most cases, there are still instances where direct answer remains a better alternative. We leave a more fine-grained study for future work.

Table 3: Comparison of average scores for original and regenerated Knowledge. Scores for correctness and relevance can be 0, 1, or 2.

| Type | Version | Correctness | Relevance |
|---|---|---|---|
| Harmful | Original | 0.3083 | 0.6293 |
| | Regenerated | 1.3795 | 1.2428 |
| Useful | Original | 1.7002 | 1.8360 |
| | Regenerated | 1.6123 | 1.4969 |

**Validation of Regeneration Quality.** To directly validate the quality of regenerated knowledge, we compare it with the original knowledge using the same annotation standard discussed in Appendix B. Table 3 shows that while regeneration results in a slight decrease for already useful knowledge, it significantly improves the correctness and relevance of original harmful knowledge. This aligns with our goal of improving low-quality knowledge and mitigating the interference of harmful information.

## 5 Conclusion

We investigate the effectiveness of knowledge introspection for commonsense reasoning in LLMs, uncovering the introspection paradox. Analysis indicates that the effectiveness of introspection results from the interplay among model capability, task difficulty, and the quality of generated knowledge. To address this, we proposed a novel, training-free strategy. It optimizes introspection via two stages: Knowledge Detection and Knowledge Regeneration. Extensive experiments across 5 LLMs and 8 benchmarks demonstrate that our method effectively mitigates the performance degradation, achieving robust gains across diverse models and tasks. Our results validate that managing the introspection process is crucial for reliably harnessing its potential to enhance LLMs.

## Limitations

While our proposed approach demonstrates promising results in mitigating the introspection paradox and improving commonsense reasoning, there are several aspects that could be improved. Firstly, our experiments are confined to the Llama family, including Llama-2-Chat and Llama-3-Instruct variants across different scales. The impact of different model architectures or alternative post-training strategies requires future investigation. Furthermore, certain hyperparameters in our method rely on empirical calibration. Future research could explore more theoretically grounded or automated methods for setting these parameters. Additionally, our reliance on a series of proxy metrics and model annotations might not fully capture all subtle relationships. Exploring alternative, potentially more nuanced metrics is a direction for future work.

## References


David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. 1985. A learning algorithm for boltzmann machines. *Cogn. Sci.*, 9:147–169.

Leonardo Berti, Flavio Giorgi, and Gjergji Kasneci. 2025. Emergent abilities in large language models: A survey. *Preprint*, arXiv:2503.05788.

Akshita Bhagia, Jiacheng Liu, Alexander Wettig, David Heineman, Oyvind Tafjord, Ananya Harsh Jha, Luca Soldaini, Noah A. Smith, Dirk Groeneveld, Pang Wei Koh, Jesse Dodge, and Hannaneh Hajishirzi. 2024. Establishing task scaling laws via compute-efficient model ladders. *CoRR*, abs/2412.04403.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Antoine Bosselut, Ronan Le Bras, and Yejin Choi. 2021. Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 35, pages 4923–4931.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Joe Davison, Joshua Feldman, and Alexander Rush. 2019. Commonsense knowledge mining from pretrained models. 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 1173–1178, Hong Kong, China. Association for Computational Linguistics.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12963–12971.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Paul Saldyt, and Anil B Murthy. 2024. Position: LLMs can't plan, but can help planning in LLM-modulo frameworks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 22895–22907. PMLR.

Jivat Kaur, Sumit Bhatia, Milan Aggarwal, Rachit Bansal, and Balaji Krishnamurthy. 2022. LM-CORE: Language models with contextually relevant external

knowledge. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 750–769, Seattle, United States. Association for Computational Linguistics.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. *arXiv:1910.11473v2*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jiachun Li, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Daojian Zeng, Kang Liu, and Jun Zhao. 2024. Focus on your question! interpreting and mitigating toxic CoT problems in commonsense reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9206–9230, Bangkok, Thailand. Association for Computational Linguistics.

