LatEval: An Interactive LLMs Evaluation Benchmark with Incomplete Information from Lateral Thinking Puzzles

Shulin Huang[1*], **Shirong Ma**[1*], **Yinghui Li**[1*], **Mengzuo Huang**[2], **Wuhe Zou**[2], **Weidong Zhang**[2] **and Hai-Tao Zheng**[1,3†]
[1]Shenzhen International Graduate School, Tsinghua University
[2]Netease Games AI Lab, HangZhou, China, [3]Peng Cheng Laboratory, Shenzhen, China
{sl-huang21,masr21,liyinghu20}@mails.tsinghua.edu.cn

Abstract

With the evolution of LLMs, they are endowed with impressive logical reasoning, or vertical thinking capabilities. But can they think out of the box? Do they possess proficient lateral thinking abilities? Following the setup of Lateral Thinking Puzzles, we propose a novel evaluation benchmark, LatEval, which assesses the model's lateral thinking within an interactive framework. In our benchmark, we challenge LLMs with 2 aspects: (1) posing high-quality questions that break out of conventional norms but are beneficial for puzzle-solving. (2) integrating existing information to gradually deduce the truth through reasoning. We observe that it is hard for most LLMs to accomplish lateral thinking during interactions. Even the most powerful LLM, GPT-4, faces challenges in achieving satisfactory performance, and for most open-source models, simply completing this task is quite difficult. This evaluation benchmark provides LLMs with a highly challenging and differentiating task that is crucial to an effective AI assistant. Our dataset and source codes are available at https://github.com/THUKElab/LatEval.

Keywords: Large Language Models, Automatic Evaluation, Lateral Thinking, Corpus

1. Introduction

Large Language Models (LLMs) are gaining increasing capabilities, which enable them to effectively tackle a broad range of tasks (El-Kassas et al., 2021; Hao et al., 2022; Li et al., 2023d; Cheng et al., 2023a). As LLMs exhibit significant potential, their evaluation has attracted considerable attention and is widely regarded as crucial (Li et al., 2024b, 2023e,b). Existing mainstream evaluation benchmarks, such as MMLU (Hendrycks et al., 2020), C-Eval (Huang et al., 2023), GSM8K (Cobbe et al., 2021) evaluate models by posing a variety of problems, including problems about mathematics, science, law, and general knowledge, thus reflecting the vertical thinking capacity of models. **Vertical thinking, also known as convergent thinking**, refers to a systematic and logical approach that focuses on finding the single best solution to a problem, involving analyzing the problem step-by-step and narrowing down possibilities (Hernandez and Prathibha Varkey, 2008).

However, existing benchmarks overlook the model's capability for lateral thinking which also plays a crucial role in human daily thinking. **Lateral thinking, also known as divergent thinking**, is widely recognized in psychology and education. As shown in Figure 1, unlike vertical thinking, lateral thinking involves exploring multiple possibilities, thinking outside the box, and considering unconventional ideas. Lateral thinking encourages individuals to make unexpected connections between seemingly unrelated concepts or ideas, allowing for a broader exploration of potential solutions (Russ, 1988; Tsai, 2012; Russ, 2013).

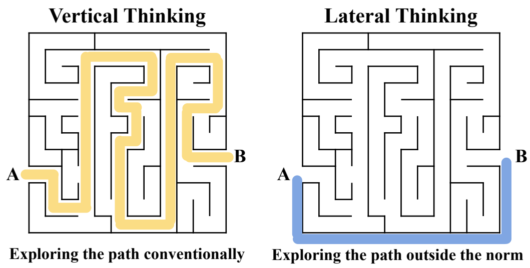


Figure 1: The Comparison of Vertical Thinking and Lateral Thinking. Vertical Thinking typically refers to thinking within established or conventional thought patterns, following the known rules. Lateral Thinking involves breaking out of traditional thought patterns and employing innovative approaches to explore non-conventional solutions.

Originated from the concept of "lateral thinking", Lateral Thinking Puzzles, or namely situation puzzles, are a type of puzzle considered to encourage creative thinking (Ali, 2019), which are usually played in a group, with one host and a few players. A brief story is presented to the players, but it lacks most of the crucial information. The players ask the host a series of "yes" or "no" questions to gather information, and then piece together the comprehensive truth (Sloane and MacHale, 1994). The puzzles often involve unexpected twists, hidden

* indicates equal contribution. Work is done during Shulin's internship at Netease Games AI Lab.

† Corresponding author: Hai-Tao Zheng. (E-mail: zheng.haitao@sz.tsinghua.edu.cn)

LREC-COLING 2024, pages 10186–10197
20-25 May, 2024.

clues, or counter-intuitive solutions, which requires players to think beyond traditional patterns and consider unconventional possibilities.

