Beyond Surface Simplicity: Revealing Hidden Reasoning Attributes for Precise Commonsense Diagnosis

Anonymous ACL submission

Abstract

Commonsense question answering (QA) are widely used to evaluate the commonsense abilities of large language models. However, answering commonsense questions correctly requires not only knowledge but also reasoning—even for seemingly simple questions. We demonstrate that such hidden reasoning attributes in commonsense questions can lead evaluation accuracy differences of up to 24.8% across different difficulty levels in the same benchmark. Current benchmarks overlook these hidden reasoning attributes, making it difficult to assess a model's specific levels of commonsense knowledge and reasoning ability. To address this issue, we introduce *ReComSBench*, a novel framework that reveals hidden reasoning attributes behind commonsense questions by leveraging the knowledge generated during the reasoning process. Additionally, *ReComSBench* proposes three new metrics for decoupled evaluation: Knowledge Balanced Accuracy, Marginal Sampling Gain, and Knowledge Coverage Ratio. Experiments show that *ReComSBench* provides insights into model performance that traditional benchmarks cannot offer. The difficulty stratification based on revealed hidden reasoning attributes performs as effectively as the model-probability-based approach but is more generalizable and better suited for improving a model's commonsense reasoning abilities. By uncovering and analyzing the hidden reasoning attributes in commonsense data, *ReComSBench* offers a new approach to enhancing existing commonsense benchmarks.

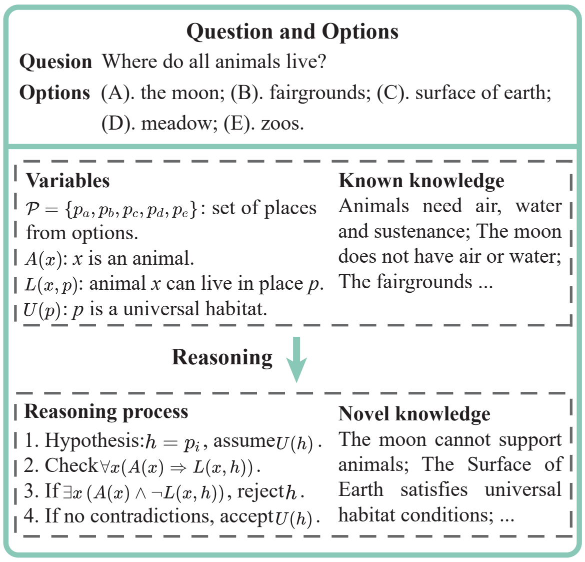


Figure 1: A QA case from CommonsenseQA, showing knowledge transformation during reasoning. Correct answers to simple commonsense questions still require reasoning.

1 Introduction

The study of commonsense involves both knowledge and reasoning (Brachman and Levesque, 2022). Large language models (LLMs) can store and retrieve commonsense knowledge effectively (Bosselut et al., 2019; Davison et al., 2019; Zhao et al., 2023b). In commonsense reasoning tasks, LLMs further exhibit the ability to make inferences based on their stored knowledge (Bhagavatula et al., 2020; Zhao et al., 2023a). To evaluate and enhance LLMs' commonsense capabilities, researchers have utilized diverse benchmarks to measure their performance across both knowledge retrieval and reasoning tasks. Despite dividing the dimensions, commonsense knowledge and reasoning are intertwined, with tasks involving simple reasoning often categorized as commonsense knowledge alone (Davis, 2024). This makes it difficult to determine the individual levels of LLMs' commonsense knowledge and commonsense reasoning abilities. Without this clarity, it is challenging to pinpoint whether a model's errors in handling commonsense tasks stem from one or both of these factors. As a result, efforts to improve both aspects simultaneously often require significant investment but yield limited results.

Another major reason is that crowdsourcing workers naturally ignore the hidden reasoning attributes of commonsense data due to the ambiguity and naturalness of commonsense. This leads to task-irrelevant noise in datasets and causes unexpected overlaps between tasks (Do et al., 2024). Researchers underestimate the impact of this neglect because even when the model answers questions without explicit reasoning, it internally performs hidden reasoning processes before generating responses, which are not directly reflected in the model's output (Ye et al., 2024). As a result, existing benchmarks only provide a macro-evaluation of the commonsense performance of LLMs and cannot effectively differentiate between commonsense knowledge and reasoning abilities. This not only undermines the clarity and effectiveness of commonsense assessment but also limits opportunities for targeted improvements through feedback.

This causes current benchmarks to often overlook two key points. First, even the simplest commonsense questions may involve reasoning attributes that require inference to answer correctly. Second, different questions vary in their reasoning attributes and difficulty levels. For example, as shown in Figure 1, a sample from the CommonsenseQA dataset demonstrates one symbolic reasoning process required to answer correctly. To answer "Where do all animals live?", one must identify exceptions among location options. But CommonsenseQA is a benchmark focused on commonsense knowledge questions.

To address these challenges, we introduce *ReComSBench*, a framework designed to enhance traditional benchmarks by making hidden reasoning attributes explicit. By defining reasoning as the process of generating new knowledge from known knowledge (as shown in Figure 1), *ReComSBench* quantifies reasoning difficulty based on the amount of knowledge required to answer questions correctly. Furthermore, it decouples the evaluation of models' commonsense knowledge and reasoning abilities through three novel metrics: Knowledge Balanced Accuracy for assessing commonsense knowledge, and Marginal Sampling Gain and Knowledge Coverage Ratio for evaluating overall domain reasoning and single inference quality.

We refine and experiment with four benchmarks: CommonsenseQA (Talmor et al., 2019), OpenBookQA (Mihaylov et al., 2018), ARC (Clark et al., 2018), and QASC (Khot et al., 2020). Experiments confirm that hidden reasoning attributes significantly impact model evaluations on existing benchmarks. Data with varying reasoning difficulties within the same benchmark consistently shows lower accuracy for models on high-difficulty data, with up to an 24.8% difference across datasets. This highlights the challenge of distinguishing whether model limitations stem from insufficient knowledge or weak reasoning abilities. The three new metrics provide fine-grained insights into models' knowledge and reasoning capabilities, with results aligning with expectations as model versions evolve, demonstrating their reference value. Using hidden reasoning attributes—measured by the amount of knowledge required during inference—as a basis for data difficulty outperforms the model-probability-based approach. This underscores the practicality of leveraging reasoning attributes for benchmark optimization.

The main contributions of this work are:

- We reveal and validate the importance of hidden reasoning attributes in commonsense data, experimentally demonstrating their impact on model evaluation.
- We propose *ReComSBench*, a framework that improves existing benchmarks by making hidden reasoning attributes explicit. It introduces three novel metrics for decoupled evaluations of commonsense knowledge and reasoning capabilities.
- Through experiments with *ReComSBench*, we confirm its effectiveness in enhancing evaluation and training, showing that organizing data based on hidden reasoning attributes improves models' commonsense abilities.

