Compound-QA: A Benchmark for Evaluating LLMs on Compound Questions

Anonymous ACL submission

Abstract

Large language models (LLMs) demonstrate remarkable performance across various tasks, prompting researchers to develop diverse evaluation benchmarks. However, existing benchmarks typically measure the ability of LLMs to respond to individual questions, neglecting the complex interactions in real-world applications. In this paper, we introduce Compound Question Synthesis (CQ-Syn) to create the Compound-QA benchmark, focusing on compound questions with multiple sub-questions. This benchmark is derived from existing QA datasets, annotated with proprietary LLMs and verified by humans for accuracy. It encompasses five categories: Factual-Statement, Cause-and-Effect, Hypothetical-Analysis, Comparison-and-Selection, and Evaluation-and-Suggestion. It evaluates the LLM capability in terms of three dimensions including understanding, reasoning, and knowledge. Our assessment of seven open-source LLMs using Compound-QA reveals distinct patterns in their responses to compound questions, which are significantly poorer than those to non-compound questions. Additionally, we investigate various methods to enhance LLMs performance on compound questions. The results indicate that these approaches significantly improve the models' comprehension and reasoning abilities on compound questions.

1 Introduction

Large language models (LLMs) have achieved remarkable success in natural language processing (NLP), demonstrating exceptional performance across a wide range of tasks due to their advanced language understanding, reasoning, and generation capabilities (Achiam et al., 2023; Dubey et al., 2024; Ouyang et al., 2022; Team et al., 2024; Guo et al., 2025; El-Kishky et al., 2025). Existing benchmarks evaluate these models' abilities across various dimensions (Kwan et al., 2024; Zhou et al., 2023; He et al., 2024a; Li et al., 2023), such as understanding (Bartolo et al., 2020; Li et al., 2023), reasoning (Yang et al., 2018; Zhu et al., 2024; Wen et al., 2024b), and knowledge (Liu et al., 2023b; Jin et al., 2019). However, these benchmarks primarily evaluate responses to individual questions or instructions, overlooking the complexity of real-world interactions (He et al., 2024c).

In real-world scenarios, users often ask a series of interrelated questions within a single query, expecting to obtain a comprehensive and precise response for each question, as illustrated in Figure 1. We refer to this as Compound Questions, which include multiple sub-questions within a single turn. These sub-questions, which may be correlated (Section 3.1). This question format is common in human-AI interactions and agent-based scenarios, where tasks are decomposed into sub-instructions that require individual responses. While humans can effectively address compound questions by answering each sub-question separately without omission or interference, LLMs face challenges such as identification of sub-questions and the elliptical phenomena in natural language (van Craenenbroeck and Temmerman, 2018). Adjacent questions and answers can cause LLMs to focus on earlier context while overlooking unanswered sub-questions. Since LLMs are susceptible to irrelevant context, they may also be influenced by other sub-questions and their answers (Wu et al., 2024).

Recent studies explore the ability of LLMs in handling multiple-problem tasks (Wang et al., 2024; Wen et al., 2024a; Liu et al., 2024b; Chen et al., 2024). However, most studies primarily focus on classification or fixed-answer tasks (Wang et al., 2024; Liu et al., 2024b; Chen et al., 2024). Furthermore, the inter-question relationships they examine are generally simple, typically involving either concatenating questions (Wang et al., 2024; Liu et al., 2024b) or using sequential instructions where one answer influences the next (Chen et al.,

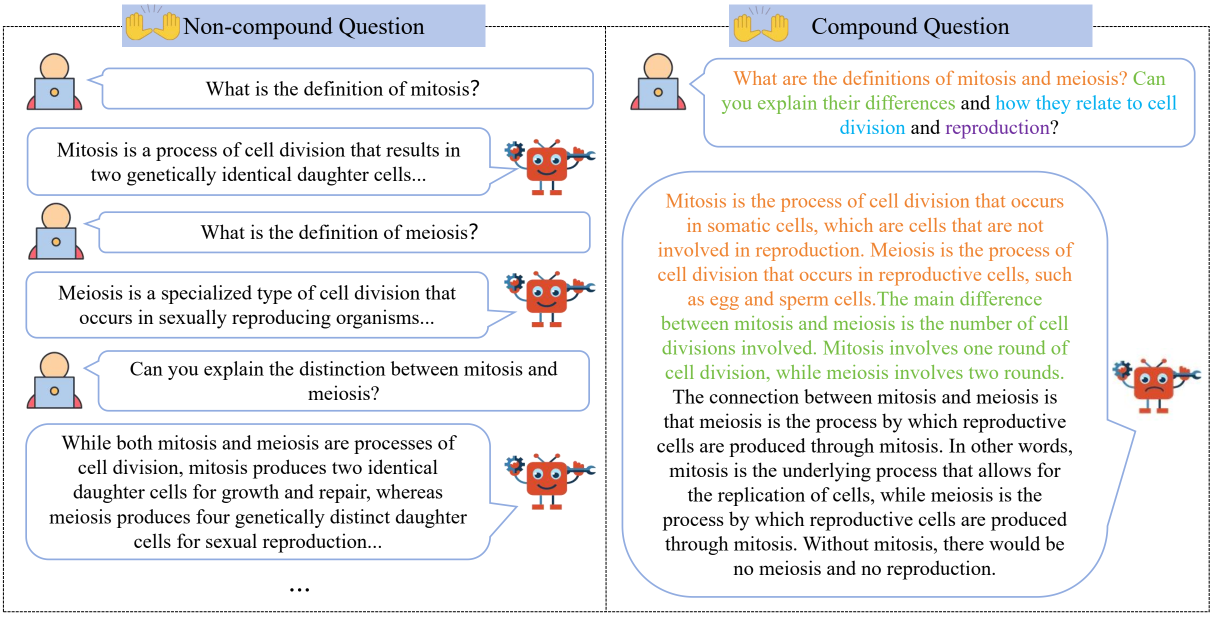

Figure 1: Examples of non-compound and compound questions. (Left) The user asks multiple questions sequentially in a multi-turn dialogue. (Right) The user asks several questions at once in a single turn.

2024). In contrast, our work tackles both issues by targeting open-ended QA tasks and by abstracting the complex, real-world logical dependencies among sub-questions.

To this end, we introduce a data synthesis framework called Compound Question Synthesis (CQ-Syn). This framework leverages LLM to generate and refine compound questions according to carefully developed guidelines, followed by a thorough human review to ensure quality. Using CQ-Syn, we construct the Compound-QA benchmark, which is designed to evaluate LLMs' ability to handle compound questions. This benchmark consists of 1,500 compound questions covering scenarios in language understanding, reasoning, and knowledge, divided into five types: Factual-Statement, Cause-and-Effect, Hypothetical-Analysis, Comparison-and-Selection, and Evaluation-and-Suggestion. Comprehensive experiments on seven mid-sized open-source LLMs reveal that their effectiveness on compound questions is significantly lower than on single-question tasks, highlighting their current limitations in handling multi-step reasoning and contextual integration. However, supervised fine-tuning with instruction data augmented by compound questions substantially improves this performance. We anticipate that this work will encourage further research and advancements to enhance the ability of LLMs to answer compound questions.[1]

[1]The dataset utilized in this study will be made publicly available to foster continued research.

2 Related Work

Evaluation of Large Language Models LLMs present superior performance on various tasks, such as question answering (Joshi et al., 2017; Yang et al., 2018; Malaviya et al., 2023), math reasoning (Cobbe et al., 2021; Hendrycks et al., 2021b), and multi-turn dialogues (Bai et al., 2024; Duan et al., 2023; Reddy et al., 2019). Previous works (Li et al., 2023; Zheng et al., 2023; Hendrycks et al., 2021a) propose different datasets and benchmarks to evaluate the capabilities of LLMs in terms of language understanding, reasoning and knowledge (He et al., 2024a). Models with better language understanding ability are preferred in tasks like reading comprehension, text classification, and multi-turn dialogue. It can be evaluated with datasets like SQuAD (Rajpurkar et al., 2018), QuAC (Choi et al., 2018) and RACE (Lai et al., 2017). The reasoning ability is important for LLM-based applications. The dataset of math reasoning (Cobbe et al., 2021; Hendrycks et al., 2021b), logical reasoning (Suzgun et al., 2023), and commonsense reasoning (Talmor et al., 2019; Geva et al., 2021) are widely used for evaluation of LLMs. Knowledge is another important capability for LLMs to survive on knowledge-intensive tasks. Benchmarks like MMLU (Hendrycks et al., 2021a), AGI-Eval (Zhong et al., 2023), and TriviaQA (Joshi et al., 2017) evaluate the disciplinary and world knowledge of different models.

As LLMs continue to demonstrate impressive performance on routine NLP tasks, researchers are now evaluating their ability to tackle complex tasks. Datasets like HotpotQA (Yang et al., 2018), ExpertQA (Malaviya et al., 2023), and GPQA (Rein et al., 2023) are used to assess how well models handle complex questions. In addition, researchers explore the ability of LLMs to follow complex instructions with multiple constraints. (He et al., 2024d,b; Aksu et al., 2023). Hu et al. (2024) proposes a sequential instruction tuning method aimed at enhancing the ability of LLMs to manage complex tasks. Wen et al. (2024a) construct ComplexBench by employing hierarchical constraints (format, semantics) and compositional question types (chain, selection) to evaluate how well models follow complex instructions, Liu et al. (2024b) introduce LongGenBench by synthesizing three datasets, MMLU (Hendrycks et al., 2021a), GSM8K (Cobbe et al., 2021), and CSQA (Talmor et al., 2019), to systematically evaluate LLMs on sequential concatenated questions.