Huanxuan Liao, Shizhu He, Yupu Hao, Xiang Li, Yuanzhe Zhang, Jun Zhao, and Kang Liu. 2025a. Skintern: Internalizing symbolic knowledge for distilling better cot capabilities into small language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3203–3221.

Huanxuan Liao, Shizhu He, Yao Xu, Yuanzhe Zhang, Shengping Liu, Kang Liu, and Jun Zhao. 2025b. Awakening augmented generation: Learning to awaken internal knowledge of large language models for question answering. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1333–1352, Abu Dhabi, UAE. Association for Computational Linguistics.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. 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 2829–2839, Hong Kong, China. Association for Computational Linguistics.

Jiacheng Liu, Skyler Hallinan, Ximing Lu, Pengfei He, Sean Welleck, Hannaneh Hajishirzi, and Yejin Choi. 2022a. Rainier: Reinforced knowledge introspector for commonsense question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8938–8958, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022b. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.

Jiacheng Liu, Ramakanth Pasunuru, Hannaneh Hajishirzi, Yejin Choi, and Asli Celikyilmaz. 2023. Crystal: Introspective reasoners reinforced with self-feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11557–11572, Singapore. Association for Computational Linguistics.

Ryan Liu, Jiayi Geng, Addison J. Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L. Griffiths. 2024. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *Preprint*, arXiv:2410.21333.

Kaixin Ma, Jonathan Francis, Quanyang Lu, Eric Nyberg, and Alessandro Oltramari. 2019. Towards generalizable neuro-symbolic systems for commonsense question answering. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 22–32, Hong Kong, China. Association for Computational Linguistics.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Arindam Mitra, Pratyay Banerjee, Kuntal Kumar Pal, Swaroop Mishra, and Chitta Baral. 2020. How additional knowledge can improve natural language commonsense question answering? *Preprint*, arXiv:1909.08855.

Francesco Maria Molfese, Simone Conia, Riccardo Orlando, and Roberto Navigli. 2024. ZEBRA: Zero-shot example-based retrieval augmentation for commonsense question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22429–22444, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI. 2024. Learning to reason with llms.

Qwen. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *Preprint*, arXiv:1907.10641.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: an atlas of machine commonsense for if-then reasoning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629, Online. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: an open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. *Preprint*, arXiv:1703.01365.

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.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. CommonsenseQA 2.0: Exposing the limits of AI through gamification. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2022. Commonsenseqa 2.0: Exposing the limits of ai through gamification. *Preprint*, arXiv:2201.05320.

Tianyi Tang, Yushuo Chen, Yifan Du, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Learning to imagine: Visually-augmented natural language generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9468–9481, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Xi Wang, Taketomo Isazawa, Liana Mikaelyan, and James Hensman. 2025. KBLam: Knowledge base augmented language model. In *The Thirteenth International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Xiaohan Xu, Chongyang Tao, Tao Shen, Can Xu, Hongbo Xu, Guodong Long, Jian-Guang Lou, and Shuai Ma. 2024. Re-reading improves reasoning in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15549–15575, Miami, Florida, USA. Association for Computational Linguistics.

Yunzhi Yao, Peng Wang, Shengyu Mao, Chuanqi Tan, Fei Huang, Huajun Chen, and Ningyu Zhang. 2023. Knowledge rumination for pre-trained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3387–3404, Singapore. Association for Computational Linguistics.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online. Association for Computational Linguistics.

Wenhao Yu, Chenguang Zhu, Zhihan Zhang, Shuohang Wang, Zhuosheng Zhang, Yuwei Fang, and Meng Jiang. 2022. Retrieval augmentation for commonsense reasoning: A unified approach. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4364–4377, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. 2022. GreaseLM: Graph REASoning enhanced language models. In *International Conference on Learning Representations*.

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. *Preprint*, arXiv:2504.05081.

Wangchunshu Zhou, Dong-Ho Lee, Ravi Kiran Selvam, Seyeon Lee, and Xiang Ren. 2021. Pre-training text-to-text transformers for concept-centric common

sense. In *International Conference on Learning Representations*.

