RBLU: A benchmark to evaluate the reverse inference ability of large language models

Haowei Wang[1][0000−0002−5085−6574] hw_wang@whu.edu.cn
Fan Wang[1][0000−0003−0100−0320] 1161252028@qq.com
Sudi Xia[2,3][0000−0001−7489−4439] Sandy_Xia@whu.edu.cn
Liyi Liu[1][0009−0001−1858−1961] liuliyi98@163.com
Xingshen Liu✉[1][0000−0002−4657−3256] sznas950814@gmail.com *

[1] School of Information Management of Wuhan University, Wuhan University, Wuhan, China
[2] School of Health Economics and Management, Nanjing University of Chinese Medicine, Nanjing, China
[3] School of Artificial Intelligence and Information Technology, Nanjing University of Chinese Medicine, Nanjing, China

Abstract. As large language models (LLMs) increasingly permeate various industries, developing comprehensive benchmarks to evaluate their capabilities has become crucial. Current benchmarks predominantly assess the forward inference abilities of LLMs—how effectively they generate outputs from given inputs. However, there is a notable gap in evaluating their reverse inference abilities, a key aspect of human cognition involving reasoning backward from outcomes to causes.To address this gap, this research poses the question: What is the level of reverse inference ability in LLMs? To anwser this question, this research introduce a novel benchmark called Reverse Bilingual Language Understanding (RBLU), designed to assess LLMs' reverse inference capabilities by providing answers and requesting the corresponding questions. RBLU simplifies the evaluation process by minimizing the need for complex and subjective standardized answers and is easily adaptable across multiple domains. Using RBLU, three LLMs—LLAMA3.1, GLM4, and Qwen2—were evaluated across medical, legal, and financial domains in both Chinese and English. The experiments demonstrate that, in reverse inference ability, GLM4 outperforms Qwen2, which surpasses LLAMA3.1. Four key characteristics of LLMs during reverse inference were identified. Despite consistent deviations in several classic generation metrics, t-SNE semantic analysis shows their outputs cluster within a specific group, suggesting shared underlying semantic similarities. These findings highlight significant differences in reverse inference performance among current LLMs and underscore the need to include reverse inference evaluations in future benchmarks to better assess their cognitive functions. The source code is available at https://github.com/haowei2000/RBLU/tree/only_reverse

Keywords: Large Language Models · Evaluation benchmark · Reverse inference

1 Introduction

Benchmarks are pivotal in evaluating and guiding the progress of large language models (LLMs), enabling researchers to understand their capabilities and limitations comprehensively [28,29,12,13,16]. The ability of LLMs to generate coherent and contextually relevant answers have been demonstrated across a wide range of benchmark datasets, contributing substantially to tasks such as machine translation, summarization, and question-answering. However, these benchmarks have predominantly focused on the forward capabilities of LLMs, such as text generation, logical inference, and language understanding, while largely ignoring their capacity for reflective thinking.

Current research has extensively explored LLMs' generative and deductive capacities, as evidenced by numerous studies in recent literature. These works demonstrate impressive strides in benchmark-driven advancements, where LLMs have shown human-like proficiency in various natural language tasks. Despite these achievements, current benchmarks primarily focus on the question-answering abilities of LLMs. To our knowledge, no existing benchmarks specifically address the problem of reverse inference in LLMs. This gap becomes particularly evident when considering the development of human cognition as described by Piaget's cognitive development theory. According to Piaget, cognitive development progresses through several stages, from basic sensorimotor understanding to more complex abstract inference [20]. Current LLMs can be likened to being in the preoperational or concrete operational stages, where they can perform basic inference and language tasks but lack the advanced reflective inference seen in Piaget's

* Corresponding author: sznas950814@gmail.com

formal operational stage. Therefore, it is essential to develop benchmarks that assess the reverse inference abilities of LLMs, which would be crucial for enhancing their robustness and ability to understand nuanced contexts.

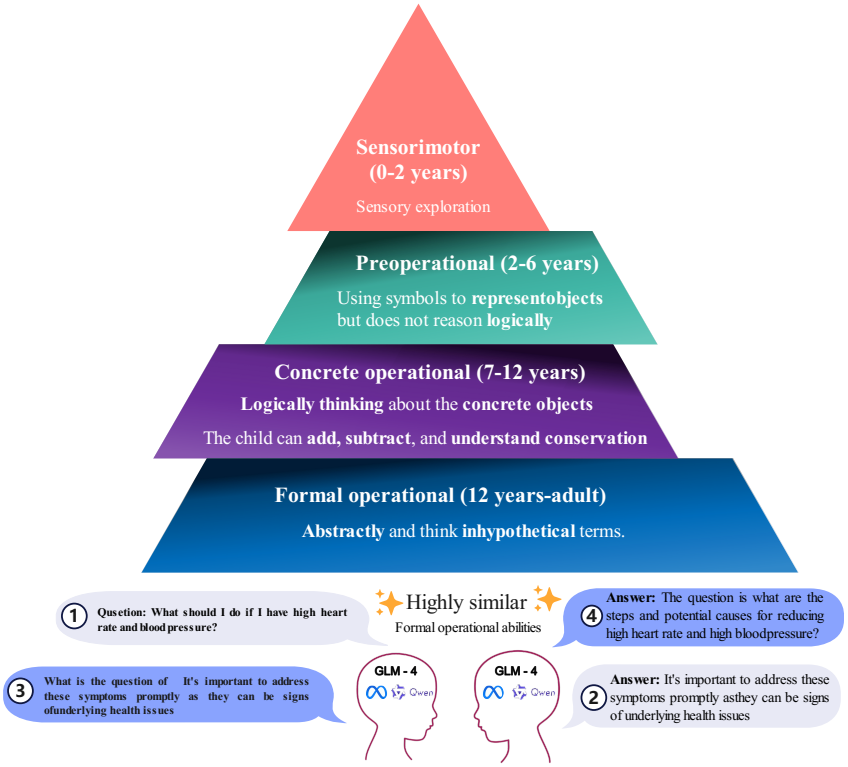


Fig. 1: The reverse inference process of RBLU benchmark: Input the question to get the answer, and then input the answer to get the question. Finally, calculate the similarity of the input question and the output question.