Current research on LLMs handling multiple questions focuses more on classification or fixed-answer tasks (Wang et al., 2024; Chen et al., 2024; Liu et al., 2024b). In contrast, our work targets real-world QA scenarios featuring open-ended questions. By reviewing diverse sources, including QA datasets, consultation manuals, and social science frameworks, we identified common patterns in sub-question relationships. This analysis underpins our proposal of five questioning strategies to evaluate LLMs' ability to tackle compound questions with complex logical dependencies, addressing key limitations of existing benchmarks.

LLM-based Data Synthesis and Verification With the advancement of LLMs, the demand for high-quality annotated data continues to grow. In this context, synthetic data generated by models or algorithms rather than directly by humans (Long et al., 2024; Liu et al., 2024a) is gradually showing its great potential. Researchers have proposed various data synthesis methods to address this trend. For example, Persona Hub (Chan et al., 2024) is a character-driven synthesis approach that organizes one billion distinct personas from online data and uses LLMs to generate questions in different voices, yielding diverse synthetic data for various scenarios. KPDDS (Huang et al., 2024) synthesizes QA pairs using key points and example pairs from real data sources. Xu et al. (2024) introduces MAGPIE, a large-scale data alignment self-synthesis method. However, existing generative methods can produce inaccurate, low-quality, or incoherent data (Lupidi et al., 2024), potentially leading to model collapse (Dohmatob et al., 2024), underscoring the importance of quality screening for synthetic data. MoDS (Du et al., 2023) and Deita (Liu et al., 2023a) employ multiple dimensions for data filtering, utilizing techniques like model scoring. Recently, PROX (Zhou et al., 2024) treats the data refinement task as a programming task for data screening, while Source2Synth (Lupidi et al., 2024) synthesizes data from real data sources and introduces automated filtering and completion mechanisms during the generation process to ensure data quality. Recent work by Tao et al. (2024) proposes a comprehensive framework for detecting LLM-generated texts, emphasizing the importance of rigorous evaluation in multilingual and operationally diverse scenarios. In our study, we adopt a combined approach of model-based filtering and manual verification for each data instance to ensure the construction of high-quality datasets.

3 Compound-QA Benchmark

3.1 Types of Compound Questions

The compound question incorporates multiple sub-questions within a single query. It is common in human-AI interaction where users might propose several questions at one time. The ability to respond to compound questions is also important in agentic applications where several sub-instructions derived from the task decomposition and planning are expected to be followed. However, it is not trivial for LLMs as they might suffer from the problem of sub-question omission and a degradation in the quality of response (see examples in Figure 1).

The compound question presents several notable characteristics:

- **Hierarchical relevance** There may be hierarchical relevance among sub-questions, requiring respondents to not only understand each sub-question individually but also to recognize their hierarchical relationships to ensure coherence and completeness in responses.
- **Interference from additional context** When answering compound questions, additional contextual information can distract respondents, causing answers to deviate from the main topic or lack precision. Respondents need to selectively

process this information to maintain the accuracy and relevance of their responses.

- **Ambiguity in sub-question reference** In compound questions, earlier sub-questions are sometimes overlooked, leading to ambiguous references that make it challenging for respondents to accurately grasp the specific meaning of each sub-question.

These features collectively amplify the complexity associated with comprehending and responding to compound questions. We group the compound questions into the five categories: Factual-Statement, Cause-and-Effect, Hypothetical-Analysis, Comparison-and-Selection, and Evaluation-and-Suggestion. Table 1 presents the specific examples of each type.

- **Factual-Statement** This compound question inquires about multiple factual information without further reasoning or deep analyses. The sub-questions have minimal internal correlation.
- **Cause-and-Effect** This compound question inquires the respondent to analyze the causes of a particular phenomenon or event. Subsequently, the respondent is asked to explain the results and impacts. It encourages in-depth exploration of the phenomenon by establishing a logical connection between cause and effect, revealing internal relationships.
- **Hypothetical-Analysis** This compound question presents a hypothetical scenario to examine potential outcomes by analyzing the impact of various conditions or roles. Through this process, it aims to uncover complex relationships between elements and propose corresponding strategies or explanations.
- **Comparison-and-Selection** This compound question involves comparative analysis to uncover similarities and differences among multiple objects, phenomena, or situations. A comprehensive evaluation of each item is required to select the solution that best meets specific criteria.
- **Evaluation-and-Suggestion** This compound question necessitates an in-depth analysis of the current state, mechanisms, and issues, exploring their underlying reasons and weighing the pros and cons. Based on this analysis, the respondent seeks specific improvement suggestions to optimize further development.

Questions	Example
Factual-Statement	What is your favorite sport? Do you have any special skills or habits when playing sport?
Cause-and-Effect	Why has online learning been able to spread rapidly in recent years? Based on these main reasons, what are the main impacts of the popularization of online learning on society and individuals?
Hypothetical -Analysis	If I am successful in organizing this event, how do I ensure that I maximize its impact? If I don't, how do I deal with the potential negative consequences?
Comparison-and -Selection	Compared with traditional classroom teaching, what are the advantages of online learning modes in terms of enhancing learning efficiency and flexibility? What is the ideal way to ensure the quality of teaching and student-teacher interaction?
Evaluation-and -Suggestion	How do you evaluate the current operational status of online education platforms? What are their advantages and disadvantages in improving learning efficiency and meeting individual needs? Based on this, what measures do you think these platforms should take in the future to promote continuous improvement and innovation?

Table 1: Examples of different types of compound questions. The core ideas of each type are highlighted in the passage.

3.2 Data Collection

The Compound-QA benchmark is designed to evaluate the LLM capability in answering compound questions. To excel in this task, LLMs are expected to possess the ability to comprehend compound questions, disassemble and sequentially and exhaustively address each sub-question. The benchmark comprises three subsets dedicated to understanding, reasoning, and knowledge, respectively. Each subset is constructed using existing related datasets and includes compound questions of each type as detailed in Table 1. Detailed sources of the dataset are provided in the Appendix A. The compound question of each type is created by LLM-based. The framework for data synthesis CQ-Syn consists of the following three phases as shown in Figure 2.

- **Step 1: Question Design** We tailor unique prompts to address the characteristics of each type of compound question. The prompt covers task description, role description, and detailed data generation guidelines, along with manually curated examples to guide the process. We also include the corresponding context and the original question in the prompt, encouraging the LLMs to generate compound questions within the simi-

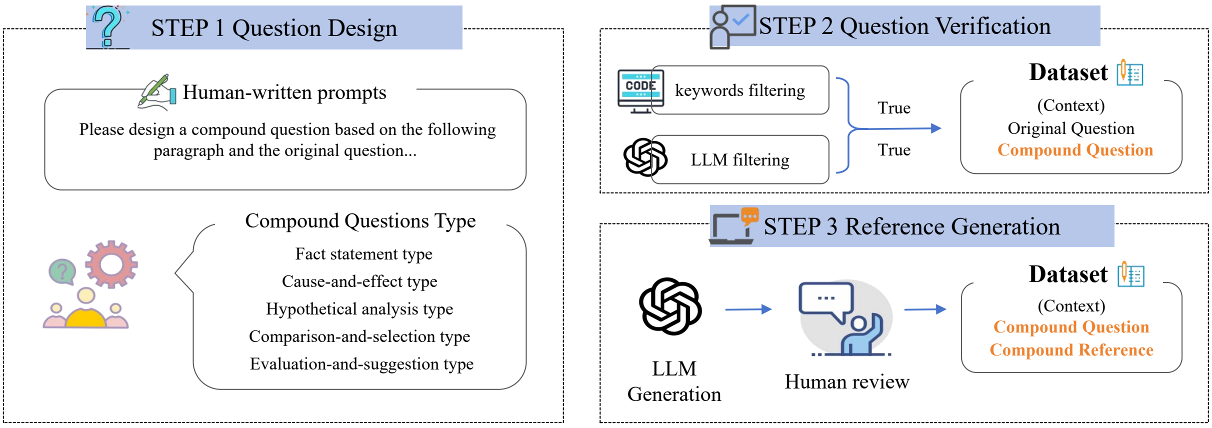

Figure 2: The overview of CQ-Syn Data Synthesis.

lar distribution of the original question based on the context. The Appendix B provides detailed descriptions of the prompts used for compound question generation.

- **Step 2: Question Verification** The generated compound questions are verified through both keyword-based and LLM-based filtering approaches. We first apply keyword-based rules to filter out the generated questions that do not contain the pre-specified keywords. The rules are manually crafted for each type separately.
- **Step 3: Reference Generation** For each filtered compound question, we prompt the proprietary LLM to obtain the reference answers for each compound question. Each compound question and its corresponding reference answer are manually reviewed to ensure accuracy and quality. For the manual validation session, we invited three students with master's degree level to participate in the validation of the final data. For more information about the validation session, please refer to the Appendix C.3.