# A Proxy Metrics

## A.1 Task Difficulty

We use the model's prediction uncertainty on questions from a specific task as a proxy for task difficulty. Specifically, we examine the relationship between the model's accuracy on a task and the mean entropy of its predictions for questions in that task. As shown in Figure 9, higher confidence for a given model generally correlates with higher accuracy. This suggests that the metric can reflect the relative difficulty of tasks. Based on the metric, the tasks rank from easiest (ARC-Easy) to hardest (WinoGrande).

## A.2 Model Capability

To quantify model capability, we follow recent works (Bhagia et al., 2024; Grattafiori et al., 2024) and use the normalized negative log-likelihood (NLL) loss on the MMLU benchmark's test set (Hendrycks et al., 2021) as a proxy metric. MMLU is a widely recognized benchmark designed to assess a model's general knowledge across a broad range of subjects, making it a suitable proxy for assessing commonsense reasoning tasks. Lower NLL on this benchmark indicates stronger capability. Figure 10 shows the MMLU NLL for the evaluated models. Figure 11 provides a more granular view, showing the NLL for each model across the specific commonsense reasoning tasks used in this paper. In our experiments, we take the negative of this NLL value, such that a higher value indicates a stronger model, aligning with the intuitive understanding of capability. Based on the metric, the models rank from weakest (Llama-2-7b-Chat) to strongest (Llama-3-70B-Instruct).

# B Knowledge Annotation

For each generated knowledge statement, we employ Deepseek-V3 (DeepSeek-AI, 2024) to assign a numerical score for both factual correctness and relevance according to the criteria provided in the prompts below. A knowledge statement was classified as *Harmful* if both correctness and relevance scores were less than 2; otherwise, it was classified as *Useful*.

**Correctness**

Please evaluate the factual correctness of the following statement: {**knowledge**}
Use the following criteria to determine your response:
*Incorrect: The statement contains factual inaccuracies or contradictions.
*Uncertain: The statement cannot be clearly verified as true or false based on the given information.
*Correct: The statement is factually accurate and consistent with reliable knowledge.
Your response must be a single number: 0 for Incorrect, 1 for Uncertain and 2 for Correct. Do not include any additional text in your answer; only provide the number.

**Relevance**

Please determine if the following knowledge is helpful for solving the problem:
*Knowledge: {**knowledge**}
*Question: {**question**}
*Answer: {**answer**}
Knowledge is related to the final answer in the following ways. Use the following criteria to determine your response:
*Helpful: The knowledge can be part of a non-trivial reasoning chain that supports the predicted answer or a trivial paraphrase of the question and the predicted answer.
*Unrelated: Any of the following: The knowledge is a mere repetition of known information given in the question; The knowledge is topically related to the question and/or the choices, but cannot be part of a reasoning chain to support or refute any of the choices; The knowledge is unrelated to the question.
*Contradict: The knowledge can be part of a reasoning chain that refutes the predicted answer, or supports a different choice.
Your response must be a single number: 0 for Contradict, 1 for Unrelated and 2 for Helpful. Do not include any additional text in your answer; only provide the number.

# C Attribution Tracing

## C.1 Attribution Scores Across Layers

As discussed in Section 3.3, we employ attribution tracing to understand the information flow from the input question context to the generated knowledge statement. Figure 12 provides a comprehensive visualization of these attribution scores across different layers for all models and tasks evaluated in this study.

## C.2 Normalized Entropy Calculation

In this appendix, we provide a detailed description of the method used to quantify the distribution of attribution scores between the question and the generated knowledge, and present the comprehensive results across all models and tasks.