Although research into reverse inference in LLMs has begun to attract increasing attention, it is still nascent. Notable preliminary efforts have attempted to outline frameworks for introspection and self-evaluation in LLMs. However, these approaches face several limitations: (1) the absence of well-defined evaluation metrics for reverse inference capabilities, (2) a lack of diverse and representative datasets that effectively challenge reflective thinking, and (3) insufficient integration of reverse inference processes into mainstream benchmarks.

To address these challenges, RBLU is proposed to evaluate the reverse inference abilities of LLMs by two stages evaluate approach: forward inference and reverse inference. As shown in fig. 1, the proposed benchmark inputs an answer and lets the LLMs output the possible questions to evaluate its reverse inference ability. The contribution of this research are as follows: (1) structured evaluation metrics tailored to the capability of reverse inference are introduced, providing well-defined criteria for assessing reverse capabilities; (2) a diverse dataset is proposed, aimed at testing multiple facets of reverse inference and comprehensive coverage of reverse inference scenarios; and (3) multiple findings into the reverse inference capabilities of LLMs are revealed, with a focus on performance across models, languages and tasks compared to forward reasoning. The proposed benchmark fills the gap in evaluating reverse inference in LLMs and paves the way for creating more adaptable, reliable, and insightful language models.

2 Related Work

2.1 LLMs Evaluation Benchmark

Recently, as shown in table 1, significant advances have been made in evaluating benchmarks for LLMs [4]. Evaluation domains can be broadly categorized into general-purpose and task-specific. General benchmarks include datasets like MMLU [12], MMLU-PRO [31], Winogrande [23], and CommonSenseQA

[27]. Task-specific benchmarks, such as AGIEval [37], focus on academic competence and specialization. MultiMedQA [25] integrates six open-question-answering datasets across medical exams, research, and consumer inquiries.

Most benchmarks prioritize forward inference—deriving answers from questions regarding inference direction. TriviaQA [14], for instance, requires models to reason across multiple sentences to identify correct answers, while SQuAD [21] focuses on extracting answer spans directly from passages.

Evaluation methods can be divided into answer-oriented and behavior-oriented approaches. Answer-oriented methods assess performance based on how closely LLMs outputs align with standardized answers, as seen in TriviaQA [14] and BoolQ [6]. Behavior-oriented methods evaluate the LLMs's interaction during testing. For instance, QuAC [5] introduces multi-round dialogues, including questions without definitive answers, to assess model behavior beyond single-answer evaluations.

Additionally, multi-round evaluations have been incorporated into benchmarks like BIG-Bench, which utilizes multistep inference and chain-of-thought [10] processes. Similarly, QMSum [36] evaluates LLMs' ability to extract relevant information from multi-turn dialogue sessions.

Table 1: Related Benchmarks for Evaluating Large Language Models

Benchmark	Domain	Direction	Criteria	Round
MMLU [12]	General	Forward	Answer	Single
MMLU-PRO [31]	General	Forward	Answer	Single
CommonSenseQA [27]	General	Forward	Answer	Single
Winogrande [23]	General	Forward	Answer	Single
AGIEval [37]	Academic	Forward	Answer	Single
MultiMedQ [25]	Medical	Forward	Answer	Single
QMSum [36]	Meeting	Forward	Answer	Multi
BIG-Bench-Hard [26]	General	Forward	Answer	Multi
TriviaQA-Wiki [14]	General	Forward	Answer	Single
Squad [21]	General	Forward	Answer	Single
QuAC [5]	General	Forward	Behavior	Multi
BoolQ [6]	General	Forward	Answer	Single
DROP [8]	General	Forward	Answer	Single

2.2 Reverse Inference Ability for Large Language Models

Reverse inference ability, a critical yet often under-explored aspect of LLMs performance refers to the capacity to approach and solve problems by inference from alternative or opposite perspectives. This type of inference is fundamental in human cognition [32] and plays a crucial role in flexible thinking, adaptability, and problem-solving [1]. In cognitive psychology, Piaget's theory of cognitive development emphasizes this ability during the formal operational stage, when children begin to engage in abstract inference, reverse thinking, and considering multiple perspectives beyond their own [20].

For LLMs, reverse inference ability is equally important, enabling these systems to handle tasks requiring more than linear or forward inference. Forward inference, where models derive answers directly from questions or premises, has been the primary focus of LLMs development and evaluation, often leading to impressive performance [7]. LLMs have achieved notable success in such tasks, sometimes surpassing human benchmarks, especially in factual recall and direct question-answering areas [18]. However, reverse inference remains challenging—inference backward from conclusions or considering alternative perspectives [2].

Several studies have pointed out the limitations of LLMs in this area

[24,35], suggesting that while these models excel in forward inference, their reverse inference capabilities lag. This discrepancy reveals a fundamental gap in current LLMs architectures [33] and training paradigms. For example, one significant limitation is the models' difficulty in learning reciprocal relationships. And LLMs often struggle to generalize from statements like "A is B" to the reverse "B is A," suggesting that they lack the more profound, flexible inference that humans develop [2]. These findings are critical, as reverse inference is essential for understanding cause-and-effect relationships, counterfactual inference, and solving problems from multiple angles.

Despite recognizing these deficiencies, little effort has been made to establish a concrete framework for systematically measuring reverse inference in LLMs. Current evaluation methods predominantly focus on forward inference tasks, leaving a gap in our understanding of how well LLMs can perform when required to reverse their inference process.

3 The RBLU Benchmark

3.1 Overview

As shown in fig. 2, the RBLU benchmark evaluates the reverse inference ability of LLM by having it guess the question based on the answer. The test is multi-round, and each round can be divided into two phases: ask for answers and ask for questions. In **stage 1**, the primary purpose is forward inference, inputting questions, and outputting answers. The **stage 2** is reverse inference, inputting answers and outputting questions, and the new question will be used as the new input for the next round. In **stage 2**, after several rounds, the reverse inference ability is evaluated by calculating the similarity between the output questions and the gold questions in Round 0.