3.3 Data Statistics

Our Compound-QA dataset consists of three main subsets: understanding, reasoning, and knowledge. Each subset includes five types of compound questions: Factual-Statement, Cause-and-Effect, Hypothetical-Analysis, Comparison-and-Selection, and Evaluation-and-Suggestion. For each question type, we generate 100 data points, totaling 1,500 data points across all types. We use `gpt-4o` to generate the QA pairs and employ `gpt-3.5-turbo` for validation.

4 Experiment Setup

In this section, we systematically evaluate the performance of open-source LLMs around 7B-9B on the Compound-QA benchmark. The evaluation focuses on five key aspects: 1) a comparative analysis of different LLMs' performance on compound questions; 2) the ability of LLMs to answer compound versus non-compound questions; 3) the ability of LLMs to answer sub-questions at different positions within compound questions; 4) the enhancement strategies for improving LLMs' capability to answer compound questions; 5) error analysize.

Settings. We conduct all inference experiments with the vLLM framework, utilizing two NVIDIA GeForce RTX 4090 GPUs. The parameters are configured with a temperature of 0.3 and a maximum token limit of 1024.

Models We evaluate seven open-sourced LLMs on the Compound-QA benchmark: DeepSeek (DeepSeek-AI, 2024), Mistral (Jiang et al., 2023), LLaMA-3.1 (Dubey et al., 2024), Gemma (Team et al., 2024), GLM-4 (GLM et al., 2024), Qwen (Team, 2024), InternLM (Cai et al., 2024). See the Appendix D for details.

Evaluation settings To quantify the model's ability in handling compound questions, we propose a multi-dimensional evaluation framework based on Comprehensiveness (explicit and complete addressing of all sub-questions), Correctness (factual and logical accuracy of each response component), and Diversity (variety in solution strategies across sub-questions). See the Appendix D.2 for details For each dimension, we use `gpt-4o-mini` as the evaluator to compare the model's response with the reference answer. The win rates are calculated as the percentage of instances where the model's response is judged to be equal or superior to the reference answer. To mitigate potential position bias, we compute the scores with the order of the model

Models	Size	Capability	Overall	FS	CE	HA	CS	ES
DeepSeek	7B	Understanding	13.1	25.8	12.8	5.5	17.8	3.5
		Reasoning	8.4	8.3	7.2	5.1	14.0	7.2
		Knowledge	12.0	25.3	9.8	5.5	14.3	5.3
Mistral	7B	Understanding	22.4	40.0	18.0	15.2	29.8	8.8
		Reasoning	13.6	10.8	11.7	11.6	22.5	11.5
		Knowledge	14.6	29.3	11.2	9.3	16.0	7.2
LLaMA	8B	Understanding	27.2	31.8	22.5	39.3	22.8	19.6
		Reasoning	22.7	15.5	22.0	21.7	32.5	21.8
		Knowledge	25.8	32.5	26.2	23.5	18.8	27.8
Gemma	9B	Understanding	32.7	29.0	26.3	43.2	39.2	25.7
		Reasoning	26.5	18.3	24.7	16.0	45.7	27.7
		Knowledge	25.0	22.7	25.3	32.3	19.2	25.5
GLM-4	9B	Understanding	56.0	51.3	**64.2**	**62.7**	**63.8**	37.8
		Reasoning	41.7	26.7	47.8	29.6	**60.3**	44.2
		Knowledge	46.7	44.8	54.5	47.2	48.0	38.8
Qwen	7B	Understanding	53.5	**56.7**	53.4	56.7	59.5	41.2
		Reasoning	**48.8**	**43.8**	52.5	**41.5**	56.2	50.2
		Knowledge	**53.1**	**55.2**	55.8	53.8	**50.3**	**50.2**
InternLM	7B	Understanding	**56.3**	54.8	62.3	60.3	61.2	**43.0**
		Reasoning	48.2	36.6	**53.2**	40.0	59.8	**51.2**
		Knowledge	52.9	50.0	**62.3**	**54.7**	48.2	49.3

Table 2: The win rates of different LLMs on Compound-QA benchmark, which covers five categories: Factual Statement (FS), Cause-and-Effect (CE), Hypothetical Analysis (HA), Comparison-and-Selection (CS), Evaluation-and-Suggestion (ES). Bold numbers indicate the highest score for that type.

response and reference answer swapped, then take the average as the final score. The overall performance is reported as the average score across all three dimensions. Additionally, we validate the reliability of the LLM evaluations by comparing their consistency with human assessments, achieving an agreement accuracy rate of 84%. See the Appendix E for details.

5 Experiment Results and Analyses

How do different LLMs perform when answering compound questions? Table 2 compares the win rates of seven open-source LLMs on five types of compound questions across the understanding, reasoning, and knowledge dimensions. InternLM and Qwen demonstrate the strongest overall performance. InternLM achieves the highest win rates in understanding tasks, while Qwen excels in reasoning and knowledge. GLM-4 also performs well, particularly in understanding, where it is comparable to InternLM. However, despite these strengths, all models exhibit weaknesses in certain types of compound questions, highlighting limitations in their reasoning and synthesis abilities.

For the five types of compound questions, models generally perform best on Factual-Statement questions. In contrast, more complex question types, such as Evaluation-and-Suggestion, pose significant challenges for all models. The win rates in this category are consistently lower, with InternLM scoring 43.0 and Qwen at 41.2. This can be attributed to the inherent difficulty of this question type, as models need to understand and integrate previous information to provide reasonable and coherent evaluations and recommendations on different topics or contexts. This requires not only strong comprehension and reasoning abilities but also the capability to generate logical and persuasive recommendations, which is a relatively complex and advanced task in NLP.

How do LLMs perform in answering compound versus non-compound questions? To further understand model behavior in answering compound questions, we compare the performance of LLaMA-3.1 and InternLM on both compound and non-compound questions. For non-compound questions, we use a multi-turn dialogue format, asking one question at a time.

First, we decompose each compound question in our Compound-QA dataset into non-compound questions. Because some questions are challenging to decompose effectively, we perform a manual

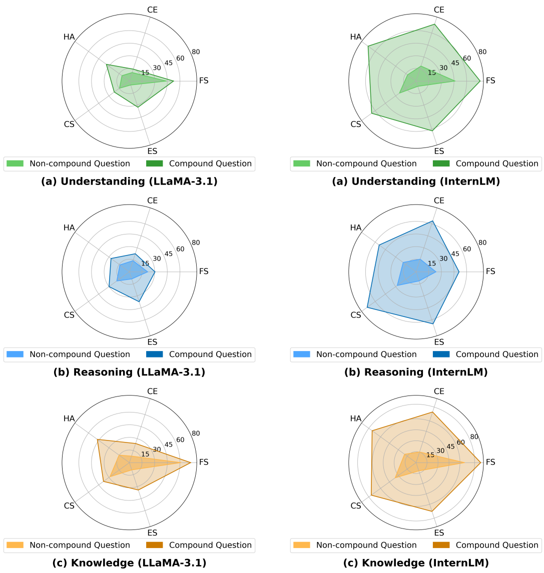

Figure 3: Performance comparison of LLaMA-3.1 and InternLM when answering compound and non-compound questions.

secondary check. Next, we use `gpt-4o` to generate reference answers for the decomposed questions. Finally, to ensure a fair comparison, we evaluate the models on this decomposed dataset by extracting answers to sub-questions from the responses to compound questions. We then report the win rates against the reference answers. Details on decomposition and extraction appear in Appendix F.

Figure 3 presents the win rates of LLaMA-3.1 and InternLM on compound and non-compound datasets. It is evident from the figure that the model's performance significantly declines when answering compound questions. We find that the answers to compound questions are much shorter than those for non-compound questions, which could contribute to the decline in answer quality. Among the five types of compound questions, Factual-Statement questions are the least affected by the compound nature, especially in terms of understanding and knowledge dimensions, where they perform more prominently. This may be due to the relatively simple nature of Factual-Statement questions and their weaker internal dependencies, resembling a combination of several independent sub-questions. However, in the reasoning dimension, this pattern is less pronounced. Factual-Statement questions here often involve combinatorial logic reasoning, demanding a deep analysis of context to arrive at the correct response. This added complexity results in lower performance.

How does LLM perform when answering sub-questions from different positions? To evaluate LLM's performance in answering sub-questions from different positions, We reorder Factual Statement questions within the understanding dimension and conduct experiments using LLaMA-3.1 and InternLM. We choose the Factual Statement questions because their sub-questions are relatively independent, and reordering them does not affect answers. Specifically, each data point contains three sub-questions, reordered so that each appears at the beginning, middle, and end of the sequence, labeled as 1XX, X1X, and XX1 (see Appendix G). We prompt the LLMs to answer these reordered compound questions.

Table 3 presents the win rates for answering sub-questions in different positions. The results show that sub-questions perform best when placed in the first and last positions, with the first position generally yielding the highest performance. Performance tends to dip in the middle position, suggesting that the model's ability to process information is stronger at the beginning and end of a sequence, but weaker in the middle. This is consistent with previous research findings that model performance steadily decreases as the number of sequential steps increases (Chen et al., 2024).

Model	Sub-question	1XX	X1X	XX1
LLaMA-3.1	Sub-question1	**40.7**	38.8	37.7
	Sub-question2	38.7	40.8	**41.8**
	Sub-question3	**44.2**	40.8	41.7
InternLM	Sub-question1	45.0	42.5	**47.5**
	Sub-question2	**51.3**	44.5	49.8
	Sub-question3	**54.7**	48.5	53.8

Table 3: Comparison of the performance of LLaMA-3.1 and InternLM in answering sub-questions at different positions in the compound questions. The positions 1XX, X1X, and XX1 correspond to the beginning, middle, and end of the compound questions.