We analyze the attribution matrix $A \in \mathbb{R}^{K \times Q}$ obtained from the method described in Section 3.3, where $A_{i,j}^{(l)}$ represents the aggregated attribution score from question token $q_i$ to knowledge token $k_j$ for a specific layer $l$ (i.e., $A_{i,j}^{(l)} = \sum_{h=1}^{H} \left| [\text{Attr}(A_h^{(l)})]_{i,j} \right|$). $K$ and $Q$ are the lengths of the knowledge and question sequence, respectively. We normalize the matrix $A^{(l)}$ by its L1 norm to obtain a probability distribution $P^{(l)} \in \mathbb{R}^{K \times Q}$:

$$P_{i,j}^{(l)} = \frac{A_{i,j}^{(l)}}{\sum_{m=1}^{K} \sum_{n=1}^{Q} A_{m,n}^{(l)}} \tag{5}$$

We then compute the normalized Shannon entropy of this distribution as follows:

$$H_{norm}(l) = \frac{-\sum_{j=1}^{K} \sum_{i=1}^{Q} P_{j,i}^{(l)} \log P_{j,i}^{(l)}}{\log(KQ)} \tag{6}$$

During implementation, we focus on the layer that exhibits the maximal difference in attribution scores between useful and harmful knowledge instances, as identified by the preliminary analysis presented in C.1.

Figure 13 presents the average normalized attribution matrix entropy for both useful and harmful knowledge across all evaluated models and tasks.

# D Implementation Details

**Tasks and Models** We use 8 diverse commonsense reasoning datasets: ARC-Easy and ARC-Challenge (Clark et al., 2018), PIQA (Bisk et al., 2020), OpenBookQA (Mihaylov et al., 2018), QASC (Khot et al., 2020), CSQA (Talmor et al., 2019), CSQA2 (Talmor et al., 2022), and WinoGrande (Sakaguchi et al., 2019). These benchmarks are chosen to represent a range of task difficulties. We evaluate Llama family models (Touvron et al., 2023; Grattafiori et al., 2024) across different scales and versions: Llama-2-7b-Chat, Llama-2-13b-Chat, Llama-2-70b-Chat, Llama-3-8B-Instruct and Llama-3-70B-Instruct.

**Our Method** In stage 1, the mean and variance for z-score normalization are calculated from the training set of each task. For correctness calculation, the model generates a knowledge statement $K = (k_1, k_2, \ldots, k_{L_K})$ of length $L_K$. For each token $k_t \in K$ (for $t = 1, \ldots, L_K$), let $P_{intro}$ be the introspective prompt and $\mathbf{p}_t = P(\cdot | P_{intro}, k_{<t})$ be the probability distribution over the vocabulary $V$ predicted by the model for the $t$-th token of the knowledge. The entropy of this distribution is

$$H(\mathbf{p}_t) = -\sum_{v \in V} p_t(v) \log_2 p_t(v) \tag{7}$$

The correctness metric is defined as the average entropy over the generated knowledge tokens:

$$Correctness = \frac{1}{L_K} \sum_{t=1}^{L_K} H(\mathbf{p}_t) \tag{8}$$

For relevance calculation, $\mathbf{e}(w_i)$ denotes the final layer hidden state embedding for a token $w_i$. We obtain aggregated hidden state representations for the knowledge $K$ and question $Q$ as:

$$\mathbf{h}_K = \text{Mean}(\{\mathbf{e}(k_i)\}_{i=1}^{L_K}) \tag{9}$$

$$\mathbf{h}_Q = \text{Mean}(\{\mathbf{e}(q_j)\}_{j=1}^{L_Q}) \tag{10}$$

where $\text{Mean}(\cdot)$ denotes the mean pooling function. The relevance metric is the cosine similarity between these representations:

$$Relevance = \frac{\mathbf{h}_K \cdot \mathbf{h}_Q}{\|\mathbf{h}_K\| \|\mathbf{h}_Q\|} \tag{11}$$

In stage 2, the hyperparameter $\alpha$ is set to 1.7. The layer selected is based on the maximal difference in attribution scores between useful and harmful knowledge. Specifically, we use 6 layers for Llama-2-7b-Chat and Llama-3-8B-Instruct, 7 layers for Llama-2-13b-Chat, and 8 layers for Llama-2-70b-Chat and Llama-3-70B-Instruct.