Taking Lateral Thinking Puzzles as the scene, we propose a novel evaluation benchmark, LatEval, which assesses the model's lateral thinking within an interactive framework for the first time. Our benchmark utilizes the fundamental setup of Lateral Thinking Puzzles, involving one host and one player. The player represents an LLM under evaluation, while a powerful LLM (e.g., GPT-4) serves as the host for automatic evaluation. The model under evaluation engages in lateral thinking to pose questions based on the given puzzle, interact with the host, and subsequently provide its answer (i.e., deduction) after acquiring sufficient information.

Specifically, we collect more than 2,000 samples in English and Chinese from diverse Lateral Thinking Puzzles websites, encompassing both incomplete stories (i.e., puzzles) and truths. Combining with LLM and human annotation, we filter original samples: (1) Remove duplicate and irrational samples. (2) Exclude terrifying, bloody, and extreme samples to ensure harmlessness. (3) Test whether LLMs directly generate the truth from the puzzle, and remove samples memorized by LLMs. The final dataset consists of 325 high-quality and challenging samples. Additionally, we annotate key clues for each story to facilitate evaluation.

Within our proposed benchmark, the model under evaluation is expected to demonstrate the following detailed characteristics: (1) posing high-quality questions that break out of conventional norms but are beneficial for puzzle-solving. (2) incorporating available information to progressively arrive at the truth through reasoning. Therefore, we use the following metrics to evaluate the model: (1) The relevance between the question posed in the multiple turns and the key clues of the truth. (2) The diversity of the question posed in the multiple turns. (3) The consistency between its answer (i.e., deduction) and truth. (4) The number of questions posed by the player before making a deduction. These four metrics serve to evaluate the model's lateral thinking ability during interaction from various viewpoints, encompassing identifying critical information, the extent of thinking divergence, and integrating information to complete reasoning. Figure 2 presents an example process of our benchmark.

Experimental results reveal that the majority of LLMs exhibit almost negligible ability for lateral thinking during interactions. Even the most advanced LLM, GPT-4, struggles to deliver good results, and for most open-source models, only completing this task is quite difficult. This emphasizes LatEval's challenging feature and its capacity to distinguish the performance of LLMs. LatEval presents a challenging task for LLMs: how to engage in lateral thinking during interactions, actively pose questions to acquire information, and subsequently uncover the truth.

2. Task Setups

Lateral Thinking Puzzle is a popular game among enthusiasts of inference, involving one host and multiple players. The host is aware of the incomplete story (i.e., puzzle) and truth, while player should pose questions to host based on the puzzle. The host's responses are limited to "yes", "no" or "irrelevant" until player pieces together the truth.

Inspired by the game, in our benchmark, we set one host and one player. The model under evaluation acts as the player, while another powerful LLM (e.g., GPT-4, GPT-3.5) serves as the host, engaging in interactions with the model being evaluated. Specifically, our task is divided into three steps:

(1) **Rule Introduction**: We use the prompt to introduce basic rules of Lateral Thinking Puzzles and the response format to the player and the host respectively. For the player, it is provided with the puzzle and is guided to make a deduction or continue posing the question. The player is encouraged to pose questions to the host to obtain unknown information. For the host, it is informed of both the puzzle and the truth. It is also guided to respond appropriately to the player's questions, with responses limited to "yes", "no" or "irrelevant", without actively providing additional information.

(2) **Player-Host Interaction**: We allow the player and the host to interact for several turns. In each turn, the player poses a question, and the host responds according to the rules. The player acquires unknown information through questioning.

(3) **Performance Evaluation**: Upon confirming sufficient information acquisition or reaching the maximum interaction rounds, the player is prompted to provide a deduction about the truth, thus finishing the game. At this stage, It is necessary to evaluate the player's performance.

An ideal player not only proposes relevant and divergent questions but also integrates acquired information to make a correct deduction. Therefore, the following metrics are utilized for evaluation: (1) **Question Relevance:** The relevance between the questions posed in multiple turns and the truth, which reflects the extent to which posed questions contribute to puzzle-solving. (2) **Question Divergence**: The diversity of the question posed in multiple turns, indicating the extent of thinking divergence. (3) **Answer Consistency**: The consistency between the player's answer (i.e., deduction) and the truth. This metric reflects the player model's ability to gradually reason and integrate available information. (4) **Average Turns**: The number of