How to improve LLM performance on Compound-QA benchmarks? In this experiment, we investigate different approaches to enhance the model's ability to answer compound questions. We test the LLaMA-3.1 model using four methods: Chain-of-Thought (CoT) (Wei et al., 2022), Decomposition strategy (Decom-S), Few-shot, and LoRA fine-tuning (Hu et al., 2021). CoT promotes step-by-step reasoning, whereas Decom-S explicitly instructs the model to break down a compound question into multiple sub-questions, address each sequentially, and synthesize the individual responses into a complete answer (see Appendix H for details). Fine-tuning is implemented via LLaMA-Factory[2] on 8 GPUs, with a LoRA rank of 8, alpha of 16, learning rate of 0.0001, and batch size of 128.

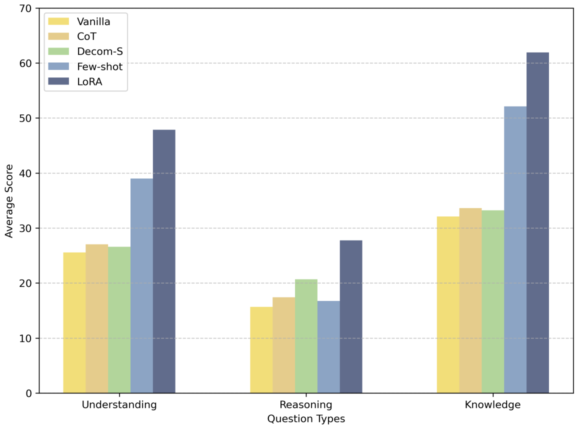

Figure 4: Comparative Performance of Different Improvement Methods on Compound-QA Benchmark.

To ensure a balanced evaluation across question types, the dataset is divided for each category and dimension (understanding, reasoning, and knowledge) in a 7:3 ratio. The 70% portion of each category and dimension is combined to form the training set, while the remaining 30% is used as the test set. CoT and Decom-S are evaluated in a zero-shot setting using the same test set.

Figure 4 compares the model's original and improved performance. LoRA fine-tuning consistently achieves the highest scores in all categories, demonstrating its effectiveness for compound question answering. Notably, Decom-S outperforms CoT in reasoning tasks, suggesting that explicit decomposition provides better handling of complex reasoning. Detailed results are presented in Figure 5 in the Appendix I.

Additionally, we further evaluate the fine-tuned model on MMLU (Hendrycks et al., 2020), GSM8K (Cobbe et al., 2021), and TruthfulQA (Lin et al., 2022) benchmarks to assess its general capabilities. As shown in Table 4, the fine-tuned model maintains strong performance across these benchmarks, with no significant degradation in its ability to handle diverse tasks. This demonstrates that fine-tuning on Compound-QA preserves the model's generalization ability while enhancing its performance on question answering.

Error Analysis Through observing the responses of open-source models, we identify common errors they tend to make when answering compound

Models	MMLU	GSM8K	TruthfulQA
LLaMA-3.1	67.78	83.00	47.61
LLaMA-3.1-LoRA	67.73	**84.50**	**52.39**

Table 4: Performance of the model on generic tasks

questions. These errors include omission of sub-questions (Error 1), confusion between related sub-questions (Error 2), and off-topic responses (Error 3). Detailed error examples appear in Table 25 (refer to Appendix J).

In Example 1, the compound question required answers to three sub-questions. However, the model addressed only the first two, omitting the third, which demonstrates the issue of sub-question omission. Example 2 shows a case where the second and third sub-questions, although closely related, were conflated in the model's response, resulting in failure to directly address the third sub-question. This pattern aligns with previous research (Chen et al., 2024). Example 3 illustrates the model's difficulty with multi-step logical reasoning: instead of following the necessary deductive steps to produce a precise permutation from logical constraints, it provided a general, explanatory narrative that lacked a definitive final answer. These observations highlight significant deficiencies in handling complex, multi-step reasoning tasks and offer valuable insights for future model refinements.

6 Conclusion

Compound questions present multiple sub-questions in a single query, which imposes challenges for LLMs to provide correct and appropriate responses to each sub-question in the Human-LLM interactive scenario. We introduce Compound-QA, a benchmark designed to evaluate the ability of LLMs on compound questions. This benchmark categorizes compound questions into five different types, with each type covering the scenarios of understanding, reasoning, and knowledge. The dataset is created using a Human-LLM collaborative framework, which includes a data synthesis process for generating and verifying compound questions. Our experiment reveals that LLMs require further improvement in effectively handling compound questions. We hope our benchmark will contribute to enhancing this capability. Additionally, we leave the exploration of evaluating compound questions in multimodal applications as future work.

[2] https://github.com/hiyouga/LLaMA-Factory

Limitations

The Compound-QA benchmark aims to evaluate the capabilities of LLMs in handling compound questions. The data creation process relies on proprietary LLMs, which introduces certain limitations. Due to resource and cost constraints, we use relatively smaller models for experimentation. Specifically, we employ `gpt-4o` to generate QA pairs and `gpt-4o-mini` for evaluation, excluding testing on closed-source models. Additionally, our dataset builds upon existing datasets, limiting its coverage across different domains and tasks and restricting it to English. In the future, we plan to extend our approach to other languages and modalities.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Taha Aksu, Devamanyu Hazarika, Shikib Mehri, Seokhwan Kim, Dilek Hakkani-Tür, Yang Liu, and Mahdi Namazifar. 2023. Cesar: Automatic induction of compositional instructions for multi-turn dialogs. *arXiv preprint arXiv:2311.17376*.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of ACL 2024*.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the ai: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Xinyi Chen, Baohao Liao, Jirui Qi, Panagiotis Eustratiadis, Christof Monz, Arianna Bisazza, and Maarten de Rijke. 2024. The sifo benchmark: Investigating the sequential instruction following ability of large language models. *arXiv preprint arXiv:2406.19999*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium.

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. *arXiv preprint arXiv:2110.14168*.

DeepSeek-AI. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Elvis Dohmatob, Yunzhen Feng, and Julia Kempe. 2024. Strong model collapse. *arXiv preprint arXiv:2410.04840*.

Qianlong Du, Chengqing Zong, and Jiajun Zhang. 2023. Mods: Model-oriented data selection for instruction tuning. *arXiv preprint arXiv:2311.15653*.

Haodong Duan, Jueqi Wei, Chonghua Wang, Hongwei Liu, Yixiao Fang, Songyang Zhang, Dahua Lin, and Kai Chen. 2023. Botchat: Evaluating llms' capabilities of having multi-turn dialogues. *arXiv preprint arXiv:2310.13650*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ahmed El-Kishky, Alexander Wei, Andre Saraiva, Borys Minaev, Daniel Selsam, David Dohan, Francis Song, Hunter Lightman, Ignasi Clavera, Jakub Pachocki, et al. 2025. Competitive programming with large reasoning models. *arXiv preprint arXiv:2502.06807*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics (TACL)*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Chaoqun He, Renjie Luo, Shengding Hu, Yuanqian Zhao, Jie Zhou, Hanghao Wu, Jiajie Zhang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2024a. Ultraeval: A lightweight platform for flexible and comprehensive evaluation for llms. *Preprint*, arXiv:2404.07584.

Qianyu He, Jie Zeng, Qianxi He, Jiaqing Liang, and Yanghua Xiao. 2024b. From complex to simple: Enhancing multi-constraint complex instruction following ability of large language models. *arXiv preprint arXiv:2404.15846*.

Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Jiaqing Liang, and Yanghua Xiao. 2024c. Can large language models understand real-world complex instructions? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18188–18196.

Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Jiaqing Liang, and Yanghua Xiao. 2024d. Can large language models understand real-world complex instructions? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18188–18196.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hanxu Hu, Pinzhen Chen, and Edoardo M Ponti. 2024. Fine-tuning large language models with sequential instructions. *arXiv preprint arXiv:2403.07794*.

Yiming Huang, Xiao Liu, Yeyun Gong, Zhibin Gou, Yelong Shen, Nan Duan, and Weizhu Chen. 2024. Key-point-driven data synthesis with its enhancement on mathematical reasoning. *arXiv preprint arXiv:2403.02333*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. *arXiv preprint arXiv:2401.16745*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. 2024a. Best practices and lessons learned on synthetic data for language models. *arXiv preprint arXiv:2404.07503*.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2023a. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. *arXiv preprint arXiv:2312.15685*.

Xiang Liu, Peijie Dong, Xuming Hu, and Xiaowen Chu. 2024b. Longgenbench: Long-context generation benchmark. *arXiv preprint arXiv:2410.04199*.

Yi Liu, Lianzhe Huang, Shicheng Li, Sishuo Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023b. Recall: A benchmark for llms robustness against external counterfactual knowledge. *arXiv preprint arXiv:2311.08147*.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126*.

Alisia Lupidi, Carlos Gemmell, Nicola Cancedda, Jane Dwivedi-Yu, Jason Weston, Jakob Foerster, Roberta Raileanu, and Maria Lomeli. 2024. Source2synth: Synthetic data generation and curation grounded in real data sources. *arXiv preprint arXiv:2409.08239*.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2023. Expertqa: Expert-curated questions and attributed answers. *arXiv preprint arXiv:2309.07852*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

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

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.