2 Related works

2.1 Challenges of commonsense benchmarks

There are now over 100 commonsense benchmarks to test AI's knowledge and reasoning abilities (Davis, 2024). While human-annotated datasets are generally high-quality, researchers have found many flaws, such as grammatical errors, incorrect answers, and noisy data. Do et al. (2024) points out that these benchmarks often focus on referenced knowledge rather than true commonsense, harming the accurate measurement of commonsense reasoning. Srivastava et al. (2023) argues that current benchmarks emphasize memory and factual knowledge, calling for "breakthrough" tasks to prepare for future models. Sakaguchi et al. (2021)

highlights spurious biases in datasets, leading to overestimation of machines' true commonsense capabilities. Veselovsky et al. (2023) shows crowd workers using LLMs to generate annotations, lowering dataset quality. Fixing these flaws helps us better understand and improve models' true capabilities. While complex problems get more attention, simple ones often involve deep reasoning processes. Even if LLMs lacks specific knowledge, it might infer correct answers through reasoning. Thus, we need to decouple knowledge and reasoning in commonsense data to evaluate models more accurately.

2.2 Hidden biases in commonsense data

The latent biases in commonsense data have significant impacts on model performance and evaluation. Existing studies reveal various types of biases. Bauer et al. (2023) identifies cultural biases using causal social commonsense knowledge. Liao and Naghizadeh (2023) investigates fairness algorithms through social and data biases. Biester (2025) highlights gender biases in LLMs within the context of Olympic sports. Lee and Kim (2024) reduces bias and performance gaps in commonsense knowledge by replacing demographic-specific words with generic terms (e.g., "Chinese -> Asian -> People"). Davis (2024) points out issues in commonsense benchmarks, such as incorrect questions, unnatural language, and expert-knowledge requirements. While research often focuses on linguistic or cultural biases in reasoning datasets, underlying reasoning attributes and differences in non-reasoning commonsense datasets remain an overlooked source of bias. Therefore, it is necessary to clarify the reasoning attributes in commonsense questions and evaluate their impact on the training and assessment of commonsense benchmarks.

2.3 Evaluation reliability for benchmarks

Multiple-choice question answering (MCQA) is widely used in existing benchmarks to evaluate the capabilities of language models (Guo et al., 2023), but its reliability is increasingly being questioned. Wang et al. (2025) found that language models tend to select the least incorrect option rather than the distinctly correct answer when responding to MCQA. Additionally, Balepur et al. (2024) demonstrated that models can solve MCQA tasks even without the actual question, suggesting the need for stronger benchmark tests. To better understand model behavior, Wang et al. (2024) proposed directly analyzing the freely generated textual outputs of models instead of relying solely on the probability of the first token. In tasks involving reasoning, the quality of the reasoning process (Cobbe et al., 2021; Weng et al., 2023) and the number of samples (Wang et al., 2023; Lin et al., 2024) are closely related to the test results. Notably, most evaluation methods focus on numerical problems because their intermediate steps are easier to verify. However, this approach does not apply well to commonsense questions, which are mostly non-numerical knowledge-based problems. Therefore, there is a need for an automated method tailored to the characteristics of commonsense tasks to improve existing benchmarks and develop new evaluation metrics that comprehensively measure both knowledge and reasoning abilities.

3 Methodology

Commonsense benchmarks typically evaluate LLMs using multiple-choice questions to assess both knowledge and reasoning abilities. However, commonsense benchmarks are crafted with data that contains varying degrees of hidden reasoning attributes. This makes it challenging to determine whether a model's shortcomings lie in knowledge or reasoning. To address this issue, we propose *ReComSBench*, a framework that explicating hidden reasoning attributes based on the principle that "knowledge reasoning is the process of using known knowledge to infer new knowledge"(Chen et al., 2020), thereby enabling a deeper and more balanced evaluation of these abilities.

3.1 Reasoning attributes explicating

Given a commonsense question Q with options $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$, we aim to find the most representative reasoning path S^* from the set of generated paths $\mathcal{S} = \{S_1, S_2, \ldots, S_n\}$. Each path S_i consists of reasoning steps $\{s_{i1}, s_{i2}, \ldots, s_{im}\}$ and produces an answer $\hat{A}_i$. The knowledge behind the reasoning steps is represented by the set of extracted knowledge triplets $\mathcal{K}(S_i)$. To ensure both correctness and conciseness, the optimal reasoning path S^* is defined as:

$$S^* = \arg\min_{S_i \in \mathcal{S}} |\mathcal{K}(S_i)| \quad \text{subject to } \mathcal{A}(S_i) = A_{\text{gt}} \tag{1}$$

where:

- $\mathcal{A}(S_i)$ denotes the answer derived from reasoning path S_i,
- A_{gt} is the ground-truth answer,

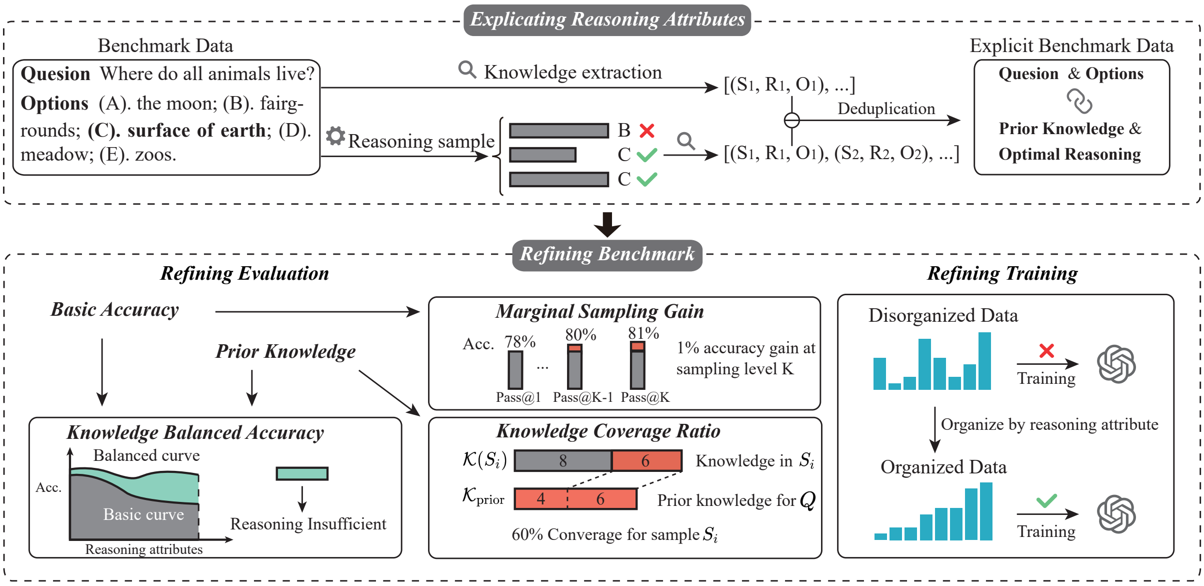


Figure 2: An overview of *ReComSBench*, which refines benchmarks with new metrics and hidden reasoning attributes. It explicates hidden reasoning attributes through optimal reasoning and prior knowledge for QA.