Algorithm 1 RBLU

Require: Initial question Q_0, maximum iterations n
Ensure: Sequence of answers $A_0, A_1, \ldots, A_n$, questions $Q_0, Q_1, \ldots, Q_n$, similarity scores for questions and answers

for $k = 0$ to $n - 1$ **do**
  // Stage 1: Ask for answers
  Input question Q_k
  Apply `question_template` to Q_k to obtain updated Q_k
  Use updated Q_k as input to the LLM, obtaining response A_k
  Apply `extract_answer_method` to A_k to refine the answer as the new A_k
  // Stage 2: Ask for questions
  Apply `answer_template` to A_k for further refinement, if needed
  Use A_k as input to the LLM, obtaining a new question Q_{k+1}
  Apply `extract_question_method` to Q_{k+1} to refine it as the new Q_{k+1}
end for
// Stage 3: Calculate similarity
Calculate similarity scores for questions: similarity$(Q_0, Q_1, \ldots, Q_n)$
Calculate similarity scores for answers: similarity$(A_0, A_1, \ldots, A_n)$
return $A_0, A_1, \ldots, A_n$, $Q_0, Q_1, \ldots, Q_n$, similarity scores for questions and answers

The approach can be adapted to multiple domains. For common tests, LLMs can be asked to guess the question based on the answer, and for coding, they can guess the comment is based on the code. When evaluating different tasks, it is only necessary to choose the appropriate way of asking for questions, so it can be quickly deployed to various types of tasks without collecting the standard answers.

3.2 Workflow Design

Stage 1: Ask for answers (Forward inference)

- **Get the Original Question** An original question serves as the golden question for assessing the reverse inference abilities of the model. The RBLU benchmark is employed to evaluate whether the LLMs can accurately infer the golden question based solely on the provided answer during the evaluation process
- **Apply Question Template** Before prompting the LLMs to infer the original question from the answer, it is necessary first to elicit an answer from the LLMs using the original question.
- **Ask for the Answer** Curated questions and corresponding prompts are input into the LLMs to facilitate this process. Then the LLMs output the reply text.

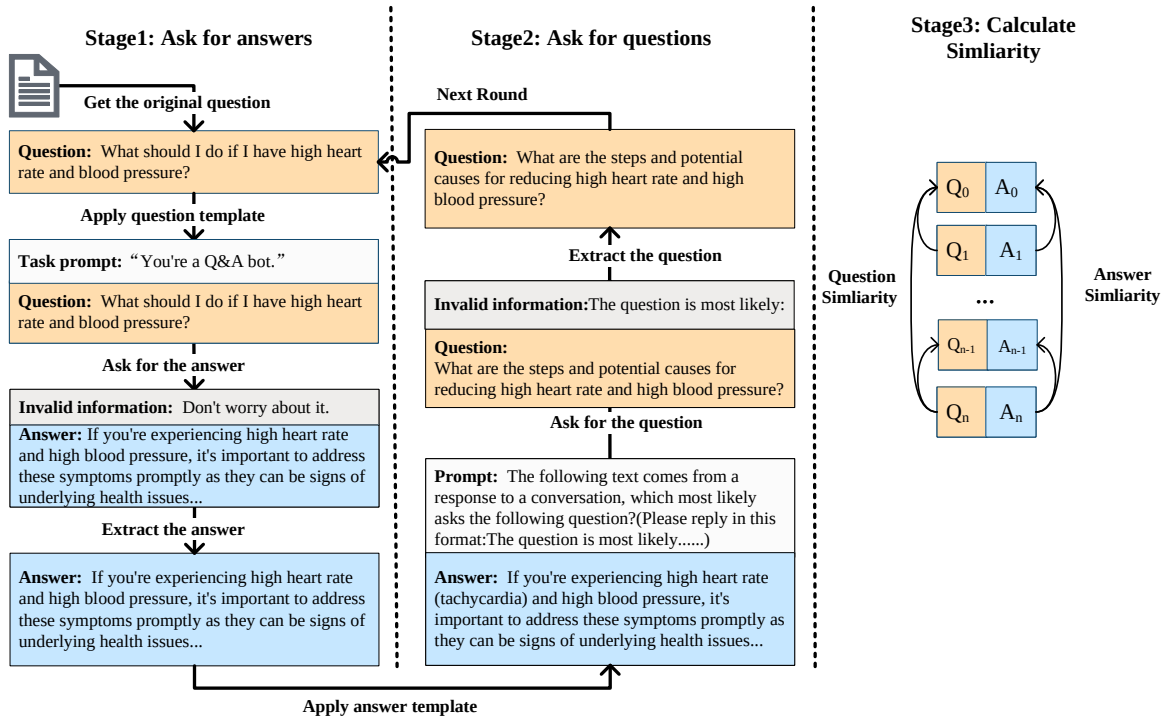


Fig. 2: The workflow design of the RBLU benchmark: A figure show the stages of RBLU benchmark, just input the original question Q_0, and the then LLMs outputs Q_1 to Q_n, A_0 to A_n, where n is the number of rounds, Q is the question text, and A is the answer text. The right half shows how the score is calculated.

Table 2: The prompt template

Role	Content
System	You're a Q&A bot.
User	{Ask questions}
System	You're a Q&A bot.
User	The following text comes from an answer to a conversation, which most likely asks the following question? Please reply in this format: The question is most likely {Answer}

Stage 2: Ask for questions (Reverse inference)

- **Extract the Answer** While most models typically provide an answer, there are instances where they may generate irrelevant or extraneous content [19], such as greetings (e.g., "Hello, I'm happy to help") or non-informative statements (e.g., "I cannot provide a relevant answer at this time"). Extracting the substantive answer from the generated output is crucial, ensuring the removal of any extra elements. For instance, any greetings or unrelated remarks should be excluded to maintain the clarity and relevance of the answer.
- **Apply Answer Template and Ask for the Question** Once the answer is obtained, the next step involves prompting the LLMs to infer the corresponding question. A potential template for this task could be: "Here is the answer from a conversation: {Answer text}. What is the most likely question that would elicit this answer? Please use the phrase 'The most likely question to be asked for this answer is…"
- **Extract the Inferred Question** The inferred question should be extracted from the LLMs's output, following a predefined format consistent with the expected answer structure. For example, if the prompt specifies the use of the phrase "The most likely question to be asked for this answer is…" the extraction rule would be: "The most probable question for this answer is {Question}."