Zhen Tao, Zhiyu Li, Dinghao Xi, and Wei Xu. 2024. Cudrt: Benchmarking the detection of human vs. large language models generated texts. *arXiv preprint arXiv:2406.09056*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Jeroen van Craenenbroeck and Tanja Temmerman. 2018. Ellipsis In Natural Language: Theoretical and empirical perspectives. In *The Oxford Handbook of Ellipsis*. Oxford University Press.

Zhengxiang Wang, Jordan Kodner, and Owen Rambow. 2024. Evaluating llms with multiple problems at once: A new paradigm for probing llm capabilities. *arXiv preprint arXiv:2406.10786*.

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

Bosi Wen, Pei Ke, Xiaotao Gu, Lindong Wu, Hao Huang, Jinfeng Zhou, Wenchuang Li, Binxin Hu, Wendy Gao, Jiaxing Xu, et al. 2024a. Benchmarking complex instruction-following with multiple constraints composition. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Bosi Wen, Pei Ke, Xiaotao Gu, Lindong Wu, Hao Huang, Jinfeng Zhou, Wenchuang Li, Binxin Hu, Wendy Gao, Jiaxing Xu, et al. 2024b. Lingoly: A benchmark of olympiad-level linguistic reasoning puzzles in low-resource and extinct languages. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. 2024. How easily do irrelevant inputs skew the responses of large language models? *arXiv preprint arXiv:2404.03302*.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *arXiv preprint arXiv:2406.08464*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

Fan Zhou, Zengzhi Wang, Qian Liu, Junlong Li, and Pengfei Liu. 2024. Programming every example: Lifting pre-training data quality like experts at scale. *arXiv preprint arXiv:2409.17115*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. 2024. Dyval: Dynamic evaluation of large language models for reasoning tasks. *arXiv preprint arXiv:2309.17167*.

A Data sources

To assess the LLM's capacity to address compound questions, we evaluate its performance across three dimensions: understanding, reasoning, and knowledge.

Understanding We modify data from Adversarial QA (Bartolo et al., 2020), a challenging QA dataset designed to test the robustness of models through adversarial examples.

Reasoning We adapt data from AGI-Eval (Zhong et al., 2023) and collect a small number of permutation-type logical reasoning questions from web pages. AGI-Eval, released by Microsoft, serves as a benchmark for evaluating the foundational capabilities of LLMs, focusing on human cognition and general problem-solving skills. We select two English datasets from AGI-Eval: LSAT (Law School Admission Test) and logical reasoning questions from civil service exams.

Knowledge We adapt data from PubMed (Jin et al., 2019), a biomedical QA dataset that includes numerous QA pairs based on PubMed articles. This dataset evaluates models' biomedical knowledge and their ability to retrieve and understand information in a specialized field.

During the automated data collection process, we manually modify the data format and expression, either independently or with the assistance of LLMs. We conduct manual verification to ensure the quality of the dataset. Table 5 presents the licensing information for each dataset used in this study.

Dataset Name	License Type
Adversarial QA	CC BY-SA 4.0
AGI-Eval	MIT
PubMedQA	MIT
Our(Compound_QA)	CC BY-SA 4.0

Table 5: Types of licenses for datasets

The datasets employed in this study comply with their respective licenses. For datasets under the CC BY-SA 4.0 license (such as Adversarial QA), we adhere to the attribution-share alike requirements, ensuring that any modifications and redistributions of these datasets are released under the same CC BY-SA 4.0 license. For datasets under the MIT license (such as PubMedQA and AGI-Eval), we comply with the license terms and ensure that these datasets are processed and extended in accordance with the license requirements.

Additionally, our newly generated dataset is released under the CC BY-SA 4.0 license, ensuring flexibility and ease of use, which allows other researchers to modify and redistribute it freely under the same conditions.

B Details of Generating Complex Questions

Table 6 to Table 10 present the detailed prompts used for generating datasets for various types of compound questions. Each prompt includes the target role setting, specific data generation rules, and manually compiled examples to guide the LLM, ensuring that the generated data meets the requirements for different types of compound questions. To maintain the reasoning nature of the generated questions within the reasoning dimension, we design a specialized prompt, as shown in Table 11, to ensure that the proposed questions exhibit reasoning characteristics.

C Filtering methods for compound questions

C.1 Details of Keywords Filtering

Keyword-based rules are employed to identify various types of compound questions. Each compound question filter relies on one or two dimensions and a specific list of keywords.

Factual-Statement The filter words include question words {*"what", "how", "why", "when", "which", "how much", "where", "who"*}

Cause-and-Effect The filter words include two dimensions: cause words {*"why", "what causes", "what are the reasons", "how does"*} and effect words {*"what", "how", "what are the effects", "what are the consequences", "how does it affect", "what results", "What impacts"*}

/ Task prompt */*
You are an expert at designing questions. Please create a compound question based on the following paragraph and the original issue. This compound question should be a coherent sentence that builds upon the original question by posing 2 to 3 additional inquiries seeking basic factual information, such as age, preferences, or the execution of specific actions. These questions should not involve complex logical reasoning or deep analysis but should focus on obtaining specific facts that the respondent can directly answer or confirm.
You can refer to these compound question examples:

/ Example */*
example1: What is your age? What are your favorite leisure activities? How do you usually spend your weekends?
example2: What is your current occupation? How long have you been working in this industry? What motivated you to choose this career path?
example3: What kind of travel mode do you usually prefer? When was your last trip? What destination did you visit during your last trip?

/ Generation */*
Based on the following actual input
context: {**context**}
org_question: {**org_question**}
please design a concise and clear compound question, containing 2 to 3 small questions, with no more than 80 characters. Avoid excessive elaboration on the text content and focus on the compound question itself:

Table 6: Prompt for generating Factual-Statement questions.

Hypothetical-Analysis The filter words include two dimensions: hypothetical words {*"if", "suppose", "imagine", "what if", "in case", "assuming", "provided that", "on the condition that"*} and question words {*"what", "how", "why", "when", "which", "how much", "where", "who"*}

Comparison-and-Selection The filter words include two dimensions: comparison words {*"compare", "compared", "comparing", "comparison", "contrast", "advantage", "disadvantage", "better", "worse", "differ", "difference", "similarities", "distinctions"*} and question words {*"what", "how", "why", "when", "which", "how much", "where", "who"*}

Evaluation-and-Suggestion The filter words include question words {*"what", "how", "why", "when", "which", "how much", "where", "who"*}

C.2 Details of LLM filtering

We design different criteria prompts to ensure the accuracy and relevance of each question. Table 12 to Table 16 show the filtering prompts for different types of questions. We use `gpt-3.5-turbo` for the filtering process.

C.3 Manual Verification Sessions

Three graduate students participate in the final data validation process, which involves evaluating both the quality of the compound questions and their corresponding reference answers. Before the validation begins, the reviewers receive specialized training on the classification criteria for compound questions and the guidelines for quality assessment. A three-tier scoring system (0/1/2 points) is employed, and only samples unanimously rated as 2 by all three reviewers are retained. The scoring criteria are as follows:

First level (0 points) These questions clearly do not meet the definition of the specified type of compound questions. For example, in Comparison-and-Selection questions, multiple comparable objects are not included.

Second level (1 point) Although the question superficially appears to meet the definition of the specified type of compound questions, it does not genuinely belong to that category. For instance, a Comparison-and-Selection questions might ask, "Compared to object A, what happens to object B under certain conditions?" However, the question could be phrased directly without involving a comparison.

Third level (2 points) These compound questions fully meet the definition, and the responses are detailed and complete, with no omissions or incomplete answers.

The overall human verification pass rate is approximately 60%, with more complex questions requiring multi-step reasoning having a pass rate as low as 50%. For each question type, only 100 verified samples are retained to ensure high-quality questions and answers.

/ Task prompt */* You are an expert at designing questions. Please create a compound question based on the following paragraph and the original issue. This compound question should be a coherent sentence that first inquiries about the reasons for a particular phenomenon or event, using phrases like "why" or "what are the reasons." Subsequently, based on this cause, it should further explore the primary impacts or consequences that may arise, employing expressions such as "what impacts will it bring" or "what results will it lead to." You can refer to these compound question examples: */* Example */* example1: Why has artificial intelligence technology made breakthrough progress in the field of medical diagnosis? How will the application of this AI-assisted diagnostic system change the future medical service model and bring about what impacts? example2: Why does the government strongly support the development of online education? What are the main socio-economic impacts that this policy support will bring? example3: Why has biotechnology been widely applied in agricultural production? How will the development of agricultural biotechnology change the future food supply pattern and bring new opportunities to farmers? */* Generation */* Based on the following actual input context:{**context**} org_question:{**org_question**} please design a concise and clear compound question, containing 2-3 small questions, with no more than 80 characters. Avoid excessive elaboration on the text content and focus on the compound question itself:

Table 7: Prompt for generating Cause-and-Effect questions.