- $|\mathcal{K}(S_i)|$ measures the size of the knowledge set extracted from S_i.

This ensures that the selected reasoning path satisfies correctness ($\mathcal{A}(S_i) = A_{\text{gt}}$) while minimizing the amount of generated knowledge ($|\mathcal{K}(S_i)|$), minimizing the provision of unnecessary knowledge that chat-oriented LLMs tend to provide (Bian et al., 2024a). As shown in Figure 2, we generate reasoning paths using Chain-of-Thought (Wei et al., 2022) and Rejection Sampling. Knowledge involved in the reasoning process is extracted by LLM. For detailed prompts templates, please refer to Table 4 in Appendix A. From the path S_i, we extract knowledge $\mathcal{K}(S_i)$ and deduplicate overlapping knowledge with the question's inherent knowledge $\mathcal{K}(Q)$, yielding novel knowledge:

$$\mathcal{K}_{\text{new}}(S_i) = \mathcal{K}(S_i) \setminus \mathcal{K}(Q) \tag{2}$$

Importantly, only the $\mathcal{K}_{\text{new}}$ derived from the optimal reasoning path S^* is regarded as $\mathcal{K}_{\text{prior}}$, which represents the prior knowledge required to answer the question Q. This distinction ensures that the extracted knowledge is both minimal and essential for reasoning.

Then the reasoning difficulty of Q is defined as $d(Q) = |\mathcal{K}_{\text{prior}}|$. This metric quantifies the complexity of inference required to answer Q, guiding subsequent evaluation and training. While the randomness inherent in the generation of new knowledge during reasoning does not directly represent the problem itself, it can still be used on a macroscopic level to compare the differences in acquired knowledge from questions to measure their reasoning attributes (Bian et al., 2024b).

3.2 Refining benchmark in evaluation

In commonsense questions, knowledge attributes and reasoning attributes are tightly intertwined, and the underlying differences in reasoning attributes can vary significantly. To disentangle the model's actual performance on the benchmark, we designed distinct indicators focusing on knowledge evaluation and reasoning evaluation separately.

Knowledge Balanced Accuracy The Knowledge Balanced Accuracy (KBA) explicitly prompts the model with the knowledge required for the answer, avoiding the hidden reasoning attributes of the question and model's hidden reasoning.

We augment the original question Q with $\mathcal{K}_{\text{prior}}$ to construct $Q_{\text{aug}} = Q \oplus \mathcal{K}_{\text{prior}}$. The KBA is computed as:

$$\text{KBA} = \frac{1}{N} \sum_{i=1}^{N} I \left(\arg\max_{A \in \mathcal{A}} P(A|Q_{\text{aug}}^{(i)}) = A_{\text{gt}}^{(i)} \right) \tag{3}$$

where $I(\cdot)$ is the indicator function, N is the total number of samples, and $A_{\text{gt}}^{(i)}$ is the ground-truth answer for the i-th question. This metric provides necessary knowledge to isolate the model's reasoning ability. It allows for a purer evaluation of the model's ability to retrieve correct answers based on question knowledge and prior knowledge, excluding the reasoning attributes. Compared to the Accuracy, it can also assess the impact of reasoning attributes on model performance. We further discuss this point in Section 4.3.

Marginal Sampling Gain By sampling, we can start from the question, generate diverse intermediate reasoning processes, and eventually arrive at a solution. However, sampling not only increases computational costs but also does not guarantee that the correct answer will be obtained. To address this issue, we introduce Marginal Sampling Gain (MSG) as a metric to evaluate the overall sampling performance of the model in the sampling reasoning space.

$$\mathrm{MSG}(K) = \mathrm{Acc}(K) - \mathrm{Acc}(K-1) \qquad (4)$$

Here, $\mathrm{Acc}(K)$ represents the accuracy achieved after K sampling trials per question in the dataset. When $\mathrm{MSG}(K) < \tau$ (a predefined threshold), it indicates that the model has reached its limit of reasoning capacity improvement through additional sampling. This implies that the accuracy gain for the given benchmark is approximately bounded by $\mathrm{Acc}(K)$ at the marginal gain threshold τ. Consequently, K serves as a reasonable threshold for the number of sampling trials, beyond which further sampling returns in an unacceptable level of diminishing returns.

Knowledge Coverage Ratio The evaluation of the quality of single reasoning sampling is also critical. Numerical validation methods for assessing reasoning steps are not applicable to most commonsense problems, as these are mostly non-numerical. Therefore, the coverage of essential knowledge in the reasoning steps becomes a natural choice for evaluation.

For single sampling, the Knowledge Coverage Ratio (KCR) evaluates single-path reasoning quality:

$$\mathrm{KCR}(S_i) = \frac{|\mathcal{K}(S_i) \cap \mathcal{K}_{\text{prior}}|}{|\mathcal{K}_{\text{prior}}|} \qquad (5)$$

Here, the formula calculates the ratio of the intersection between the knowledge set $\mathcal{K}(S_i)$ derived from the reasoning path S_i and the prior knowledge set $\mathcal{K}_{\text{prior}}$, relative to the size of $\mathcal{K}_{\text{prior}}$. A higher KCR value indicates that the reasoning paths align more closely with the critical knowledge required for the task, ensuring high-quality reasoning.

3.3 Refining benchmark in training

To further improve training effectiveness, we partition the data into individual difficulty levels based on reasoning attributes. Inspired by curriculum learning (Bengio et al., 2009), we design a progressive training strategy that allows the model to transition gradually from simpler to more complex commonsense question-answering tasks. This structured approach outperforms random shuffled data distribution in handling data with varying reasoning difficulties.

Specifically, we define L difficulty levels $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_L$, where:

$$\mathcal{D}_l = \{Q \mid d(Q) = l\}. \qquad (6)$$

The training sequence follows:

$$\mathcal{D}_{\text{train}} = \mathcal{D}_1 \rightarrow \mathcal{D}_2 \rightarrow \cdots \rightarrow \mathcal{D}_L. \qquad (7)$$

During sampling, we use dynamic weighting to address data imbalance and ensure diversity.