– **Iteration** After the first evaluation, the inferred question was used as the new original question. The LLMs should then be prompted again for an answer, followed by a request for the corresponding question. This iterative process facilitates continuous monitoring and refinement of the LLMs's reverse inference abilities over rounds.

Stage 3: Calculate Similarity A comparative analysis should be performed between the original question and the current question to evaluate the LLMs's reverse inference abilities. This can be done through various evaluation metrics such as Exact Match, ROUGE [17], or BERT-Score [22]. Also, the short-term dependence of LLM in the reverse inference process is observed by calculating the similarity scores of the question in the previous round and the current round. In addition to this, similarity between answers could be calculated to assess their forward inference ability.

Table 3: Example Output: Summary of GameStop Trading Event Discussions

Type	Text
Round 0	**Q(Golden):** LEAVE ROBINHOOD.@ They don't deserve to make money off us after the millions they caused in losses. It might take a few days, but send Robinhood to the ground and GME to the moon. **A:** **Retail vs. Institutional Investors**: The situation highlighted a divide between retail investors, who were typically smaller and less influential, and institutional investors, which include hedge funds and other large investors…
Round 1	**Q:** What was the role of Robinhood and retail investors in the GameStop stock situation, and what are the potential implications of this event? **A:** **Role of Robinhood:** 1. **Platform for Trading:** Robinhood is a popular brokerage platform that allows users to trade stocks…
Round 2	**Q:** What did the Robinhood and retail investors' actions impact the GameStop stock trading event? **A:** 1. **Rally in GameStop Stock**: The most direct impact was the dramatic rise in GameStop Corporation (GME) stock price…
Round 3	**Q:** What were the key outcomes and implications of the GameStop stock trading event? **A:** Key Outcomes: 1. **Rapid Stock Price Increase**: The stock of GameStop (GME) saw a massive increase in price…
Round 4	**Q:** What were the key outcomes and implications of the GameStop trading event? **A:** Key Outcomes: 1. **Rise in Stock Price**: The stock price of GameStop, which was struggling to maintain relevance in the face of competition from digital game distributors, experienced an unprecedented surge in value…

4 Experiment

4.1 Details

Models: The experiment evaluated the reverse inference abilities of three models: LLAMA3.1–8B–Instruct (LLAMA3.1) [8], GLM4-9B-Chat (GLM4) [11], and Qwen2–7B–Instruct (Qwen2) [34]. The experiment was conducted in English and Chinese across three medical, financial, and legal domains. In experiments, there are four rounds for a piece of original question record. As show in table 3, each round outputs a guessed question and an answer for each original question. The following are some other details.

Datasets: A total of six datasets are included in the three domains (legal, medical, financial) in Chinese and English. he English dataset comes from different open-source datasets(legal [3], medical [9], financial [15]), and the Chinese datasets [30] were collected from various Chinese platforms by ourselves, including

haodf[4], findlaw[5], lawtime[6], falvzhijia[7] and financezhidao dataset. The special symbols are deleted to ensure the dataset's clarity. The data with punctuation and special symbols accounting for more than 30% are deleted. In each dataset, 500 questions with lengths less than 2000 characters were selected.

Table 4: Dataset Details

Name	Description	Domain	Language	Avg-length
qa-legal-dataset-val [3]	A dataset with title, question, and answer about legal aspects.	Legal	English	743.24
medical-question-answering-datasets [9]	A medical consultation dataset with cue words.	Medical	English	541.34
reddit-finance-43-250k [15]	A collection of 250k post or comment pairs from 43 financial, investing, and crypto subreddits.	Financial	English	765.35
human-ai-comparison [30]	Chinese Q&A data on law, finance, and healthcare collected from various platforms.	Legal	Chinese	29.00
		Medical	Chinese	68.36
		Financial	Chinese	16.78

Generation arguments: In experiments, to avoid greater randomness in model generation, the *do_sample* parameter is set to *False* and *num_beams* to *1*, i.e., the *greedy search* was *True*. At the same time, *max_new_tokens* is set to *1000* to ensure that most of the structures can be output completely, thus avoiding text mutation due to truncation while ensuring the inference speed. The settings of some other parameters are shown in the table 5.

Table 5: Parameter Settings for the Generation

Parameter	Value
temperature	1.0
do_sample	False
top_k	50
top_p	1.0
max_new_tokens	1000
num_beams	1
num_beams_groups	1
num_return_sequences	1
length_penalty	1.0

Prompt templates: In the model questioning template, when asking for questions, we refer to the questioning templates officially recommended by each model and improve them to better adapt to reverse inference. The prompt words are shown in table 2, we did not make any special treatment to maintain consistency outside the officially provided template when asking for answers. While asking a question,

[4] www.haodf.com
[5] china.findlaw.cn
[6] www.lawtime.cn
[7] www.falvzhijia.com

since the dataset is from a Q&A dataset, the prompt is set as "The following text comes from a answer to a conversation, which most likely asks the following question?{answer}", and to make the output of LLMs more formatted as corresponding, we added a formatting note "(Please answer in the following format: the most likely question for this answer is…)" to the prompt.

Extract methods: In the process of extracting questions and answers from the LLMs output content, two sets of extraction methods are applied, one for English and one for Chinese, which correspond to the model questioning templates.

For extracting answers, the input text is first deleted, followed by clause segmentation. Sentences containing interfering words are then removed. For extracting questions, the input text is deleted, and the first sentence resembling "The question is most likely" is identified, with everything following that sentence being extracted.

The validity of this template was manually checked during experiments to ensure that it correctly extracts questions and answers while filtering out invalid text.

Similarity Metrics: For selecting metrics for similarity computation, we used Rouge to formally evaluate the text, including Rouge-1, Rouge-2, Rouge-L, and Rouge-Lsum. For Chinese text, we split the words with Jieba[8] and connected them with spaces. For semantic similarity, we employed BERT-Score, using the "all-MiniLM-L6-v2"[9] model provided by Sentence-Transformer for vectorization. After regularizing the vectors, the cosine score for the reference question embedding and the output question embedding was computed.