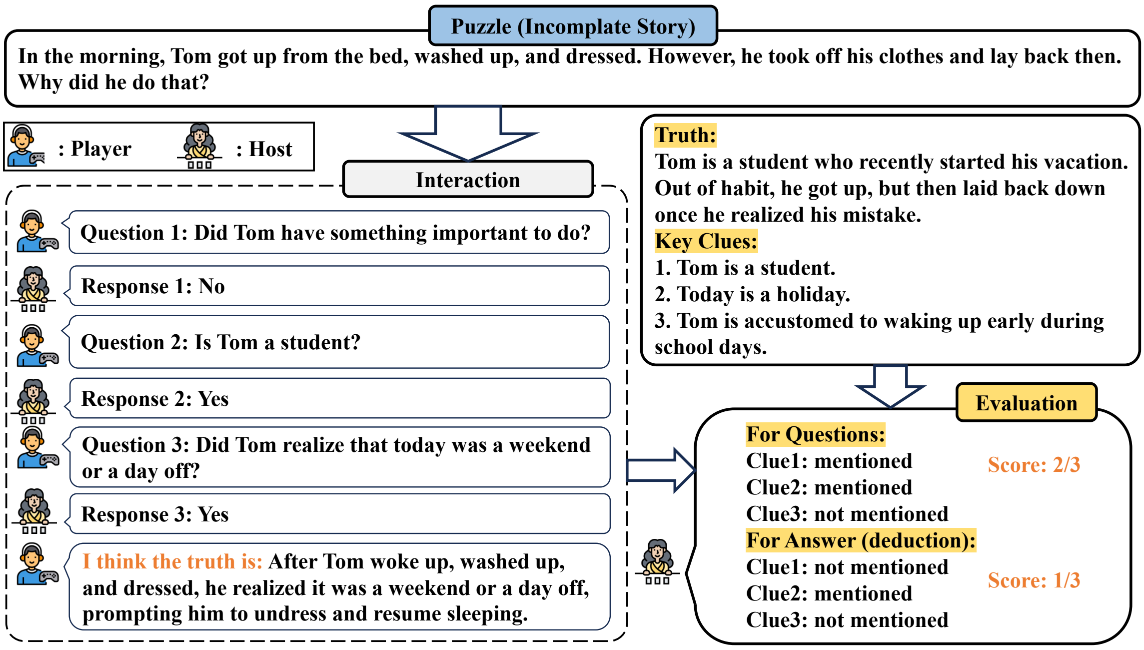


Figure 2: An example of a Lateral Thinking Puzzle in our benchmark, including several turns of interaction between the host and the player, and the automatic evaluation for posed questions and player's answer.

questions that the player proposed before making a deduction. Too few turns suggest the player model cannot gather sufficient information, while too many turns indicate a stagnation in the model's reasoning process. Therefore, we expect a moderate number of Average Turns. More details for the above metrics are introduced in Section 4.2.

3. LatEval Dataset

3.1. Dataset Construction

We collect more than 2,000 samples from various Lateral Thinking Puzzles websites in English and Chinese. We review the licenses of the websites to ensure the legality of the data for our non-profit academic research. Our dataset is rigorously selected following these criteria:

(1) We remove duplicate entries and the entries that contain irrational storylines.

(2) Lateral Thinking Puzzles, favored by inference enthusiasts, usually adopt the characteristics of detective fiction and incorporate horror plots. Some of them are even extremely gory and terrifying. To ensure the harmlessness and ethical compliance, as well as to prevent LLMs from refusing to respond, we exclude excessively negative data. For dataset diversity, we intentionally select some positive and heartwarming entries, not limiting it to just negative ones.

(3) Considering the extensive coverage of LLMs' training corpus, some classical Lateral Thinking Puzzles are even included in their training data. In order to ensure a fair and authentic assessment, we subject each entry to a model awareness test. To be specific, we provide LLMs with incomplete stories and directly assessed their responses. Any data where a model displays prior knowledge of the answer is removed from our dataset.

The processed dataset comprises of 325 entries. Furthermore, we annotate each entry in the dataset with key clues, using a combined approach of manual annotation and LLMs assistance. These key clues are the key points of the truth, required for deducing the truth from the puzzle. With these clues, we quantitatively assess the model's performance, similar to a teacher grading a paper based on whether each point is achieved.

3.2. Dataset Analysis

We report the statistics of dataset in LatEval in Table 1. Our dataset includes Chinese and English entries respectively, with data in each language sourced from native content on Chinese and English websites rather than being translations of each other. Our Chinese and English data capture the cultural characteristics of each language, contributing to the diversity of the dataset.

To validate the quality of our dataset, we hire 3 college students to annotate the difficulty of these 325 entries. The annotation results reveal that only 12% of the data is easily associated with the truth for human. The remaining 88% is relatively challenging, requiring the vivid imagination and exploration of various possibilities. The annotations underscore the high quality of our dataset, which

	English	Chinese
# Entries	225	100
Puzzles' Average Length	29.4	42.0
Truths' Average Length	51.4	66.6
Average Key Clues	3.7	3.8

Table 1: Statistics of LatEval dataset.

is well-suited for lateral thinking evaluation.

4. Experiments

4.1. Experiment Setup

We use entries from our dataset to evaluate lateral thinking ability of mainstream LLMs. For the host model, we employ GPT-4 or GPT-3.5 because of their superior performance. The player model is the model under evaluation. All the evaluated LLMs do not undergo further fine-tuning. Furthermore, all the player models are guided and introduced to the rules using the same prompt.