4 Experiments and Analysis

4.1 Datasets and experimental setup

We evaluate our framework on two categories of commonsense benchmarks, which are knowledge-oriented and reasoning-oriented. **CommonsenseQA** (Talmor et al., 2019) and **OpenBookQA** (Mihaylov et al., 2018) focus on factual knowledge retrieval. Specifically, CommonsenseQA tests minimal reasoning over factual knowledge, while OpenBookQA combines core scientific facts with crowdsourced multiple-choice questions. In contrast, **ARC** (Clark et al., 2018) and **QASC** (Khot et al., 2020) emphasize complex multi-step reasoning. ARC contains challenging science questions requiring multi-step inference, and QASC involves integrating multiple facts for multi-hop inference. All datasets exhibit varying levels of hidden reasoning attributes, and only the challenge subset of ARC is used in our evaluation.

All experiments employ consistent prompts and are conducted on *Llama3.1-8B* (Dubey et al., 2024), *Gemma2-9B* (Rivière et al., 2024), *Gemma-7b* (Mesnard et al., 2024), and *Llama2-7B* (Touvron et al., 2023). We employ *LoRA* (Hu et al., 2022) for efficient training. For sampling, both greedy and random (with temperature 0.7) methods are used. Hidden reasoning attributes of commonsense data are generated by *Llama3.1-8B* and serve as the sole basis. Knowledge similarity for coverage

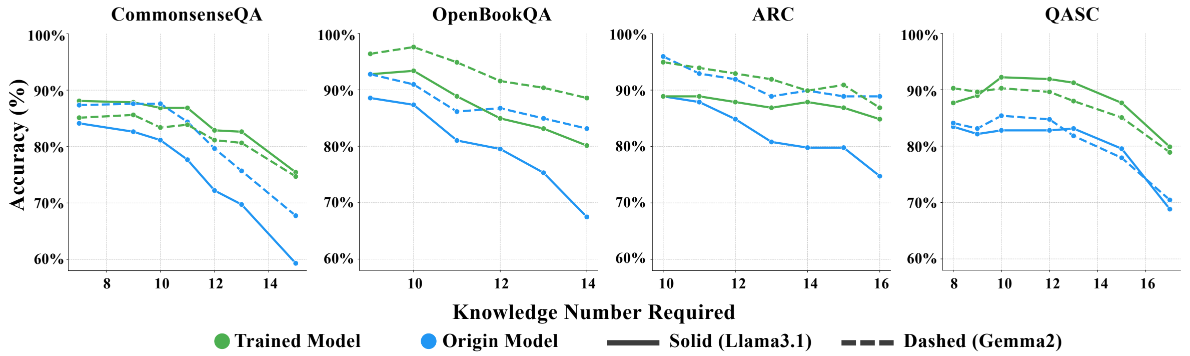


Figure 3: Sliding window accuracy of *Llama3.1* and *Gemma2* on commonsense benchmarks. The x-axis represents the knowledge number required to answer questions, calculated from K_{prior}.

calculation is computed using *all-MiniLm-L6-v2* (Wang et al., 2020).

4.2 Impact analysis of hidden reasoning attributes

We analyze the accuracy changes of different models across reasoning difficulties $d(Q)$ to examine the impact of hidden reasoning attributes. The validation set is sorted by $d(Q)$, from easy to hard. A sliding window approach is used to calculate LLM accuracy without reasoning: the window length is one-third of the dataset size, and the step size is one-third of the window length. The accuracy difference between the first window (starting point, Easy part) and the last window (endpoint, Hard part) reflects model performance on data with varying hidden reasoning attributes. The Easy part contains more low-reasoning data, while the Hard part contains more high-reasoning data.

In Figure 3, the y-axis shows accuracy, and the x-axis shows knowledge levels corresponding to $d(Q)$. Both *Llama3.1* and *Gemma2* exhibit declining accuracy as $d(Q)$ increases across datasets. This highlights the consistent correlation between hidden reasoning difficulty and lower accuracy in LLM benchmarks. Traditional benchmarks often overlook this, making it hard to analyze reasoning and knowledge proportions in incorrect responses based on basic accuracy alone.

Further experiments in Table 1 and Table 3 show that the accuracy gap between Easy and Hard cases persists post-training. In CommonsenseQA, for *Llama3.1*, the accuracy gap is 24.8% pre-training and 12.7% post-training, with accuracy dropping from 84.1% (Easy) to 59.3% (Hard). Significant differences exist for both knowledge-oriented and reasoning-oriented benchmarks, emphasizing the importance of hidden reasoning properties. These findings confirm that hidden reasoning influences all aspects of model evaluation and training.

Dataset	Model	Accuracy (%)		Difference (%)
		Easy	Hard	
CommonsenseQA	llama3.1	84.1	59.3	24.8
	llama3.1†	88.1	75.4	12.7
	gemma2	87.3	67.7	19.6
	gemma2†	85.1	74.7	10.4
OpenBookQA	llama3.1	88.6	67.5	21.1
	llama3.1†	92.8	80.1	12.7
	gemma2	92.8	83.1	9.7
	gemma2†	96.4	88.6	7.8
ARC	llama3.1	88.9	74.7	14.2
	llama3.1†	88.9	84.8	4.1
	gemma2	96.0	88.9	7.1
	gemma2†	94.9	86.9	8.0
QASC	llama3.1	83.4	68.8	14.6
	llama3.1†	87.7	79.9	7.8
	gemma2	84.1	70.5	13.6
	gemma2†	90.3	78.9	11.4

Table 1: Sliding window accuracy of *Llama3.1* and *Gemma2* on different datasets (†indicates trained models). The sliding window progresses from Easy (first window) to Hard (last window).

4.3 New metrics in ReComSBench

Metric 1: Knowledge Balanced Accuracy KBA evaluates models' commonsense knowledge capabilities by decoupling the assessment of commonsense knowledge from reasoning demands through explicit knowledge prompting. During prompting, necessary prior knowledge is explicitly passed to the model to support factual commonsense answering, thereby bypassing hidden reasoning.