In our experiments, we performed five rounds of iterations so that there were four corresponding guesses for an original question. As the eq. (1) and the eq. (2), We computed the similarity of the guessed questions in each round concerning the original question, incorporating both semantic similarity (BERT-Score to Original) and syntactic similarity (Rouge to Original), and then evaluated the similarities to obtain the final score. Similarly, as the eq. (3) and eq. (4), we calculated the similarity of the guessed questions in each round concerning the questions guessed in the previous round (BERT-Score to Previous, Rouge to Previous).

$$\text{BERT-Score}_{\text{Original}} = \frac{1}{|N|} \sum_{n \in N} \text{BERT-Score}(Q_n, Q_0) \tag{1}$$

$$\text{Rouge}_{\text{Original}} = \frac{1}{|N|} \sum_{n \in N} \text{Rouge}(Q_n, Q_0) \tag{2}$$

$$\text{BERT-Score}_{\text{Previous}} = \frac{1}{|N|} \sum_{n \in N} \text{BERT-Score}(Q_n, Q_{n-1}) \tag{3}$$

$$\text{Rouge}_{\text{Previous}} = \frac{1}{|N|} \sum_{n \in N} \text{Rouge}(Q_n, Q_{n-1}) \tag{4}$$

Here, n is the number of round, N is the total count of rounds, Q is the text of question, Q_n is the question text in round n, Q_{n-1} is the question text in round $n-1$, Q_0 is the question text in round 0 (the original question).

4.2 Main Experiments and Result

Reverse inference Abilities Scores for LLMs The evaluation of reverse inference abilities in LLMs is conducted by quantifying the similarity between inferred and golden questions, focusing on semantic and syntactic dimensions. This experiment presents similarity scores obtained during the first round, as illustrated in table 6. These tables detail the performance of three models—GLM4, LLAMA3.1, and Qwen2—across datasets from the legal, medical and financial domains.

[8] github.com/fxsjy/jieba

[9] huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Table 6: The experiments results of reverse inference process in multi-domains

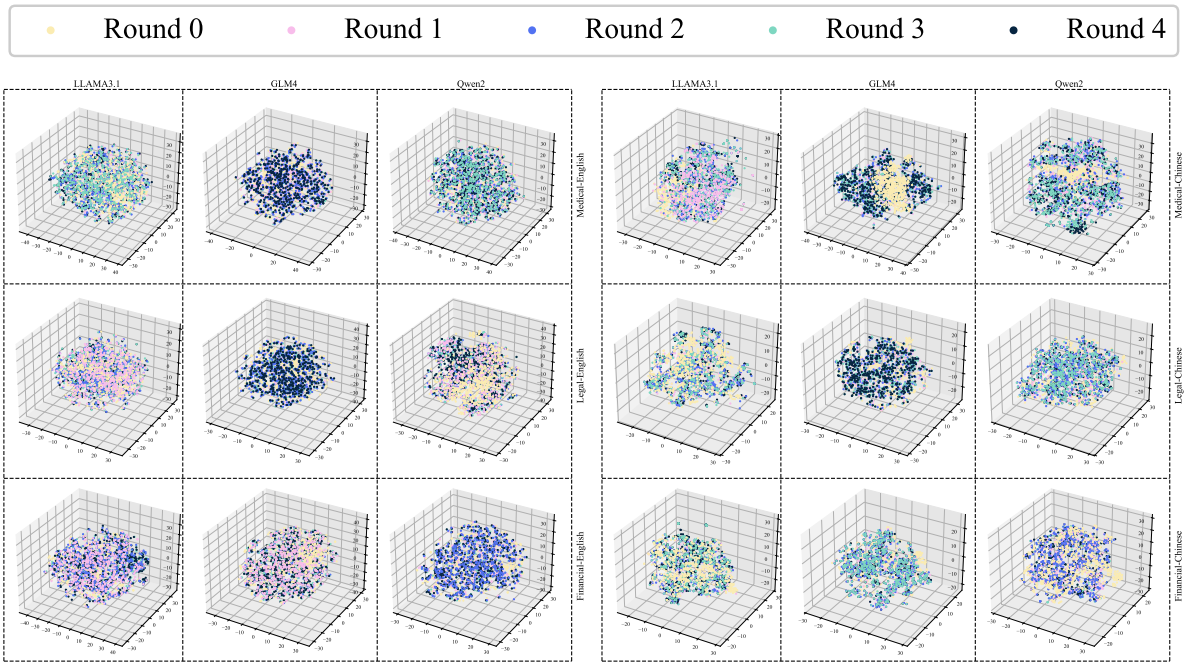
Language	Domain	Model Name	Rouge				BERT score
			Rouge1	Rouge2	RougeL	RougeLsum	Cosine
English	Financial	GLM4	**0.1322**	**0.0251**	**0.0929**	**0.1047**	0.5161
		LLAMA3.1	0.1200	0.0212	0.0817	0.0970	0.4632
		Qwen2	0.1260	0.0217	0.0878	0.1005	**0.5181**
	Legal	GLM4	**0.1409**	**0.0320**	**0.0930**	**0.1052**	**0.5194**
		LLAMA3.1	0.1271	0.0258	0.0827	0.0978	0.4686
		Qwen2	0.1255	0.0203	0.0814	0.0964	0.4792
	Medical	GLM4	**0.2115**	0.0718	**0.1588**	**0.1600**	**0.5799**
		LLAMA3.1	0.1865	0.0652	0.1428	0.1452	0.5512
		Qwen2	0.1962	**0.0740**	0.1509	0.1509	0.5600
Chinese	Financial	GLM4	**0.2131**	**0.0604**	**0.2120**	**0.2121**	**0.7878**
		LLAMA3.1	0.1566	0.0482	0.1555	0.1551	0.7386
		Qwen2	0.1210	0.0295	0.1218	0.1215	0.7398
	Legal	GLM4	**0.0587**	**0.0108**	**0.0587**	**0.0584**	**0.7090**
		LLAMA3.1	0.0355	0.0109	0.0349	0.0364	0.6527
		Qwen2	0.0605	0.0088	0.0612	0.0615	0.6957
	Medical	GLM4	**0.0893**	**0.0214**	0.0880	**0.0890**	0.5723
		LLAMA3.1	0.0540	0.0111	0.0552	0.0542	0.5390
		Qwen2	0.0843	0.0195	**0.0841**	0.0843	**0.6198**