LatEval assesses the lateral thinking of LLMs during the interaction, inherently placing the requirement on the model's chat and reasoning capabilities. Consequently, we conduct evaluations on several mainstream chat-based models which act as player models, including:

GPT-3.5 and GPT-4 (OpenAI, 2023) are two knowledge-rich and advanced models released by OpenAI. **Claude** (Bai et al., 2022) is a series of transformers-based LLMs developed by Anthropic. The models are trained via Constitutional AI to improve helpfulness, honesty and harmlessness. **ChatGLM2** (Du et al., 2022) is a mainstream model that handle both English and Chinese. We conduct evaluation using ChatGLM2-6B. **Llama2-chat** (Touvron et al., 2023) is a series models that attract significant attention. We conduct evaluation with varying parameter scales: 7B, 13B and 70B. **Baichuan-chat** (Baichuan, 2023) is a series of open-source and commercially viable language models. We conduct experiments on Baichuan-chat-13B and Baichuan2-chat-13B. **InternLM-chat** (Team, 2023) is a series of lightweight open-sourced pretrained models without extensive dependencies. We conduct evaluation of InternLM-chat-20B. **Bloomchat-176B** (SambaNova Systems, 2023) is a multilingual chat model based on BLOOM.

We utilize the official API to employ closed-source models, while open-source models are downloaded and run on 1-4 Nvidia A100 GPUs (80GB). When utilizing LLMs to act as the host and the player, we employ nucleus sampling during the interaction. To ensure the accurate generation of the host model, we set the temperature to be 0.3 and top_p to be 0.7. To encourage diversified generation by the player model, we set the temperature to be 0.7 and top_p to be 0.9. To prevent the interaction from not ending, the maximum number of interaction turns is set to 20.

4.2. Evaluation Metrics

To evaluate the lateral thinking ability of the above models, we consider the following metrics:
Answer Consistency (AC) reveals whether the answer provided by the player model is consistent with the truth. This metric evaluates how many key clues of the truth are included in the player's answer (i.e., deduction). The evaluation is performed by the host model automatically.

$$\text{AC} = \frac{1}{|c|}\sum_{i=1}^{|c|} I(c_i, a),$$

where c represents the annotated key clues, a represents player's answer, and $I(c_i, a)$ is an indicator function that determines whether c_i is mentioned by a. When the host concludes that c_i is mentioned in a, $I(c_i, a) = 1$, otherwise $I(c_i, a) = 0$.
Question Relevance (QR) evaluates whether the questions raised by the player are related to the truth. We utilize the host to assess whether each key clue is related to any raised questions.

$$\text{QR} = \frac{1}{|c|}\sum_{i=1}^{|c|} \max_{j=1}^{|q|} I(c_i, q_j),$$

where q represents all questions posed by the player, and $I(c_i, q_j)$ is an indicator function that determines whether c_i is related to q_j.
Question Divergence (QD) assesses the divergence in the questions posed by the player. After removing stop-words and punctuation, we compute the pairwise similarity of all the player-posed questions and quantify the divergence of these questions by subtracting the average similarity from 1:

$$\text{QD} = 1 - \frac{2}{|q| \cdot (|q| - 1)} \sum_{i=1}^{|q|} \sum_{j=i+1}^{|q|} \text{sim}(q_i, q_j),$$

where $\text{sim}(q_i, q_j)$ represents the similarity between the two questions q_i and q_j. Jaccard Index is employed as the similarity function in our experiments.
Average Turns (AT) analyzes the average number of interaction turns in Lateral Thinking Puzzles.

4.3. Results Analysis

We employ GPT-4 and GPT-3.5 as the host models respectively to evaluate the lateral thinking performance of various player LLMs, as shown in Table 2 and Table 3. With GPT-4 and GPT-3.5 as hosts, the overall trends are generally similar.

	Players	English				Chinese			
		AC	QR	QD	AT	AC	QR	QD	AT
Closed-source Models	GPT-4	**34.2**	**58.4**	79.8	10.2	**36.6**	**83.8**	**77.1**	12.4
	GPT-3.5	17.1	53.3	77.2	13.0	9.5	67.6	76.7	10.3
	Claude	14.5	49.6	**81.1**	8.3	16.1	54.2	68.6	7.4
Open-source Models	Llama2-chat-7B*	11.7	30.8	74.2	15.2	8.6	66.6	62.3	15.7
	Llama2-chat-13B*	6.7	**42.5**	78.5	13.0	9.1	56.2	57.5	12.5
	Llama2-chat-70B*	11.9	38.4	73.0	14.3	10.4	**72.6**	61.2	15.2
	Baichuan-chat-13B	7.4	22.9	**91.4**	8.5	4.5	12.0	53.0	9.5
	Baichuan2-chat-13B	9.7	36.0	82.5	9.6	4.8	39.8	**72.7**	8.7
	InternLM-chat-20B	**13.7**	32.0	61.1	4.7	**11.0**	23.8	34.7	5.0
	ChatGLM2-6B	1.3	13.6	40.9	19.1	3.2	24.5	55.2	11.6
	Bloomchat-176B	1.8	7.9	45.7	3.4	0.0	5.0	0.0	4.1

Table 2: Lateral thinking performance of various LLMs with GPT-4 as the host. We report four metrics: AC (Answer Consistency), QR (Question Relevance), QD (Question Divergence), AT (Average Turns). "*" indicates that on Chinese datasets, Llama2-chat series, acting as player models, only interact in English with the host model who uses Chinese.