**Prompts** For the knowledge generation step, we follow the original settings specified by each baseline. For few-shot knowledge generation, we aligned ZEBRA's (Molfese et al., 2024) prompt format with GKP's (Liu et al., 2022b). For the Question Answering step, we employ standardized prompt templates and use greedy decoding for answer generation across all models and methods. The prompts are as follows:

**Direct Answering**

**System:** You are a helpful assistant for question answering. You are given a question and up to 4 options (labeled A, B, C, and D). Your task is to choose the label corresponding to the best answer for the question.

**User:** Do you understand the task?

**Assistant:** Yes, I understand. Please provide the question and the possible choices.

**User:**
Question: {question}
Options: {choices}
You must always give an answer and only pick one answer choice.

**Assistant:** Among A through D, the answer is

**Knowledge-Utilized Answering**

**System:** You are a helpful assistant for question answering. You are given a question, up to 4 options (labeled A, B, C, and D), and a list of explanations. Your task is to choose the label corresponding to the best answer for the question based on the given explanations.

**User:** Do you understand the task?

**Assistant:** Yes, I understand. Please provide the question and the possible choices.

**User:**
Question: {question}
Options: {choices}
Explanations: {knowledge}
You must always give an answer and only pick one answer choice.

**Assistant:** Among A through D, the answer is

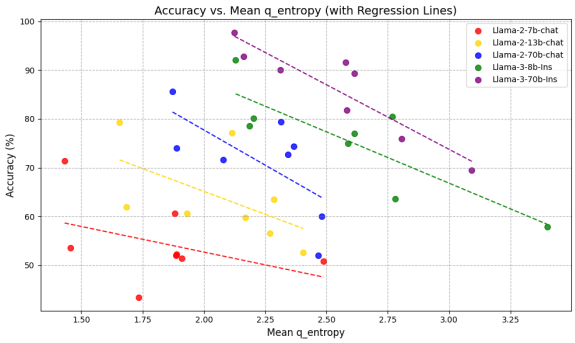



Figure 9: Model accuracy versus mean question entropy. Each point represents a (model, task) pair. Dashed lines are regression lines for each model. Generally, lower entropy correlates with higher accuracy.

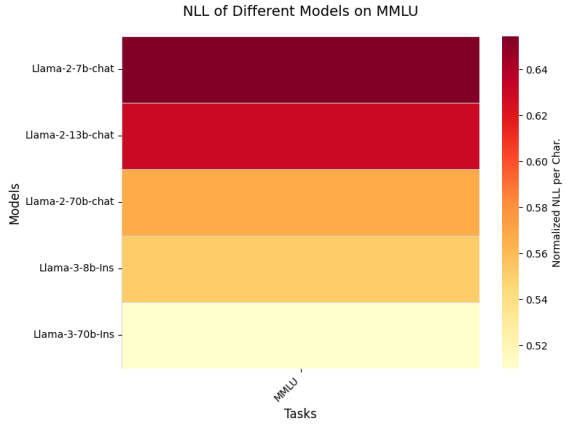



Figure 10: Normalized Negative Log-Likelihood (NLL) scores of different models on the MMLU benchmark. Lower scores indicate stronger overall model capability. This metric is used to rank models (from weakest: Llama-2-7b-chat to strongest: Llama-3-70B-Instruct).

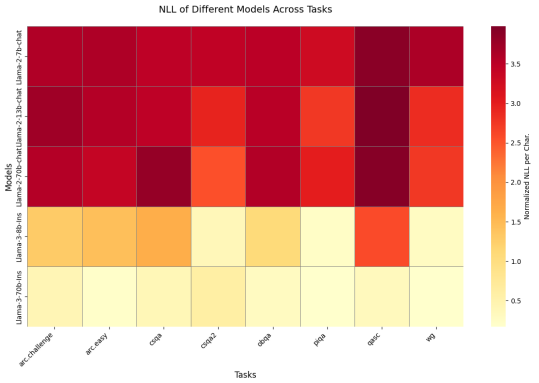



Figure 11: Normalized Negative Log-Likelihood (NLL) scores of different models across the specific commonsense reasoning tasks. This heatmap illustrates performance variations across tasks for each model, indirectly reflecting the relative difficulty of these tasks for different models.