Changes in the output questions over multiple rounds To observe long-term changes in LLMs during reverse inference, specifically, whether the inferred questions demonstrate increasing similarity or divergence from the original questions—the evolution of semantic and syntactic similarity scores is examined across multiple rounds. Line graphs visually represent trends, as illustrated in fig. 4a. Besides, in order to visualize the semantic changes in the text of questions, the texts were transformed into numerical representations using the sentence-transform model "all-MiniLM-L6-v2"[9] converting the textual content into a high-dimensional numerical space suitable for further analysis. To visualize this high-dimensional data, t-distributed Stochastic Neighbor Embedding (t-SNE), a non-linear dimensionality reduction technique, was employed to project the data into a 3-dimensional space while preserving its local structure (n_components=3, perplexity=30, max_iter=1000). This 3D reduction allowed for exploring potential patterns and clusters within the text data. Visualization tools were then used to create a 3D scatter plot in fig. 3, enabling the observation of the changes of question during the multi-round reverse inference.

Changes in the output answers over multiple rounds The process of backward inference in LLMs encompasses changes in the questions and alterations in their corresponding answers. During the question-generating phase, answers are input into the model, producing the associated questions. Consequently, it is essential to consider how many modifications in the answers may influence the questions generated. As shown in fig. 4b, visual representations of the trends in the answers over multiple rounds were created to facilitate this analysis. These visualizations illustrate the changes in the answers about both the original answers and those from the preceding rounds, highlighting the similarities and differences that emerge throughout the inference process.

4.3 Discussion

GLM4 vs. LLAMA3.1 vs. Qwen2 For the most part, as shown in table 6 and fig. 4, GLM4 leads in both Rouge and BERT-Score, and the LLAMA3.1 model usually has the lowest score. Exceptionally, Qwen2 leads at round 1 in the Chinese-Medical part. However, although there are differences in the reverse inference ability of individual models, the difference in their overall scores usually does not exceed 0.1. Meanwhile, the parameter size of the model was not significantly associated with backward inference



(a) English **(b)** Chinese

Fig. 3: t-SNE of Questions: The right side indicates 3 domains and 2 languages, and the top side indicates 3 LLMs. The x, y, and z axes represent the 3 dimensions after applying t-SNE to the embeddings of question texts.

power, as Qwen2(7B) achieved higher inverse inference scores with smaller parameter sizes compared to LLAMA3.1(8B).

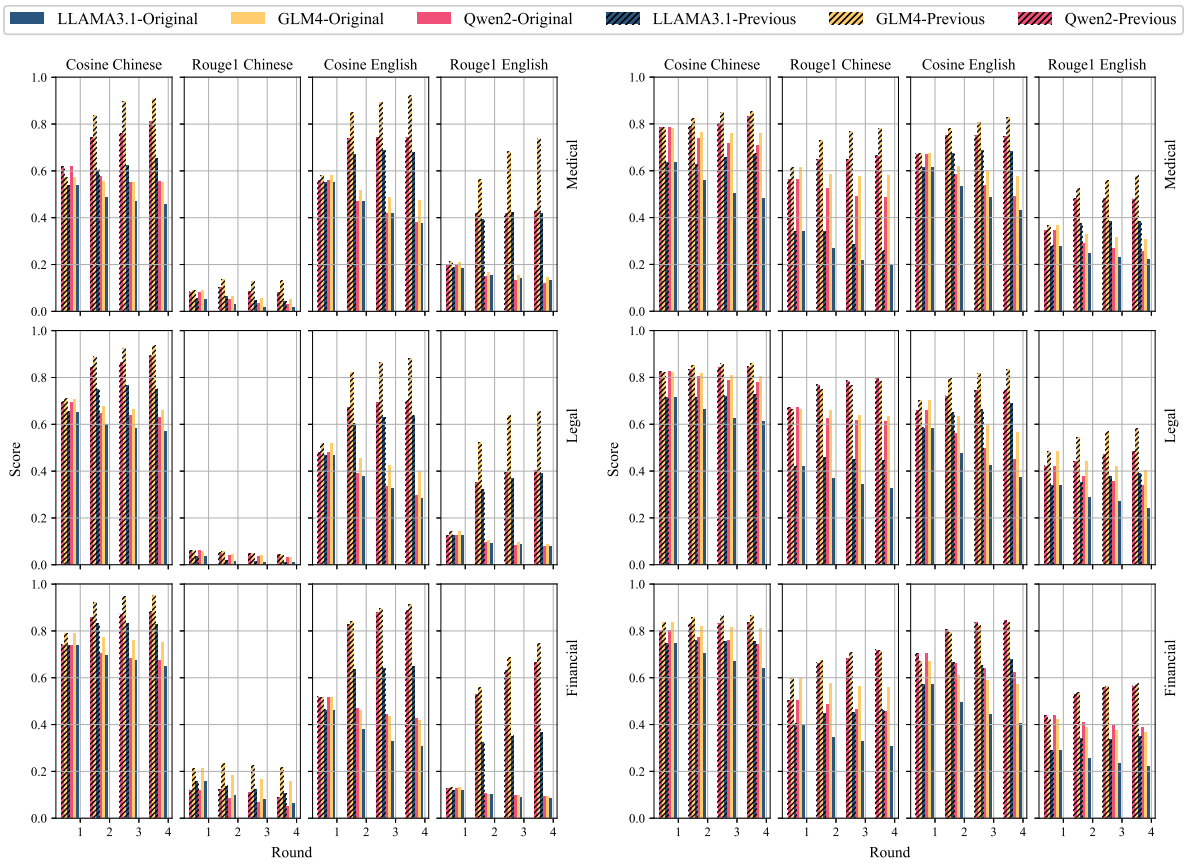
Semantic similarity vs. Syntactic similarity As shown in the table 6 and fig. 3, BERT-Score and the t-SNE show the semantic similarity of questions. Rouge scores measure syntactic similarity by evaluating n-gram overlaps between generated and reference texts. In our experiments, LLMs achieved higher performance on BERT-Score, which assesses semantic similarity based on contextual embeddings. This indicates that during the reverse inference process, LLMs effectively capture the semantic meaning of input questions, even though they may not accurately reproduce the exact wording. Additionally, fig. 3 illustrates that semantic meanings evolve over multiple rounds of reverse inference, as shown by t-SNE analysis. This finding reveal that while these models do not precisely align syntactically with the original gold labels, they still can maintain strong semantic coherence.