We systematically tested *Llama2*, *Llama3.1*, *Gemma*, and *Gemma2* models. To mitigate variance from stochastic knowledge selection, all knowledge generated as standard snippets was incorporated into prompts. KBA demonstrates its ability to evaluate knowledge while mitigating the

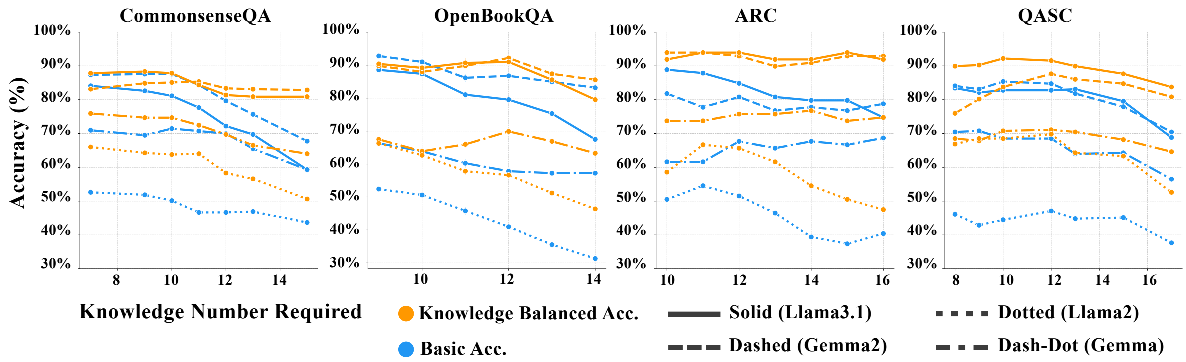


Figure 4: KBA curves and basic accuracy curves of *Llama* and *Gemma* families on commonsense benchmarks

influence of hidden reasoning attributes in the data. As Figure 4 demonstrates, The KBA curve consistently surpasses and is flatter than the basic accuracy curve across all datasets, confirming its effectiveness in isolating knowledge assessment from reasoning demands. The alignment of KBA and basic accuracy curve trends across model generations confirms KBA's equivalent analytical power. By analyzing the differences between KBA and basic accuracy curves at easy and hard parts, we can identify whether knowledge or reasoning has a greater impact on accuracy. Larger gaps in the easy part indicate insufficient knowledge, while larger gaps in the hard part suggest insufficient reasoning. On commonsense benchmarks, previous-generation models had deficiencies in both areas, while advanced-generation models show more reasoning limitations. These all confirm that KBA has unique diagnostic value and can evaluate the model from a broader and deeper perspective. For more numerical details, please refer to Table 5 in Appendix B.

Dataset	Model	MSG(K) (%)				Sum
		K=2	K=3	K=4	K=5	
CommonsenseQA	llama2	13.4	6.5	4.7	3.0	27.6
	llama3.1	9.4	4.0	2.4	1.9	17.7
	gemma	5.4	3.1	1.9	0.9	11.3
	gemma2	5.8	3.0	1.1	1.1	11.0
OpenBookQA	llama2	11.2	8.0	4.2	2.4	25.8
	llama3.1	8.6	3.4	2.8	0.6	15.4
	gemma	6.4	3.8	2.2	3.4	15.8
	gemma2	7.8	2.6	1.4	0.8	12.6
ARC	llama2	12.0	9.3	5.1	6.0	32.4
	llama3.1	7.7	2.4	1.3	0.7	12.1
	gemma	6.7	1.6	1.7	2.0	12.0
	gemma2	6.4	3.0	1.0	1.0	11.4
QASC	llama2	12.6	6.7	4.1	4.3	27.7
	llama3.1	14.7	4.5	2.3	1.0	22.5
	gemma	6.2	3.4	1.7	1.6	12.9
	gemma2	9.9	4.9	1.6	1.4	17.8

Table 2: MSG and sum for different models on commonsense benchmarks

Metric 2: Marginal Sampling Gain An ideal high-performance model maintains low MSG values at high accuracy levels, demonstrating confidence. Conversely, the combination of low accuracy with high MSG indicates suboptimal model performance. We sample K times of inference on models in the commonsense benchmark, where the first sampling is greedy sampling, and calculate the model accuracy under pass@K and MSG(K). As show in Table 2, our analysis of *Llama* and *Gemma* model families reveals progressively diminishing MSG values across iterations. Specifically, when $K = 5$, the improvement in accuracy is close to 1%. Notably, advanced models in each series demonstrate lower MSG values indicating enhanced confidence (e.g., MSG(3): *Llama3.1* at 2.3% vs. *Llama2* at 9.3% in ARC). The difference in MSG metric is consistent with the performance differences of different generations of models. This is because MSG metric effectively evaluate the model's sampling level in the reasoning sampling space.

Metric 3: Knowledge Coverage Ratio KCR can effectively evaluate the quality of sampled commonsense reasoning. In our experiments, we calculated the knowledge coverage of all inferences made by the *Llama3.1* model on the commonsense benchmarks, with a sampling size of 5. The similarity threshold for determining whether knowledge is similar was set to 0.75. Based on the correctness of answer, we grouped the data into correct and incorrect groups and plotted the boxplots shown in Figure 5. In the boxplots, the median knowledge coverage of the correct group is consistently higher than that of the incorrect group across all four datasets. Additionally, the U-statistic test indicates a substantial advantage for the correct group, with $p < 0.05$. These results demonstrate the effectiveness of knowledge coverage as a metric for

Method	CommonsenseQA (%)				OpenBookQA (%)				ARC (%)				QASC (%)			
	Acc.	KBA	Δ	Δ^*	Acc.	KBA	Δ	Δ^*	Acc.	KBA	Δ	Δ^*	Acc.	KBA	Δ	Δ^*
Base	73.2	83.8	24.8	6.9	79.4	87.2	21.1	10.8	81.3	92.0	14.1	0.0	78.0	88.2	14.6	6.2
RandSample	82.4	87.1	14.6	8.9	86.4	92.8	9.6	6.0	81.9	90.6	7.1	2.0	84.4	89.0	8.4	6.8
Score-CL	81.4	87.1	15.1	7.9	86.4	**93.2**	12.7	5.4	**85.6**	90.7	5.1	4.0	86.3	**90.2**	9.7	2.6
Reason-CL	**82.7**	**88.2**	**13.4**	7.9	**86.8**	92.8	**7.2**	5.4	85.3	**92.3**	**1.0**	5.1	**86.6**	88.0	**7.5**	4.5

Table 3: Performance comparison of different training strategies (Score-CL: score-based curriculum learning using model's negative log-likelihood scores; Reason-CL: reasoning-based curriculum learning) across four datasets. Metrics include: Accuracy (Acc.), Knowledge Balanced Accuracy (KBA), Easy/Hard accuracy difference (Δ), and its knowledge balanced version (Δ^*).

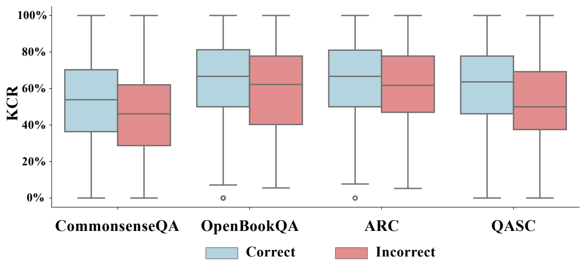


Figure 5: Boxplot of Knowledge Coverage Ratio differences between correct and incorrect reasoning groups on commonsense benchmarks

evaluating reasoning quality and highlight the importance of knowledge generation during the reasoning process.