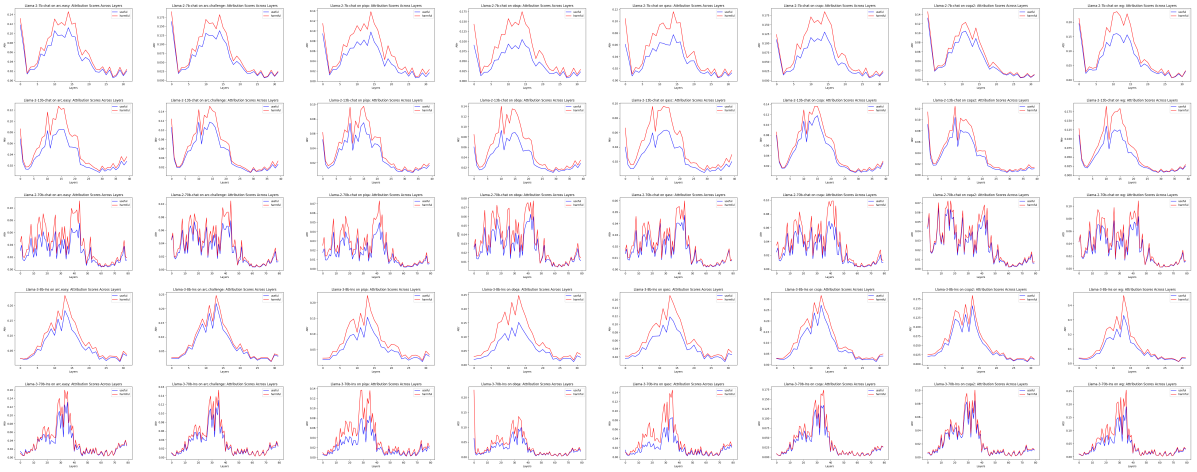
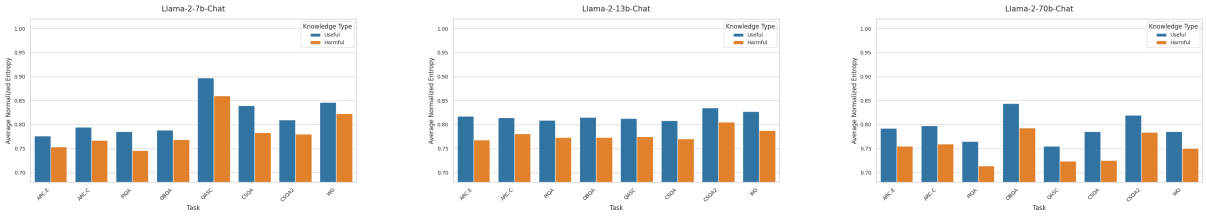
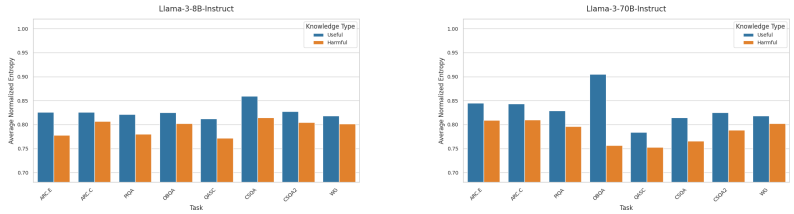


Figure 12: Layer-wise attribution scores from the input question context to the generated knowledge across all evaluated models and tasks.




(a) Llama-2-7b-Chat (b) Llama-2-13b-Chat (c) Llama-2-70b-Chat




(d) Llama-3-8B-Instruct (e) Llama-3-70B-Instruct

Figure 13: Comparison of normalized entropy across models and tasks.