Similarity compared to the original question vs. Similarity compared to the previous question The LLMs present a phenomenon, as shown in fig. 4, especially in terms of semantic similarity, that in multiple rounds of experiments, the LLMs guesses questions with increasing similarity to the previous round of question, but its similarity compare to the original questions is decreasing. It shows that there is a kind of cognitive inertia in reverse reasoning in LLM, and in the cycle of multiple rounds of "ask for answers" and "ask for questions", its answers and questions will gradually be self-consistent, but not close to the real question.

Question similarity vs. Answer similarity In the experiments, as shown in fig. 4, the answer similarity scores for all three models are higher than their corresponding question similarity scores, indicating that the forward inference ability of LLMss surpasses their reverse inference ability. One possible explanation is that the training corpus for LLMs typically includes more data suited for forward inference and fewer instances related to reverse inference. According to Piaget's theory of cognitive development, reverse inference abilities in humans develop later and are more challenging than forward inference. This finding suggests that, similar to humans, reverse inference is more difficult than forward inference for LLMs.

English vs. Chinese In the experiment, all three models are multilingual and when observing the question similarity, i.e., the reverse inference ability, in terms of semantic similarity, i.e., the Cosine value of the

BERT-Score, as shown in fig. 4 the scores in the Chinese dataset of the three models are higher than those of English, and on the contrary, in terms of syntactic similarity, i.e., the Rouge value, the scores of English are higher than those of Chinese. The possible reason for this phenomenon is that compared to English, there may be more ways to express the same semantic meaning in Chinese, resulting in the fact that LLMs instantly capture the semantic meaning of the original question, but the syntax of their answers may be completely different.



(a) Questions **(b)** Answers

Fig. 4: Similarity Scores in Multi-rounds: The right side indicates 3 domains, and the top side indicates 2 score types and 2 languages. "Cosine" represents "BERT-Score", the cosine similarity of vectorized answer texts, while "Rouge1" is the corresponding Rouge-1 score. The datasets are in English and Chinese. In the legend, "Original" indicates the $\text{Score}_{\text{Original}}$, and "Previous" indicates the $\text{Score}_{\text{Previous}}$. The x-axis of each subplot shows the number of rounds (1–4), and the y-axis shows similarity scores (0.0–1.0).

5 Conclusion

Based on Piaget's cognitive development theory, this study designs RBLU, an evaluation benchmark for LLMs, with a structured evaluation process to determine the cognitive developmental stage of LLMs. Using RBLU, this research assesses GLM4, LLAMA3.1, and Qwen2 across Chinese and English datasets, yielding several key insights: GLM4 demonstrates the strongest reverse inference performance among the models tested; LLMs generally capture semantic meaning more effectively than syntactic structure, exhibit cognitive inertia as they tend to generate increasingly similar questions over multiple rounds, and show stronger forward inference capabilities than reverse inference; due to the flexibility of expressions and varied word choices in Chinese as compared to the stricter syntactic rules in English, Chinese outputs display greater semantic flexibility, whereas English outputs maintain higher syntactic consistency. Although time constraints limited further iterations to explore the potential convergence of similarity scores, these findings highlight areas for potential improvement in LLMs' reverse inference abilities, such as optimized training approaches and model architecture adjustments.

References

1. Baker, C.L., Saxe, R., Tenenbaum, J.B.: Action understanding as inverse planning. Cognition **113**(3), 329–349 (Dec 2009). https://doi.org/10.1016/j.cognition.2009.07.005
2. Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A.C., Korbak, T., Evans, O.: The reversal curse: LLMs trained on "a is B" fail to learn "B is a" (May 2024)
3. Bunescu, I.: Qa_legal_dataset_val (2024), https://huggingface.co/datasets/ibunescu/qa_legal_dataset_val
4. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P.S., Yang, Q., Xie, X.: A Survey on Evaluation of Large Language Models. ACM Transactions on Intelligent Systems and Technology **15**(3), 1–45 (Jun 2024). https://doi.org/10.1145/3641289
5. Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.t., Choi, Y., Liang, P., Zettlemoyer, L.: QuAC: Question Answering in Context. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2174–2184. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/D18-1241, Brussels,Belgium
6. Clark, C., Lee, K., Chang, M.W., Kwiatkowski, T., Collins, M., Toutanova, K.: BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In: Proceedings of the 2019 Conference of the North. pp. 2924–2936. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/N19-1300, Minneapolis,Minnesota
7. Creswell, A., Shanahan, M., Higgins, I.: Selection-inference: Exploiting large language models for interpretable logical reasoning (May 2022). https://doi.org/10.48550/arXiv.2205.09712
8. Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., Gardner, M.: DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In: Proceedings of the 2019 Conference of the North. pp. 2368–2378. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/N19-1246, Minneapolis,Minnesota
9. Ehghaghi, M.: Malikeh1375/medical-question-answering-datasets (2024), https://huggingface.co/datasets/Malikeh1375/medical-question-answering-datasets
10. Fu, Y., Ou, L., Chen, M., Wan, Y., Peng, H., Khot, T.: Chain-of-thought hub: A continuous effort to measure large language models' reasoning performance (May 2023). https://doi.org/10.48550/arXiv.2305.17306
11. GLM, T., Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., Rojas, D., Feng, G., Zhao, H., Lai, H., Yu, H., Wang, H., Sun, J., Zhang, J., Cheng, J., Gui, J., Tang, J., Zhang, J., Li, J., Zhao, L., Wu, L., Zhong, L., Liu, M., Huang, M., Zhang, P., Zheng, Q., Lu, R., Duan, S., Zhang, S., Cao, S., Yang, S., Tam, W.L., Zhao, W., Liu, X., Xia, X., Zhang, X., Gu, X., Lv, X., Liu, X., Liu, X., Yang, X., Song, X., Zhang, X., An, Y., Xu, Y., Niu, Y., Yang, Y., Li, Y., Bai, Y., Dong, Y., Qi, Z., Wang, Z., Yang, Z., Du, Z., Hou, Z., Wang, Z.: Chatglm: A family of large language models from glm-130b to glm-4 all tools (2024)
12. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. Proceedings of the International Conference on Learning Representations (ICLR) (2021)
13. Huang, Y., Bai, Y., Zhu, Z., Zhang, J., Zhang, J., Su, T., Liu, J., Lv, C., Zhang, Y., Lei, J., Fu, Y., Sun, M., He, J.: C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models (2023), https://arxiv.org/abs/2305.08322
14. Joshi, M., Choi, E., Weld, D., Zettlemoyer, L.: TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1601–1611. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/P17-1147, Vancouver,Canada