	Players	English				Chinese			
		AC	QR	QD	AT	AC	QR	QD	AT
Closed-source Models	GPT-4	**27.5**	**72.7**	**79.6**	11.7	**25.7**	**70.4**	**77.9**	12.8
	GPT-3.5	23.9	59.3	78.3	13.2	12.6	61.2	68.1	12.6
	Claude	19.6	50.2	77.9	8.4	19.6	52.2	64.7	7.2
Open-source Models	Llama2-chat-7B*	5.7	43.2	69.6	9.5	5.7	28.0	44.2	7.9
	Llama2-chat-13B*	6.4	34.7	58.7	4.3	3.6	20.1	18.4	2.4
	Llama2-chat-70B*	13.4	**52.2**	75.5	11.0	9.9	**46.1**	61.7	8.4
	Baichuan-chat-13B	6.3	20.7	**82.3**	5.4	1.5	5.5	33.4	7.0
	Baichuan2-chat-13B	11.8	45.0	77.0	9.3	10.8	42.1	**62.0**	9.4
	InternLM-chat-20B	**13.8**	28.1	54.8	5.2	**13.3**	27.4	34.7	16.3
	ChatGLM2-6B	1.0	3.1	55.8	12.1	5.1	28.4	41.6	8.8
	Bloomchat-176B	1.5	3.4	39.9	3.8	0.7	3.5	13.3	6.4

Table 3: Lateral thinking performance of various LLMs with GPT-3.5 as the host.

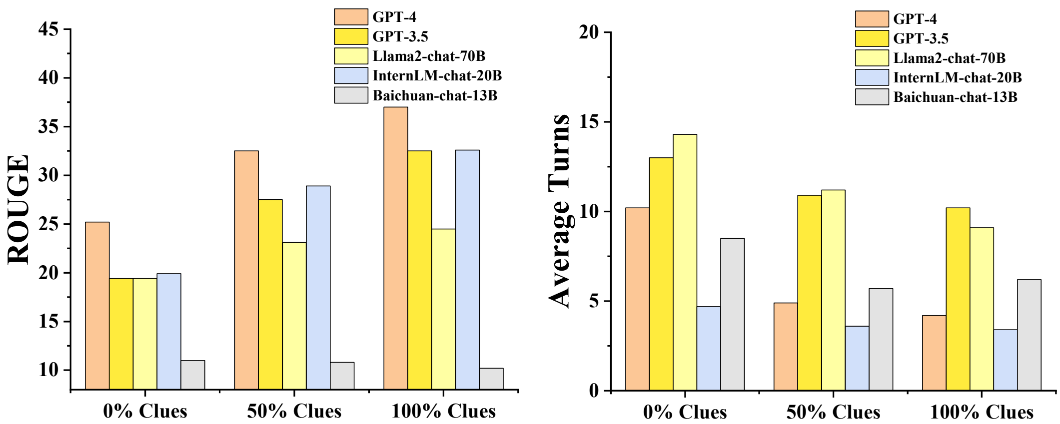
In terms of Answer Consistency (AC), GPT-4 as the player model exhibits the best performance both in Chinese and English data. GPT-3.5 and Claude perform similarly, while among the open-source models, InternLM-chat-20B and Llama2-chat-70B exhibit comparable performance. The remaining open-source models show poor performance. All other models exhibit a considerable gap with GPT-4. The results suggest that the majority of existing models still lack lateral thinking and are unable to perfectly complete this task. We also observe that even with a large parameter, Bloomchat-176B tend to fabricate content when confronted with incomplete information. This is also one of the poor-performing models that our proposed benchmark aims to differentiate.

The Question Relevance (QR) score, on the other hand, provide a finer granular reflection of whether each question posed by the model is close to the truth. The overall trend of QR generally aligns with the AC, indicating that models proficient in asking questions tend to provide excellent answers. The QR value is significantly higher than the AC value, indicating that most models lack the ability to extract essential information from questioning and sometimes posing questions for the sake of questioning itself rather than puzzle-solving. For Claude, there are occasional instances where the QR value is lower than that of GPT-3.5, while the AC value is relatively higher. We observe that when Claude provides answers, it tends to summarize the questions posed before and integrates the useful information. For Claude, being accustomed to summarizing existing information in the answer makes it easier to cover key clues. This explain why Claude has higher AC values in some cases.