D Details on Models and evaluation settings

D.1 Models

The details of the models used are as follows:

- **DeepSeek**: DeepSeek (DeepSeek-AI, 2024) is a mixture-of-experts language model developed by DeepSeek AI. We employ the `DeepSeek-llm-7b-chat` model.
- **Mistral**: Mistral (Jiang et al., 2023) is a series of open source models developed by the French company Mistral AI. We employ the `Mistral-7B-Instruct-v0.3` model.
- **LLaMA**: LLaMA-3.1 (Dubey et al., 2024) is a series of open source models proposed by Meta, which have been improved in terms of reasoning ability and multilingual support, with their context length increased to 128K. We employ the `Meta-LLaMA-3.1-8B-Instruct` model.
- **Gemma**: Gemma (Team et al., 2024) is a series of lightweight open-source models developed by Google. We utilize the `gemma-2-9b-it` model.
- **GLM-4**: GLM-4 (GLM et al., 2024) is a series of open source models introduced by Zhipu AI with powerful Agent capabilities, support for longer contexts, faster inference, and reduced inference costs. We employ the `glm-4-9b-chat` model.
- **Qwen**: Qwen (Team, 2024) is a series of open source models developed by Alibaba Cloud as part of the Tongyi Qianwen series. This series includes multiple versions and scales, such as Base and Chat models, to meet different computational needs. We employ the `Qwen2.5-7B-Instruct` model.
- **InternLM**: InternLM (Cai et al., 2024) is an open source, lightweight training framework designed to support large model training without extensive dependencies. We employ the `internlm2-5-7b-chat` model.

D.2 Evaluation prompt

The evaluation prompt details are given in the Table 17 to Table 19. It asks the judge to compare a given question, a reference answer, and a model-generated answer, categorizing the evaluation results into three labels: $A > B$, $A = B$, $B > A$.

E LLM-Based Evaluation and Human Evaluation Consistency Analysis

To further verify the accuracy of using LLMs as evaluators, we involve three graduate-level annotators with expertise in the relevant field. Each annotator independently assesses a set of questions paired with two responses (Response A and Response B), using two criteria: (1) clarity of format and (2) completeness of content.

The annotators assign one of three labels to compare the responses: A > B, A = B, and B > A. To ensure consistency, a label is considered valid only if at least two annotators agree on it. In cases of

/ Task prompt */*
You are an expert at designing questions. Please design a compound question based on the following paragraph and the original question. The compound question should be a coherent sentence that first establishes two hypothetical scenarios or conditions, introducing them with words like "if," and then inquiries about the different potential outcomes, impacts, or implications that may arise under these hypothetical premises. When conceiving the compound question, please focus on these two hypothetical scenarios. Explore the different consequences that may occur in the case of success and failure. Alternatively, analyze how different factors or roles influence potential outcomes and propose corresponding strategies or explanations.
/ Example */*
example1: If I successfully organize this public welfare event, how can I ensure the maximum impact of the event? If it fails, how should I address the possible negative consequences? example2: If you were the CEO of a company facing malicious price competition strategies from competitors, how would you respond? Conversely, if you were a newly-hired marketing officer in this company, what measures would you take to tackle this challenge? example3: If the government significantly increases financial support for the research, development, and promotion of new energy vehicles, how will car manufacturers, power companies, and consumers respond to this change? Conversely, if the government reduces relevant policy support, what challenges will this bring to all stakeholders?
/ Generation */*
Based on the following actual input context:{**context**} org_question:{**org_question**} please design a concise and clear compound question, containing 2-3 small questions, with no more than 80 characters. Avoid excessive elaboration on the text content and focus on the compound question itself:

Table 8: Prompt for generating Hypothetical-Analysis questions.

disagreement, the majority decision determines the final label.

After evaluating 100 data points, we compare the human assessments with those from the LLM-based evaluation. To quantify the performance of the LLM evaluator, we treat it as a multi-class classifier, where each label (A > B, A = B, B > A) represents a distinct class. We calculate evaluation metrics such as accuracy, True Positive Rate (TPR), and False Positive Rate (FPR) for each class. The results show an overall accuracy of 84%, with an average TPR of 0.8530 and an average FPR of 0.1421, indicating strong alignment between LLM and human evaluations.

F Prompt for decomposing compound question and extracting sub-question answers

Table 20 presents the prompt used for decomposing compound questions into non-compound questions. To facilitate a fair comparison between compound and non-compound questions, we extract sub-question answers from the responses to compound questions. Table 21 displays the detailed prompt for this answer extraction.

G Four Orderings of Compound Questions

This section details the construction of four distinct orderings of compound questions designed to assess sub-questions in various positions, as illustrated in Table 22. Each ordering ensures that every sub-question appears at the beginning, middle, and end of the compound question, providing a comprehensive evaluation framework.

H Prompt variants

We also present the details of prompts using CoT, Decom-S and few-shot prompting, as shown in Tables 23 to Table 24.

I Improvement Methods on Compound-QA

The Figure 5 presents a detailed comparison of five improvement methods (Vanilla, CoT, Decom-S, Few-shot, LoRA) across all question categories: Factual-Statement, Cause-and-Effect, Hypothetical-Analysis, Comparison-and-Selection, and Evaluation-and-Suggestion.

J Error Analysis

Table 25 shows the types of errors made by the model in answering the compound questions.

/ Task prompt */*
You are an expert at designing questions. Please design a compound question based on the following paragraph and the original question. This compound question should be a coherent sentence, which first poses 2-3 comparison questions that contrast different objects using words like "compared with" and "in contrast to" to analyze their similarities and differences. Then, inquire about the object or solution that best meets specific criteria or conditions.

/ Example */*
example1: Compared with AI-assisted diagnosis, how is the accuracy and reliability of independent diagnosis by doctors? In terms of improving medical efficiency and reducing medical costs, which diagnostic model will generate greater social value?
example2: Compared with conventional fuel vehicles, how do new energy vehicles differ in terms of range, charging/refueling convenience and environmental friendliness? Which model will become the mainstream choice in the next 10 years?
example3: Compared with traditional classroom teaching, what are the advantages of online learning modes in terms of enhancing learning efficiency and flexibility? What is the ideal way to ensure the quality of teaching and student-teacher interaction?

/ Generation */*
Based on the following actual input
context:{**context**}
org_question:{**org_question**}
please design a concise and clear compound question, containing 2-3 small questions, with no more than 80 characters. Avoid excessive elaboration on the text content and focus on the compound question itself:

Table 9: Prompt for generating Comparison-and-Selection questions.

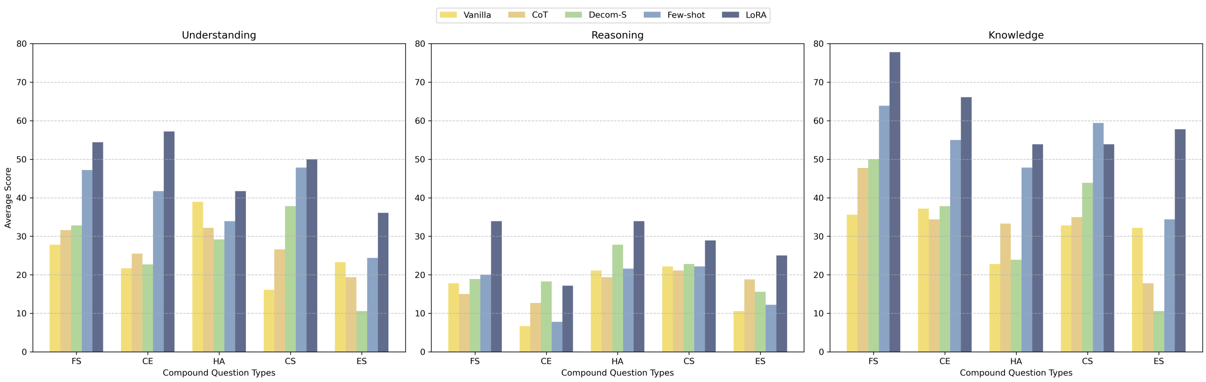

Figure 5: Comparative Performance of Different Improvement Methods on Compound-QA Benchmark

/ Task prompt */*
You are an expert at designing questions. Please create a compound question based on the following paragraph and the original issue. The compound question should be a coherent sentence that first asks the respondent to comprehensively evaluate the current status of the matter, including its operational mechanism, achieved results, and existing problems. Then, it should guide the respondent to deeply analyze the reasons behind these statuses, exploring their advantages and disadvantages, in order to more accurately grasp the essence of the matter. Finally, based on the previous analysis, the question will require the respondent to propose targeted improvement suggestions or future development measures aimed at promoting the optimization and development of the matter.
/ Example */*
example1: How do you evaluate the current operational status of online education platforms? What are their advantages and disadvantages in improving learning efficiency and meeting individual needs? Based on this, what measures do you think these platforms should take in the future to promote continuous improvement and innovation? example2: What is the current status of artificial intelligence application in the field of medical diagnosis? What potential risks and challenges might this AI technology bring? In response to these issues, what coping strategies do you suggest medical institutions and regulatory authorities should adopt? example3: What are the advantages and disadvantages of shopping in physical stores and online shopping respectively? In terms of protecting privacy and security, what aspects do e-commerce platforms still need to further improve and perfect? How do you think consumers' shopping habits will change in the future, and how should enterprises respond?
/ Generation */*
Based on the following actual input context:{**context**} org_question:{**org_question**} please design a concise and clear compound question, containing 2-3 small questions, with no more than 80 characters. Avoid excessive elaboration on the text content and focus on the compound question itself:

Table 10: Prompt for generating Evaluation-and-Suggestion questions.