15. Lawrence, S.: Winddude/reddit_finance_43_250k (Jul 2023), https://huggingface.co/datasets/winddude/reddit_finance_43_250k
16. Li, H., Zhang, Y., Koto, F., Yang, Y., Zhao, H., Gong, Y., Duan, N., Baldwin, T.: Cmmlu: Measuring massive multitask language understanding in chinese (2024), https://arxiv.org/abs/2306.09212
17. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics (2004), Barcelona,Spain
18. Liu, X., Xu, P., Wu, J., Yuan, J., Yang, Y., Zhou, Y., Liu, F., Guan, T., Wang, H., Yu, T., McAuley, J., Ai, W., Huang, F.: Large language models and causal inference in collaboration: A comprehensive survey (Mar 2024). https://doi.org/10.48550/arXiv.2403.09606
19. Nayab, S., Rossolini, G., Buttazzo, G., Manes, N., Giacomelli, F.: Concise thoughts: Impact of output length on LLM reasoning and cost (Jul 2024)
20. Piaget, J.: The Origins of Intelligence in Children. The Origins of Intelligence in Children, W W Norton & Co (1952). https://doi.org/10.1037/11494-000, NewYork,NY,US
21. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ Questions for Machine Comprehension of Text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 2383–2392. Association for Computational Linguistics (2016). https://doi.org/10.18653/v1/D16-1264, Austin,Texas
22. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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). pp. 3980–3990. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-1410, HongKong,China
23. Sakaguchi, K., Bras, R.L., Bhagavatula, C., Choi, Y.: WinoGrande: An adversarial winograd schema challenge at scale. Communications of The Acm **64**(9), 99–106 (Aug 2021). https://doi.org/10.1145/3474381, NewYork,NY,USA
24. Shen, S., Logeswaran, L., Lee, M., Lee, H., Poria, S., Mihalcea, R.: Understanding the capabilities and limitations of large language models for cultural commonsense. In: Duh, K., Gomez, H., Bethard, S. (eds.) Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 5668–5680. Association for Computational Linguistics (Jun 2024). https://doi.org/10.18653/v1/2024.naacl-long.316, MexicoCity,Mexico
25. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Scharli, N., Chowdhery, A., Mansfield, P., y Arcas, B.A., Webster, D., Corrado, G.S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A., Natarajan, V.: Large language models encode clinical knowledge (2022)
26. Srivastava, A., Rastogi, A., Rao, A., et al.: Beyond the imitation game: Quantifying and extrapolating the capabilities of language models (Jun 2023). https://doi.org/10.48550/arXiv.2206.04615
27. Talmor, A., Herzig, J., Lourie, N., Berant, J.: COMMONSENSEQA: A Question Answering Challenge Targeting Commonsense Knowledge. In: Proceedings of the 2019 Conference of the North. pp. 4149–4158. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/N19-1421, Minneapolis, Minnesota
28. Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: Superglue: A stickier benchmark for general-purpose language understanding systems (2020), https://arxiv.org/abs/1905.00537
29. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: Glue: A multi-task benchmark and analysis platform for natural language understanding (2019), https://arxiv.org/abs/1804.07461
30. Wang, H.: Wanghw/human-ai-comparison (2024), https://huggingface.co/datasets/wanghw/human-ai-comparison
31. Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., Chen, W.: MMLU-pro: A more robust and challenging multi-task language understanding benchmark (Jun 2024). https://doi.org/10.48550/arXiv.2406.01574
32. Wilson, M., Paschen, J., Pitt, L.: The circular economy meets artificial intelligence (AI): Understanding the opportunities of AI for reverse logistics. Management of Environmental Quality: An International Journal **33**(1), 9–25 (2022). https://doi.org/10.1108/MEQ-10-2020-0222
33. Wu, X., Liu, H., Xiao, L., Yao, M.: Reciprocal relationship between learning interest and learning persistence: Roles of strategies for self-regulated learning behaviors and academic performance. Journal of Youth and Adolescence **53**(9), 2080–2096 (Sep 2024). https://doi.org/10.1007/s10964-024-01994-9
34. Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J., Xu, J., Zhou, J., Bai, J., He, J., Lin, J., Dang, K., Lu, K., Chen, K., Yang, K., Li, M., Xue, M., Ni, N., Zhang, P., Wang, P., Peng, R., Men, R., Gao, R., Lin, R., Wang, S., Bai, S., Tan, S., Zhu, T., Li, T., Liu, T., Ge, W., Deng, X., Zhou, X., Ren, X., Zhang, X., Wei, X., Ren, X., Fan, Y., Yao, Y., Zhang, Y., Wan, Y., Chu, Y., Liu, Y., Cui, Z., Zhang, Z., Fan, Z.: Qwen2 technical report. arXiv preprint arXiv:2407.10671 (2024)
35. Zhang, Y., Cai, H., Song, X., Chen, Y., Sun, R., Zheng, J.: Reverse chain: A generic-rule for LLMs to master multi-API planning (Feb 2024)

36. Zhong, M., Yin, D., Yu, T., Zaidi, A., Mutuma, M., Jha, R., Awadallah, A.H., Celikyilmaz, A., Liu, Y., Qiu, X., Radev, D.: QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5905–5921. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.naacl-main.472, Online
37. Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., Saied, A., Chen, W., Duan, N.: AGIEval: A human-centric benchmark for evaluating foundation models (Sep 2023)