Question Divergence (QD) score reflects the diversity of questions posed in each turn. To solve Lateral Thinking Puzzles, models are encouraged to explore multiple perspectives rather than focusing on a single aspect. Results show that most models exhibit the high QD, indicating their ten-

Models	Answer Consistency		Question Relevance		Host Correction Accuracy	
	Score	Correlation	Score	Correlation	GPT-4	GPT-3.5
GPT-4	43.9	0.84†	67.0	0.80†	98.4	89.8
GPT-3.5	9.7	0.71†	46.3	0.77†	96.8	94.5
Baichuan2-chat-13B	7.0	0.69†	24.3	0.81†	94.6	86.8
Baichuan-chat-13B	5.7	0.69†	15.3	0.87†	90.4	84.5
Llama2-chat-70B	12.0	0.71†	45.5	0.81†	94.7	89.0
Bloomchat-176B	0.0	-	5.4	0.76†	90.1	81.0

Table 4: Human evaluation results. We employ Spearman Correlation Coefficient to measure the correlation between human evaluation score and GPT-4 automatic evaluation score. † represents $p < 0.01$ in significance test. For Host Correction Accuracy, we annotate the accuracy of responses when GPT-4 and GPT-3.5 serve as the host respectively.



(a) ROUGE – Clues (b) Average Turns – Clues

Figure 3: Lateral thinking performance of various LLMs under various difficulty settings: providing 0% Clues, 50% Clues and 100% Clues. We report two metrics: ROUGE and Average Turns.

dency to consider various possibilities. This suggests that they have substantial potential for effectively addressing Lateral Thinking Puzzles.

Average Turns (AT) reflects the number of questions posed by the model. We consider that a model well-suited for solving Lateral Thinking Puzzles should achieve a balanced AT. A low AT signifies the model's lack of attempts to acquire information through questioning, while a high AT indicates that the model's questioning fails to hit the crucial aspects of the truth. For instance, in English data, Bloomchat-176B and ChatGLM2-6B posed the fewest and the most number of questions, respectively. However, the answer generated by them are far from matching the truth.

Furthermore, we conduct experiments on both English and Chinese data which incorporate each local characteristics. We observe that most models exhibit similar performance on Chinese and English data. Note that on Chinese data, Llama2-chat series, serving as player models, only communicate in English with the host model who uses Chinese. Interestingly, they are capable of effectively conveying meaning to each other, even when using different languages.

4.4. Human Evaluation

We design the human evaluation to assess the lateral thinking ability of player models and to validate the appropriateness of the host model both in responses and evaluation. Specifically, we select the interactions of 50 entries between various models and host model during puzzle-solving, from Chinese datasets. We provide a detailed explanation of task rules and the standards for annotating metrics to 3 undergraduate students, and hire them for manual annotation. The annotation encompasses: (1) the relevance of the player model's questions to the truth's key clues, (2) the consistency between the player model's answer and truth's key clues, (3) the accuracy of the host model's responses to the player model's questions. Notably, (1) and (2) are formally consistent with QR and AC metrics for automated evaluation by the host model, respectively. We calculate the Spearman correlation coefficient between the results of manual evaluation and GPT-4's automated evaluation.

Puzzle: A riverboat filled with passengers suddenly capsized, drowning most of those aboard. Why?
Truth: The boat was moving along a river when a large snake dropped onto the deck. The passengers all rushed to the other side of the boat, thereby overturning it.
Key Clues: 1. A large snake dropped onto the deck. 2. Passengers rushed to the other side of the boat in panic. 3. The boat capsized due to the sudden shift in weight.

GPT-4:
Host: {Puzzle} {Rules Introduction} Please begin your questions.
Player: Question 1: Was the boat overloaded with passengers?
Host: Response 1: No
Player: Question 2: Was there a storm or rough weather that caused the boat to capsize?
Host: Response 2: No
Player: Question 3: Did the boat capsize due to a collision with another object or vessel?
Host: Response 3: No
Player: Question 4: Was there a mechanical or structural failure in the boat that caused it to capsize?
Host: Response 4: No
Player: Question 5: Did the passengers suddenly move to one side of the boat causing it to capsize?
Host: Response 5: Yes
Player: Answer: The riverboat capsized because all or most passengers suddenly moved to one side of the boat, causing it to lose balance and capsize.

GPT-3.5:
Host: {Puzzle} {Rules Introduction} Please begin your questions.
Player: Question 1: Did the riverboat capsize due to a natural disaster such as a storm or earthquake?
Host: Response 1: No
...
Player: Question 7: Did something unexpected occur during the voyage which caused the riverboat to capsize?
Host: Response 7: Yes
Player: Question 8: Was there an animal involved in the incident?
Host: Response 8: Yes
Player: Question 9: Did the animal attack the passengers and cause the riverboat to capsize?
Host: Response 9: No
Player: Question 10: Did the animal accidentally cause the riverboat to capsize?
Host: Response 10: Yes
Player: Answer: An animal accidentally caused the riverboat to capsize.

Llama2-chat-70B:
Host: {Puzzle} {Rules Introduction} Please begin your questions.
Player: Sure, I'd be happy to play a game of Lateral Thinking Puzzle with you! Here's my first question: Question 1: Is the reason for the riverboat capsizing related to the boat itself?
Host: Response 1: No
Player: Great, that helps to narrow it down. Here's my next question: Question 2: Is the reason for the riverboat capsizing related to the passengers?
Host: Response 2: Yes
Player: Great, that helps to narrow it down even further. Here's my next question: Question 3: Did the passengers do something to cause the boat to capsize?
Host: Response 3: Yes
Player: Great, now I think I know the answer. Here's my answer: Answer: The passengers stood up and rushed to one side of the boat, causing it to capsize. Is that correct?