4.4 Stratified data for training

To evaluate the effectiveness of difficulty stratification based on reasoning attributes, we conducted experiments using the *Llama3.1* model as the base model. We compared four training strategies: (1) base model performance, (2) random sampling, (3) curriculum learning based on data score difficulty, and (4) curriculum learning based on data reasoning difficulty. Here, data reasoning difficulty was defined by the number of knowledge elements in hidden reasoning attributes (proposed in this study), while data score difficulty was calculated using the negative log-likelihood scores of correct answers from *Llama3.1*, following the approach of Maharana and Bansal (2022).

As shown in Table 3, training with difficulty stratification based on reasoning attributes achieves performance improvements comparable to those of model-probability-based stratification. By leveraging the hidden reasoning attributes in the data, the model performs stronger on datasets (e.g., CommonsenseQA, OpenBookQA) that require hidden reasoning perception. Notably, across all datasets, the model trained with hidden reasoning attributes exhibits the smallest difference δ between Easy and Hard accuracies, indicating its enhanced focus on high-reasoning-difficulty samples. This demonstrates the method's generality and effectiveness in improving reasoning capabilities. Thus, these results indicate that integrating hidden reasoning attributes into data organization strategies may enhance model performance and reasoning capabilities.

5 Conclusion

Simple commonsense data may still require reasoning to arrive at the correct answer, which aligns with the hidden reasoning phenomena observed in LLMs. This characteristic makes existing commonsense benchmarks insufficient for distinguishing whether a model's poor performance is due to a lack of commonsense knowledge or inadequate reasoning ability. In this study, we explored the hidden reasoning attributes within commonsense benchmarks. Our findings confirmed that these attributes significantly impact the evaluation and training of a model's commonsense capabilities. To address this challenge, we proposed *ReComSBench*, a framework for refining existing commonsense benchmarks. *ReComSBench* transforms the differences in hidden reasoning attributes within benchmark data into explicit representations of reasoning and knowledge. It not only identifies variations in reasoning difficulty of "simple" commonsense QA but also introduces three specialized metrics designed to decouple and deeply evaluate a model's commonsense knowledge and reasoning abilities. Through experiments, we validated the effectiveness of these metrics and demonstrated the feasibility of leveraging the hidden reasoning attributes in benchmark data to enhance a model's commonsense capabilities.

Limitations

The limitations of the proposed method lie in the fact that a Large Language Model is used to automatically generate the prior knowledge required for answering questions. Thus, this approach is still not entirely model-independent. Compared to methods that assess question difficulty based on model probabilities, the difference in overall performance improvement is less significant than expected, although it still shows advantages on reasoning-related data. Moreover, the prior knowledge generated by the model does not fully represent the actual prior knowledge required for the questions. However, within the scope of benchmark data, it can still reflect the overall reasoning properties and differences of the data. Additionally, the Marginal Sampling Gain (MSG) metric involves randomness in sampling, leading to potential result fluctuations, though these still indicate model sampling performance. For future work, extending ReComSBench to areas such as empathetic dialogue or legal reasoning could test its generalizability and improve the metrics.

Ethical Considerations

Our work aims to improve the evaluation of LLMs' commonsense abilities, which could lead to more reliable and robust AI systems. However, there are potential ethical concerns that warrant discussion. First, the use of LLMs for generating prior knowledge may inadvertently propagate biases present in the training data. To mitigate this, we recommend incorporating diverse datasets and regularly auditing model outputs for fairness and inclusivity. Second, our framework relies on benchmark datasets that may not fully represent real-world scenarios. Therefore, when applying the evaluation results to real-world application scenarios, the specific needs and limitations of the target domain need to be carefully considered.

References

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. 2024. Artifacts or abduction: How do llms answer multiple-choice questions without the question? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 10308–10330. Association for Computational Linguistics.

Lisa Bauer, Hanna Tischer, and Mohit Bansal. 2023. Social commonsense for explanation and cultural bias discovery. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 3727–3742. Association for Computational Linguistics.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pages 41–48. ACM.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2020. Abductive Commonsense Reasoning. *Preprint*, arXiv:1908.05739.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, Ben He, Shanshan Jiang, and Bin Dong. 2024a. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 3098–3110. ELRA and ICCL.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, Ben He, Shanshan Jiang, and Bin Dong. 2024b. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 3098–3110. ELRA and ICCL.

Laura Biester. 2025. Sports and women's sports: Gender bias in text generation with olympic data. *Preprint*, arXiv:2502.04218.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. *Preprint*, arXiv:1906.05317.

Ronald J. Brachman and Hector J. Levesque. 2022. Toward a New Science of Common Sense. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12245–12249.

Xiaojun Chen, Shengbin Jia, and Yang Xiang. 2020. A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications*, 141:112948.

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. *CoRR*, abs/1803.05457.

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

Ernest Davis. 2024. Benchmarks for Automated Commonsense Reasoning: A Survey. *ACM Computing Surveys*, 56(4):1–41.

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.

Quyet V. Do, Junze Li, Tung-Duong Vuong, Zhaowei Wang, Yangqiu Song, and Xiaojuan Ma. 2024. What Really is Commonsense Knowledge? *Preprint*, arXiv:2411.03964.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating large language models: A comprehensive survey. *CoRR*, abs/2310.19736.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. QASC: A dataset for question answering via sentence composition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8082–8090. AAAI Press.

Jinkyu Lee and Jihie Kim. 2024. Improving commonsense bias classification by mitigating the influence of demographic terms. *IEEE Access*, 12:161480–161489.

Yiqiao Liao and Parinaz Naghizadeh. 2023. Social bias meets data bias: The impacts of labeling and measurement errors on fairness criteria. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 8764–8772. AAAI Press.

Lei Lin, Jia-Yi Fu, Pengli Liu, Qingyang Li, Yan Gong, Junchen Wan, Fuzheng Zhang, Zhongyuan Wang, Di Zhang, and Kun Gai. 2024. Just ask one more time! self-agreement improves reasoning of language models in (almost) all scenarios. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 3829–3852. Association for Computational Linguistics.

Adyasha Maharana and Mohit Bansal. 2022. On curriculum learning for commonsense reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 983–992. Association for Computational Linguistics.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Cristian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, and et al. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

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, Brussels, Belgium, October 31 - November 4, 2018*, pages 2381–2391. Association for Computational Linguistics.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozinska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucinska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjösund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, and Lilly McNealus. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartlomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cèsar Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan J. Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, François Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse H. Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, José Hernández-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory W. Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, María José Ramírez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael I. Ivanitskiy, Michael Starritt, Michael Strube, Michal Swedrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T., Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Milkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima (Shammie) Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay V. Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Trans. Mach. Learn. Res.*, 2023.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. 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, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *CoRR*, abs/2306.07899.

Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. 2025. Llms may perform MCQA by selecting the least incorrect option. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 5852–5862. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep selfattention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. "my answer is c": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 7407–7416. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 2550–2575. Association for Computational Linguistics.

Tian Ye, Zicheng Xu, Yuanzhi Li, and Zeyuan Allen-Zhu. 2024. Physics of Language Models: Part 2.1, Grade-School Math and the Hidden Reasoning Process. *ArXiv e-prints*, abs/2407.20311. Full version available at `http://arxiv.org/abs/2407.20311`.

Wenting Zhao, Justin Chiu, Claire Cardie, and Alexander Rush. 2023a. Abductive Commonsense Reasoning Exploiting Mutually Exclusive Explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14883–14896, Toronto, Canada. Association for Computational Linguistics.

Zirui Zhao, Wee Sun Lee, and David Hsu. 2023b. Large language models as commonsense knowledge for large-scale task planning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

A Prompt Templates

In this appendix, as show on Figure 4, we list the prompt templates used in this document along with their corresponding purposes. Large language models may be sensitive to differences in prompts, so we use a consistent prompt template.

Prompt Template and Purpose
Template: Please read the multiple-choice question below carefully and select ONE of the listed options. Provide the final answer starting with 'The correct answer is OPTION'. {QA}. **Purpose:** To guide the model directly choose the answer.
Template: Please read the multiple-choice question below carefully and select ONE of the listed options. Let's think step by step. Each step should start with 'THOUGHT:'. After all thoughts, provide the final answer starting with 'The correct answer is OPTION'. {QA}. **Purpose:** To guide the model choose the answer inferentially.
Template: "Please read the multiple-choice question below carefully and select ONE of the listed options. Provide the final answer starting with 'The correct answer is OPTION'. Knowledge hints: {HINT}\n{QA}". **Purpose:** To guide the model choose the answer under the knowledge hints.
Template: You are an expert in knowledge extraction. Please extract knowledge from text in the form of triples (subject, predicate, object). Guidelines: 1. Extract only knowledge explicitly stated in the text. Do not infer or derive information from context, common sense, or options unless explicitly mentioned. 2. Avoid overgeneralization or assumptions. Stick strictly to what is directly expressed in the text. 3. If no knowledge is extractable, return an empty list. Format: Return the extracted knowledge in JSON format under the key extracted_knowledge. Use an empty list if no knowledge is extractable. Examples: {FEW_SHOT} Now, extract knowledge from the following text: {TEXT}. **Purpose:** To guide the model so that it can extract knowledge properly and in a valid style.

Table 4: Prompt templates and their purposes

B Details of Experiments

We provide additional details of the experimental results here. Table 5 shows the numerical data corresponding to Figure 4. By comparing the differences (diff), we observe that the accuracy changes are generally smaller after knowledge balancing. Moreover, the improvement in KBA overall accuracy is more concentrated in the Hard part, where the Hard part's accuracy increases more than the Easy part, making the KBA curve in Figure 4 flatter. We define the Easy and Hard parts as the first and last window values, rather than the maximum and minimum values within the sliding window. These findings demonstrate that the KBA metric provides additional insights into model performance beyond standard accuracy.

Table 6 additionally shows the pass@K ($Acc(K)$) required before computing MSG. For the Knowledge Coverage Ratio, the U statistic is significant, as shown in Figure 7. The horizontal axis is the similarity threshold that measures whether the knowledge is similar. It can be seen that the advantage is significant under most thresholds. We also analyzed the redundancy of knowledge, defined as the proportion of dissimilar knowledge generated during inference. As shown in Figure 6, correct groups have higher redundancy. However, since redundancy has no upper limit and increases with more generated knowledge, its reference value is slightly lower than coverage.

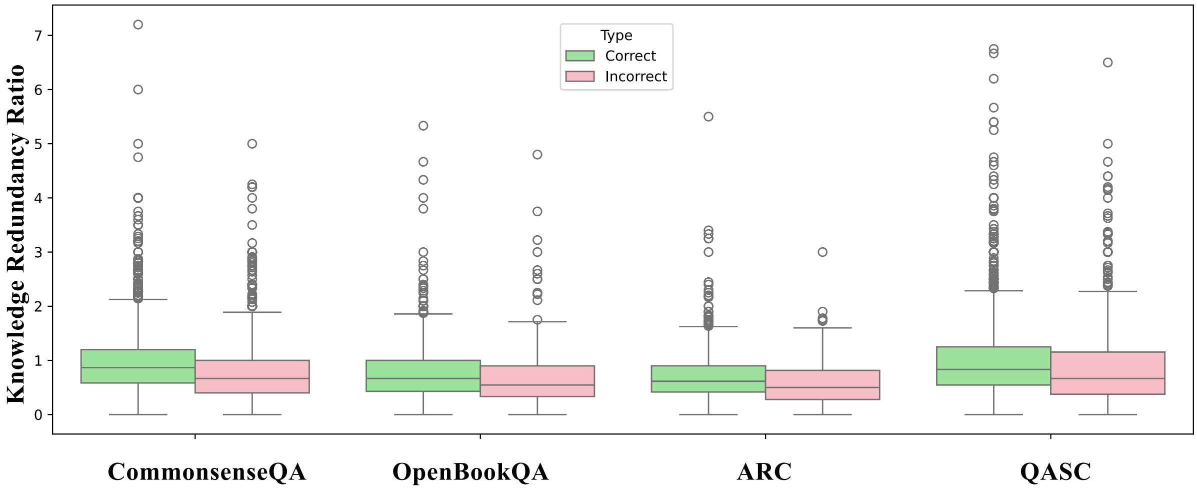
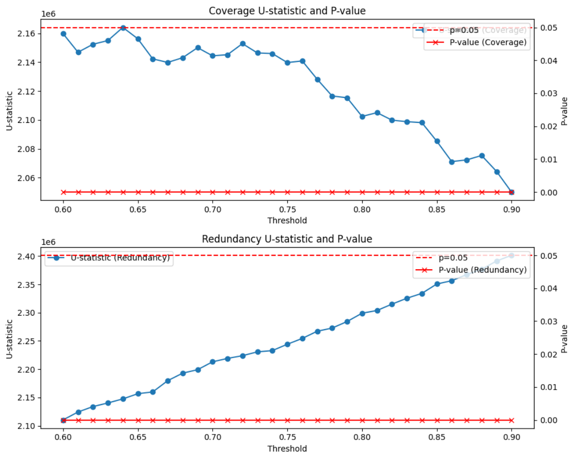
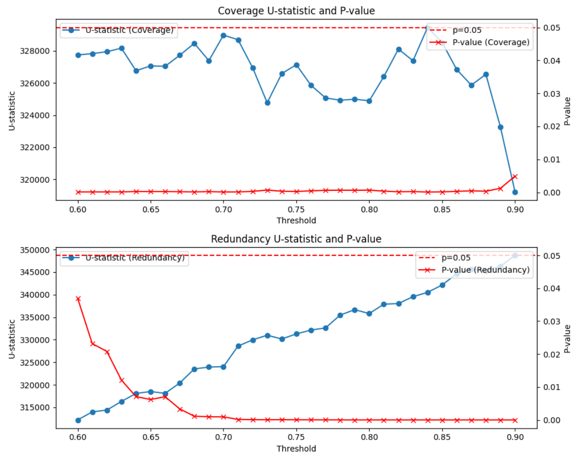