/ Task prompt */*
You are an expert in designing inference-based questions. Please design a hypothetical analysis compound question based on the following context and the original question. The compound question should present two hypothetical scenarios or conditions using words like "if" or "assuming," and each scenario must be analyzed using logical inference to explore different outcomes or impacts. The analysis should follow an inference-driven process based on known facts and logical reasoning. When designing the compound question, focus on how each hypothetical scenario leads to different potential outcomes or implications through multi-step reasoning. Each sub-question must logically follow from the established conditions. You can refer to these inference-based compound question examples:
/ Example */*
example 1: If player F, G, and J are already participating, which additional players can be selected? How many combinations of players meet the selection criteria if F, G, and J are participating? Which of these combinations include player H? Reasoning steps: First, infer which players are already selected, analyze the selection criteria, and then logically deduce the valid combinations. Example 2: Assuming that the budget for a project is cut by 20%, which components will be affected? How many of the components can still be implemented under the reduced budget? Which of these components are essential for project success? Reasoning steps: First, analyze the budget and current costs, then deduce the components that fit within the budget, and infer the essential ones. example 3: If I successfully organize this public welfare event, how can I ensure the maximum impact? What are the logical steps to ensure this success based on current resources? If it fails, what are the most likely causes, and how can I mitigate the negative impact? Reasoning steps: First, evaluate the resources available, infer success factors, then logically analyze possible causes of failure.
/ Generation */*
Based on the following actual input context: {**context**} org_question: {**org_question**} Please design an inference-based hypothetical analysis compound question containing 2-3 sub-questions. Each sub-question must be derived through logical reasoning, and the question must be concise and clear. Avoid excessive elaboration on the text content and focus on the compound question itself.

Table 11: Prompt for generating hypothetical analysis compound questions under the reasoning dimension

/ Task prompt */*
To evaluate the quality of a given factual statement compound question, I will assess the question based on the following criteria:

/ Screening criteria */*
1. Compound Structure: The question should consist of two to three interrelated sub-questions, forming a compound question.
2. Factual Content: The question should ask for basic factual information such as age, preferences, or specific behaviors.
3. Direct Answerability: The question should be directly answerable or confirmable without requiring complex logical reasoning or in-depth analysis.

/ Generation */*
Question: {**com_question**}
Provide your judgement in the following format:
result: True or False
judgement: <Explain the reasons behind your judgement>

Table 12: Screening prompt for Factual-Statement questions.

/ Task prompt */*
To effectively filter cause-and-effect compound questions, I will evaluate questions based on the following criteria:

/ Screening criteria */*
1. Compound Structure: The question must consist of two to three interrelated sub-questions, with a clear cause-and-effect structure.
2. Cause Identification: The first part of the question should aim to identify and describe the reasons or causes behind a specific phenomenon or event, typically beginning with phrases like "Why" or "What are the reasons."
3.Outcome Exploration: Following the identification of causes, the question should logically inquire about the consequences or impacts that arise from these causes, using expressions such as "What impacts will it bring" or "What results will it lead to."

/ Generation */*
Question: {**com_question**}
Provide your judgement in the following format:
result: True or False
judgement: <Explain the reasons behind your judgement>

Table 13: Screening prompt for Cause-and-Effect questions.

/ Task prompt */*
To evaluate the quality of a given hypothetical analysis compound question, I will assess the question based on the following criteria:

/ Screening criteria */*
1. Compound Structure: The question should establish two to three interrelated hypothetical scenarios or conditions, typically introduced with phrases like "if", "suppose", or "assuming".
2. Analytical Inquiry: Following the establishment of hypothetical scenarios, the question should explore the potential outcomes, impacts, or implications that may arise under these conditions.
3. Complex Relationship Analysis: The question should delve into analyzing the complex relationships and dynamics between different factors, roles, or stakeholders involved in the hypothetical scenarios.

/ Generation */*
Question: {**com_question**}
Provide your judgement in the following format:
result: True or False
judgement: <Explain the reasons behind your judgement>

Table 14: Screening prompt for Hypothetical-Analytical questions.

/ Task prompt */*
To determine whether a given question belongs to a comparative selection compound question, it must fulfill the following criteria:

/ Screening criteria */*
1. Compound Structure: The question should consist of two or more interrelated sub-questions, forming a compound question.
2. Comparative Elements: The question should involve two or more objects, phenomena, or scenarios that are being compared.
3. Selection Requirement: The question should require the selection of the most suitable object or solution based on specific criteria or conditions.

/ Generation */*
Question: {**com_question**}
Provide your judgement in the following format:
result: True or False
judgement: <Explain the reasons behind your judgement>

Table 15: Screening prompt for Comparison-and-Selection questions.

/ Task prompt */*
To effectively assess and filter evaluation-suggestion type compound questions, I will apply the following criteria:

/ Screening criteria */*
1. Comprehensive Evaluation: The question must initiate with a request for a thorough assessment of the current situation of the subject, covering aspects like its operational mechanisms, effectiveness, and any issues it faces.
2. Deep Analysis: It should then guide the respondent to explore the underlying reasons for the current situation, identifying advantages and disadvantages to better understand the essence of the subject.
3. Constructive Suggestions: Finally, the question should culminate by asking the respondent to propose targeted improvements or future development actions based on the analysis, aiming to enhance or evolve the subject matter.

/ Generation */*
Question: {**com_question**}
Provide your judgement in the following format:
result: True or False
judgement: <Explain the reasons behind your judgement>

Table 16: Screening prompt for Evaluation-and-Suggestion questions.

/ Task prompt */*
Please act as an impartial judge to evaluate responses to a compound question. You will compare Assistant A and Assistant B based on the following structured evaluation framework. **Important: Do NOT let the length of an assistant's answer influence your evaluation. A longer response is not necessarily better. ** Your evaluation must be concise. Keep it within 150 characters unless a slightly longer response is necessary for clarity.

/ Evaluation Guidelines */*
- Identify all sub-questions embedded in the compound question.
- Compare whether Assistant A or B fully addresses more sub-questions.
- Assess whether the response is well-structured and logically organized.
- **Penalize unnecessary verbosity or repetition that does not add value.**
- Provide a final judgment based on completeness and organization.

/ Final Judgment Format */*
- [[A>B]] (Assistant A is more comprehensive)
- [[B>A]] (Assistant B is more comprehensive)
- [[A=B]] (Both are equally comprehensive)

/ Input */*
`<|User Prompt|>`
{sample.com_question}
`<|The Start of Assistant A's Answer|>`
{sample.com_answer if not swap_position else sample.com_reference}
`<|The End of Assistant A's Answer|>`
`<|The Start of Assistant B's Answer|>`
{sample.com_reference if not swap_position else sample.com_answer}
`<|The End of Assistant B's Answer|>`

Table 17: Comprehensiveness Evaluation of Model Responses to Compound Questions.

/ Task prompt */*
Please act as an impartial judge to evaluate responses to a compound question. You will compare Assistant A and Assistant B based on the following structured evaluation framework. **Important: Do NOT let the length of an assistant's answer influence your evaluation. A longer response is not necessarily better. ** Your evaluation must be concise. Keep it within 150 characters unless a slightly longer response is necessary for clarity.

/ Evaluation Guidelines */*
- Verify factual accuracy and logical consistency of each response.
- Identify any errors, unsupported claims, or misleading information.
- Compare whether Assistant A or B provides more precise and error-free reasoning.
- **Reward concise and accurate responses over lengthy but less precise ones.**
- Provide a final judgment based on factual correctness and reasoning clarity.

/ Final Judgment Format */*
- [[A>B]] (Assistant A is more correct)
- [[B>A]] (Assistant B is more correct)
- [[A=B]] (Both are equally correct)

/ Input */*
<|User Prompt|>
{sample.com_question}
<|The Start of Assistant A's Answer|>
{sample.com_answer if not swap_position else sample.com_reference}
<|The End of Assistant A's Answer|>
<|The Start of Assistant B's Answer|>
{sample.com_reference if not swap_position else sample.com_answer}
<|The End of Assistant B's Answer|>

Table 18: Correctness Evaluation of Model Responses to Compound Questions.

/ Task prompt */*
Please act as an impartial judge to evaluate responses to a compound question. You will compare Assistant A and Assistant B based on the following structured evaluation framework. **Important: Do NOT let the length of an assistant's answer influence your evaluation. A longer response is not necessarily better. ** Your evaluation must be concise. Keep it within 150 characters unless a slightly longer response is necessary for clarity.

/ Evaluation Guidelines */*
- Compare whether Assistant A or B presents a wider range of perspectives, solutions, or approaches.
- Evaluate the richness of expression (e.g., use of examples, analogies, or data).
- Check for creative or innovative elements in the response.
- **Do not equate length with diversity; focus on the meaningful variety of content.**
- Provide a final judgment based on the variety and depth of ideas presented.

/ Final Judgment Format */*
- [[A>B]] (Assistant A is more diverse)
- [[B>A]] (Assistant B is more diverse)
- [[A=B]] (Both are equally diverse)

/ Input */*
<|User Prompt|>
{sample.com_question}
<|The Start of Assistant A's Answer|>
{sample.com_answer if not swap_position else sample.com_reference}
<|The End of Assistant A's Answer|>
<|The Start of Assistant B's Answer|>
{sample.com_reference if not swap_position else sample.com_answer}
<|The End of Assistant B's Answer|>

Table 19: Diversity Evaluation of Model Responses to Compound Questions.