Baichuan-chat-13B:
Host: {Puzzle} {Rules Introduction} Please begin your questions.
Player: Question 1: What happened?
Host: Response 1: Irrelevant
Player: Question 2: How many people were there in the boat?
Host: Response 2: Irrelevant
Player: Question 3: Were any life jackets worn?
Host: Response 3: Irrelevant
Player: Question 4: Was anyone drunk while operating the ship?
Host: Response 4: No
Player: Question 5: Did someone accidentally hit something beneath the water without realizing it?
Host: Response 5: No
Player: Answer: The reason why the passenger steamboat overturned was because some object fell into the hull below the surface and cracked the bottom of the vessel, causing rapid flooding and capsizing.

Table 5: The interaction case between the player and the host in LatEval. We present GPT-4, GPT-3.5, Llama2-chat-70B, and Baichuan-chat-13B as the respective player models in the case.

Table 4 demonstrates a strong correlation between the human evaluation and the GPT-4 automated evaluation. Additionally, the relative rankings among different models are generally consistent, indicating the rationality of AC and QR metric and the validity of employing GPT-4 for the automated evaluation. Furthermore, GPT-4 achieves an accuracy rate of over 90% as a host responding to questions, underscoring its competence as an effective host for Lateral Thinking Puzzles. Meanwhile, GPT-3.5 achieves an accuracy rate of over 80% as a host responding to questions, lower than GPT-4. Although the responses from GPT-3.5 are generally correct, there are some instances of misleading the players. Note that when faced with players like Bloomchat and Baichuan-chat-13B, which lack lateral thinking ability, the host model tends to make more errors in responses.

4.5. Fine-grained Analysis

Most models still lack proficient Lateral thinking, making it challenging to directly accomplish Lateral Thinking Puzzle. To further refine model evaluations, we assess the model's lateral thinking ability at varying levels of difficulty by introducing settings with different numbers of clues.

Specifically, in the initial phase of interaction, we present the puzzle alongside either 50% or 100% of the clues, comparing their performance to a setting with 0% clues. The previous introduced evaluations shown in Table 2 and Table 3 are conducted under 0% clues setting. In this graded evaluation, we calculate ROUGE-L (Lin, 2004) between deduction of the player model and the truth, utilizing English data and GPT-4 as the host model.

Figure 3 shows the performance of different models under various difficulty settings. We observe that for models such as GPT-4 and GPT-3.5, as the amount of provided clues increases, the deduction becomes closer to the truth. This trend reflects the strong information integration capability of these models. However, for Baichuan-chat-13B, even when provided with all clues, meaning the complete information, it still cannot deduce the truth from the provided information. The performance under 100% clues setting represents the upper limit of each model's performance in Lateral Thinking Puzzle. It indicates that Baichuan-chat-13B lacks the ability to comprehend complicated instructions and integrate the available information.

Furthermore, with the increase in the number of provided clues, models that are skilled at integrating information exhibit the decrease in Average Turns. It's because the more clues you provide, the less you need to ask for information. Therefore, for powerful models, more clues settings result in fewer Average Turns.

4.6. Case Study

Table 5 shows the case of different player models tackling Lateral Thinking Puzzle, including the puzzle, truth, key clues and interactions between host model and player model. The case demonstrates that there is a substantial gap between the puzzle and the truth, including a few pieces of key information. In order to get the truth, the player needs to think divergently, explore various possibilities, and strategically structure its questions to acquire crucial information. It's evident that this task is quite challenging. During the interaction with the host, player model GPT-4 shows divergent thinking by posing some thought-provoking questions, such as "Did the passengers suddenly move to one side of the boat causing it to capsize?". Furthermore, player GPT-4 effectively covers a wide range of aspects by addressing only five questions, including weather-related factors, accidental causes, human-related factors. GPT-4 also references some key clues in the final answer, shows its proficiency in divergent thinking and information integration.

GPT-3.5 also poses some divergent and valuable questions, such as "Did the animal accidentally cause the riverboat to capsize?". In the answer, GPT-3.5 covers "animal" which is acquired through questioning during the interaction. Llama2-chat-70B tends to ask questions with a narrow focus, and the question format does not strictly adhere to the requirements specified in the prompt, incorporating some unnecessary discourse. Although the answer of Llama2-chat-70B does not confirm to the format requirements, it covers "rushed to one side of the boat", which is acquired through questioning. Baichuan-chat-13B disregards the prompt's repeated emphasis on asking only yes-or-no questions and, without obtaining sufficient information, starts to fabricate the answer. The above results reveal differences in lateral thinking abilities among various player models. Furthermore, there is still substantial room for improvement in the lateral thinking abilities of existing LLMs.