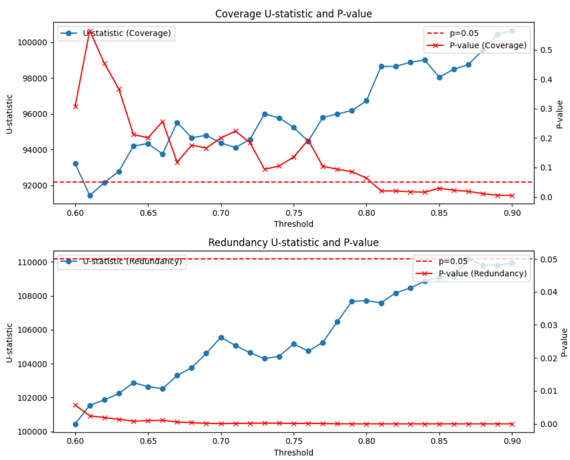
Figure 6: Boxplot of Knowledge Redundancy Ratio differences between correct and incorrect reasoning groups on commonsense benchmarks



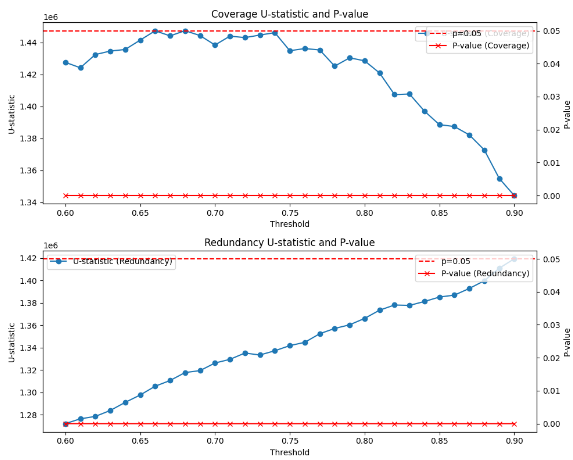
(a) CommonsenseQA Dataset



(b) OpenBookQA Dataset



(c) ARC Dataset



(d) QASC Dataset

Figure 7: U statistic for knowledge coverage (upper) and redundancy (lower) under different similarity thresholds in four datasets. The left axis shows statistical advantage, while the right axis shows P values.

Dataset	Model	Accuracy (%)				KBA (%)			
		Overall	Easy	Hard	Diff	Overall	Easy	Hard	Diff
CommonsenseQA	llama2	47.4	52.6	43.7	8.9	60.3	66.0	50.6	15.4
	llama3.1	73.2	84.1	59.3	24.8	83.8	87.8	80.9	6.9
	gemma	66.6	71.0	59.3	11.7	70.6	75.9	64.0	11.9
	gemma2	79.7	87.3	67.7	19.6	83.6	83.1	82.9	0.2
OpenBookQA	llama2	42.8	52.4	31.3	21.1	56.4	66.3	46.4	19.9
	llama3.1	79.4	88.6	67.5	21.1	87.2	90.4	79.5	10.8
	gemma	61.0	66.3	57.2	9.1	65.8	67.5	63.3	4.2
	gemma2	87.0	92.8	83.1	9.7	88.4	92.8	83.1	9.6
ARC	llama2	45.8	50.5	40.4	10.1	56.2	58.6	47.5	11.1
	llama3.1	81.3	88.9	74.7	14.1	92.0	91.9	91.9	0.0
	gemma	65.2	61.6	68.7	-7.1	74.9	73.7	74.7	-1.0
	gemma2	91.3	96.0	88.9	7.1	92.3	93.9	92.9	1.0
QASC	llama2	43.5	46.1	37.7	8.4	62.7	66.9	52.6	14.3
	llama3.1	78.0	83.4	68.8	14.6	88.2	89.9	83.8	6.2
	gemma	65.0	70.5	56.5	14.0	67.8	68.5	64.6	3.9
	gemma2	79.6	84.1	70.5	13.6	81.4	76.0	80.8	-4.9

Table 5: Accuracy and KBA for different models on commonsense benchmarks

Dataset	Model	Accuracy (%)					MSG(K) (%)			
		pass@1	pass@2	pass@3	pass@4	pass@5	K=2	K=3	K=4	K=5
CommonsenseQA	llama2	52.8	66.2	72.7	77.4	80.4	13.4	6.5	4.7	3.0
	llama3.1	71.0	80.4	84.4	86.8	88.7	9.4	4.0	2.4	1.9
	gemma	65.4	70.8	73.9	75.8	76.7	5.4	3.1	1.9	0.9
	gemma2	75.4	81.2	84.2	85.3	86.4	5.8	3.0	1.1	1.1
OpenBookQA	llama2	53.4	64.6	72.6	76.8	79.2	11.2	8.0	4.2	2.4
	llama3.1	79.8	88.4	91.8	94.6	95.2	8.6	3.4	2.8	0.6
	gemma	61.6	68.0	71.8	74.0	77.4	6.4	3.8	2.2	3.4
	gemma2	80.0	87.8	90.4	91.8	92.6	7.8	2.6	1.4	0.8
ARC	llama2	50.2	62.2	71.5	76.6	82.6	12.0	9.3	5.1	6.0
	llama3.1	82.9	90.6	93.0	94.3	95.0	7.7	2.4	1.3	0.7
	gemma	65.9	72.6	74.2	75.9	77.9	6.7	1.6	1.7	2.0
	gemma2	83.6	90.0	93.0	94.0	95.0	6.4	3.0	1.0	1.0
QASC	llama2	43.1	55.7	62.4	66.5	70.8	12.6	6.7	4.1	4.3
	llama3.1	69.9	84.6	89.1	91.4	92.4	14.7	4.5	2.3	1.0
	gemma	61.4	67.6	71.0	72.7	74.3	6.2	3.4	1.7	1.6
	gemma2	66.8	76.7	81.6	83.2	84.6	9.9	4.9	1.6	1.4

Table 6: Accuracy and MSG for different models on commonsense benchmarks