/* *Task prompt* */
You excel at text processing, especially in identifying and extracting questions. Now, please process the following text:

/* *Input* */
Text: com_question

/* *Decomposition criteria* */
Based on the provided information, please complete the following tasks:
Firstly, determine if the text contains at least one question. If it does, proceed to the next step; if not, simply output "This text does not contain any questions."
If the text contains multiple questions, attempt to split it into individual, meaningful questions. Each question should end with a question mark '?' and ensure that the split questions are logically complete and coherent.
If the question in the text cannot be effectively split, or if the split questions are not logically complete or coherent, then output the original text as a single question.

/* *Generation* */
Please ensure that your output is accurate, clear, and retains the key information from the original text as much as possible.

Table 20: Prompt for decomposing a compound question into non-compound questions.

/* *Task prompt* */
You are an assistant with expertise in text extraction.

/* *Extraction criteria* */
Given the following context, extract the answer to the specified question. Ensure that the answer is directly related to the question and avoid including information that answers other questions. If the answer is not found, output "None".

/* *Generation* */
Question: question
Com_Answer: com_answer
Answer:

Table 21: Prompt for extracting the answer to a sub-question in the answer to a compound question.

Order	**Example**
Order1	What is the significance of the Amazon Rainforest in carbon sequestration? How does the Amazon Rainforest contribute to global oxygen production? What are the current threats to the preservation of the Amazon Rainforest?
Order2	How does the Amazon Rainforest contribute to global oxygen production? What is the significance of the Amazon Rainforest in carbon sequestration?What are the current threats to the preservation of the Amazon Rainforest?
Order3	How does the Amazon Rainforest contribute to global oxygen production? What are the current threats to the preservation of the Amazon Rainforest?What is the significance of the Amazon Rainforest in carbon sequestration?
Order4	What are the current threats to the preservation of the Amazon Rainforest? What is the significance of the Amazon Rainforest in carbon sequestration?How does the Amazon Rainforest contribute to global oxygen production?

Table 22: Four Orderings of Compound Questions

/* *vanilla prompt* */
Please answer the following question: com_question,You can refer to the following content {context}

/* *CoT Prompt* */
Please answer the following questions respectively: com_question,**Let's think step by step.** You can refer to the following content {context}.

/* *Decom-S Prompt* */
Please decompose the following compound question into multiple sub-questions, answer each sub-question step by step, and finally synthesize the answers to form a complete response without repeating information.
Compound Question:{com_question}
Context for reference: {context}

Table 23: CoT, Decom-S and vanilla prompt details.

/ Context */*
The quick and decisive defeat of the Arab troops during the Six-Day War by Israeli troops constituted a pivotal event in the Arab Muslim world. The defeat along with economic stagnation in the defeated countries, was blamed on the secular Arab nationalism of the ruling regimes. A steep and steady decline in the popularity and credibility of secular, socialist and nationalist politics ensued. Ba'athism, Arab socialism, and Arab nationalism suffered, and different democratic and anti-democratic Islamist movements inspired by Maududi and Sayyid Qutb gained ground.

/ com_question */*
What led to the decline of secular ideologies post-Six-Day War? What contributed to the adoption of Islamist movements in addition to the 6-Day War? Are Islamist movements dominating?

/ com_reference */*
1. **What led to the decline of secular ideologies post-Six-Day War?**
- The quick and decisive defeat of the Arab troops during the Six-Day War by Israeli troops was a pivotal event that severely undermined the credibility of the secular Arab nationalism espoused by the ruling regimes. This defeat, coupled with economic stagnation in the defeated countries, was blamed on these secular ideologies, leading to a steep and steady decline in their popularity and credibility. Specifically, ideologies such as Ba'athism, Arab socialism, and Arab nationalism suffered as a result of this combination of military and economic failures.
2. **What contributed to the adoption of Islamist movements in addition to the Six-Day War?**
- In addition to the impact of the Six-Day War, the economic stagnation experienced by the defeated countries played a significant role in discrediting secular regimes. The failure of secular, socialist, and nationalist politics to deliver economic prosperity and stability created a vacuum that was filled by different democratic and anti-democratic Islamist movements. These movements, inspired by figures like Maududi and Sayyid Qutb, offered an alternative ideological framework that promised to address the failures of the secular regimes.
3. **Are Islamist movements dominating?**
- While the context provided indicates that Islamist movements gained ground following the decline of secular ideologies, it does not explicitly state that they are dominating. However, the phrase "different democratic and anti-democratic Islamist movements...gained ground" suggests that these movements became increasingly influential in the political landscape, challenging the previously dominant secular ideologies. Therefore, it can be inferred that Islamist movements became more prominent, but the extent of their dominance would require further context-specific information.

Table 24: Few-shot prompt for the Adversarial QA dataset.

Error	Details
Error 1	**Context:** From the death of Augustus in AD 14 until after AD 70, Rome accepted as her Germanic frontier the water-boundary of the Rhine and upper Danube. Beyond these rivers she held only the fertile plain of Frankfurt, opposite the Roman border fortress of Moguntiacum (Mainz), the southernmost slopes of the Black Forest and a few scattered bridge-heads. The northern section of this frontier, where the Rhine is deep and broad, remained the Roman boundary until the empire fell. The southern part was different. The upper Rhine and upper Danube are easily crossed. The frontier which they form is inconveniently long, enclosing an acute-angled wedge of foreign territory between the modern Baden and Württemberg. The Germanic populations of these lands seem in Roman times to have been scanty, and Roman subjects from the modern Alsace-Lorraine had drifted across the river eastwards. **Compound Questions:** What is the first place mentioned? Where is it located? What rivers define Rome's Germanic frontier? **Compound Answer:** The first place mentioned in the provided content is "Frankfurt," which is located in the fertile plain opposite the Roman border fortress of Moguntiacum (Mainz).
Error 2	**Context:** The following four timelines show the geologic time scale. The first shows the entire time from the formation of the Earth to the present, but this compresses the most recent eon. Therefore, the second scale shows the most recent eon with an expanded scale. The second scale compresses the most recent era, so the most recent era is expanded in the third scale. Since the Quaternary is a very short period with short epochs, it is further expanded in the fourth scale. The second, third, and fourth timelines are therefore each subsections of their preceding timeline as indicated by asterisks. The Holocene (the latest epoch) is too small to be shown clearly on the third timeline on the right, another reason for expanding the fourth scale. The Pleistocene (P) epoch. Q stands for the Quaternary period. **Compound Questions:** What do the second, third, and fourth timelines represent? How are they interconnected? Are they expansions of the preceding timelines? **Compound Answer:** The second, third, and fourth timelines represent the geologic time scale, which is an expansion and compression of the time line of the corresponding period. The second timeline represents the Pleistocene epoch, the third timeline represents the Quaternary period, and the fourth timeline represents the Holocene epoch. The fourth timeline is an expansion of the third timeline, and the third timeline is an expansion of the second timeline. The second timeline is an expansion of the first timeline. The first timeline represents the entire time from the formation of the Earth to the present, but this compresses the most recent eon. The fourth timeline is too small to be shown clearly on the third timeline, so it is further expanded. The four timelines are interconnected, and they are all part of the geologic time scale.
Error 3	**Context:** Five graduate students Zhang, Wang, Li, Zhao, and Liu graduated from a linguistics major in a college of arts of a university in 2014, and were hired by one of the three employers, Tianshu, Tianji, and Tianxuan, and at least one of them was hired by each unit.name.Known? (1) Li was hired by Tianshu; (2) Li and Zhao were not hired by the same unit; (3) Liu and Zhao were hired by the same unit; (4) If Zhang was hired by Tianxuan, then Wang Ye Was hired by Tianxuan. **Compound Questions:** Please give a combination of graduate students and recruiting companies. Why do such combinations appear? **Compound Answer:** A combination of graduate students and recruiting companies can appear due to various reasons, such as:1. Skillset Match: Companies often look for candidates with specific skill sets that align with their job requirements. Graduate students, having completed their studies in a particular field, possess the necessary expertise and knowledge to meet these requirements.2. Networking: Graduate students often have connections with companies through their professors, research projects, or internships. These connections can lead to job opportunities for the students, and vice versa.3. Talent Pool: Companies may seek to build a strong talent pool of young, skilled professionals who can contribute to their growth and development. Hiring graduate students can help them achieve this goal.4. Innovation and Creativity: Graduate students are known for their innovative thinking and problem-solving skills, which can be valuable assets for companies looking to stay ahead in their respective industries.5. Diversity and Inclusion: Hiring graduate students from diverse backgrounds and educational institutions can help companies foster a more inclusive and diverse workplace culture.In the given scenario, the combination of graduate students Zhang, Wang, Li, Zhao, and Liu with recruiting companies Tianshu, Tianji, and Tianxuan can be explained by the following facts:1. Li was hired by Tianshu.2. Li and Zhao were not hired by the same unit.3. Liu and Zhao were hired by the same unit.4. If Zhang was hired by Tianxuan, then Wang was also hired by Tianxuan.These facts provide a logical explanation for the combination of graduate students and recruiting companies

Table 25: Case study of different LLMs for answering one compound questions.