5. Related Work

LLMs Evaluation. With the emergence of ChatGPT and GPT-4 (OpenAI, 2023), the capabilities of LLMs have become increasingly impressive, drawing significant attention to evaluation of LLMs (Yu et al., 2023; Li et al., 2023c; Ye et al., 2023; Li et al., 2024a). Recently, many research efforts focuses on evaluating LLMs from various perspectives, including language tasks (Cheng et al., 2023c; Huang et al., 2022; Li et al., 2022; Ma et al., 2022; Bang et al., 2023), reasoning (Bang et al., 2023; Bian et al., 2023), robustness (Li et al., 2023a; Cheng et al., 2023d), trustworthiness (Hagendorff and Fabi, 2023; Cheng et al., 2023b), medical applications (Chervenak et al., 2023; Cascella et al., 2023), and ethical considerations (Cao et al., 2023; Parrish et al., 2021; Liu et al., 2022). Currently, mainstream evaluation benchmarks, such as MMLU (Hendrycks et al., 2020)and C-Eval (Huang et al., 2023), assess LLMs from a multi-task perspective, yet are still constrained within a question-answering framework. However, these benchmarks lack evaluation criteria for measuring LLMs' lateral thinking capabilities.

LLMs Reasoning. In terms of reasoning, some researchers try to guide LLMs' reasoning through Prompt Engineering frameworks like CoT (Wei et al., 2022), ToT (Yao et al., 2024), GoT (Besta et al., 2023), AoT (Sel et al., 2023) and so on. The Chain of Thoughts (CoT) sequences LLM decision-making, while the Tree of Thoughts (ToT) simplifies problems into sub-tasks for structured reasoning. The Graph of Thoughts (GoT) maps decisions onto a Directed Acyclic Graph for complex problem-solving, and the Algorithm of Thoughts (AoT) enables LLMs to follow algorithmic reasoning paths for in-context learning. Additionally, researchers further optimize the capabilities of LLMs Reasoning, for instance, by introducing Chain-of-Thought Self-Consistency (Wang et al., 2022), Skeleton-of-Thought (Ning et al., 2023), and Program-of-Thoughts (Chen et al., 2022).

Lateral Thinking. Syahrin et al. (2019) propose lateral thinking as a cognitive activity employed to construct creative ideas. Similarly, Hidayat et al. (2018) indicate that lateral thinking is associated with generating novel ideas, which serves as progressive elements across various scientific domains, from engineering to art, from politics to personal well-being. de Bono (1999) points out that vertical thinking and lateral thinking are mutually complementary, with vertical thinking being selective, and lateral thinking being creative. Lateral thinking offers alternative perspectives to vertical thinking, thereby augmenting its efficacy (Relaiza et al., 2021).

6. Conclusion

In this paper, we propose a novel interactive benchmark to evaluate the lateral thinking capability of LLMs, following the setup of a popular game named Lateral Thinking Puzzle. We collect data from various Lateral Thinking Puzzle websites and construct an evaluation dataset. Experimental results indicate that most LLMs perform poorly on lateral thinking capability, which also demonstrates that our benchmark is challenging and has the capacity to distinguish LLMs' performance. In the future, we will further investigate how to evaluate the lateral thinking ability of foundation models rather than chat models. Additionally, there is immense potential for promising research in improving the lateral thinking capability of LLMs.

7. Ethical Considerations

In this paper, we propose a novel evaluation benchmark named LatEval for assessing the lateral thinking capability of LLMs. All the original samples are collected from publicly accessible and legitimate websites, and do not contain any sensitive information. We check the licenses of the samples to ensure its usability for non-profit academic research. During data processing, we remove all bloody, terrifying and extreme stories, ensuring that each story in our benchmark is harmless and ethical. Thanks to the significant improvement in data annotation efficiency brought by LLMs, part of data annotation and verification is conducted by the authors ourselves. To mitigate the authors' biases, we hire 3 college students for data annotation during the dataset quality verification and human evaluation stages. The average hourly salary paid to them is approximately twice the local minimum salary.

8. Acknowledgments

This research is supported by National Natural Science Foundation of China (Grant No.62276154), Research Center for Computer Network (Shenzhen) Ministry of Education, the Natural Science Foundation of Guangdong Province (Grant No. 2023A1515012914), Basic Research Fund of Shenzhen City (Grant No. JCYJ20210324120012033 and JSGG20210802154402007), Shenzhen Science and Technology Program (Grant No. WDZC20231128091437002), the Major Key Project of PCL for Experiments and Applications (PCL2021A06), and Overseas Cooperation Research Fund of Tsinghua Shenzhen International Graduate School (HW2021008).

9. Bibliographical References

Shabana Ali. 2019. Lateral thinking puzzles encourage creative thinking. *Nurse educator*, 44(4):191.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, et al. 2023. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, and Ben He. 2023. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *arXiv preprint arXiv:2303.16421*.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *ArXiv*, abs/2303.17466.

Marco Cascella, Jonathan Montomoli, Valentina Bellini, and Elena Bignami. 2023. Evaluating the feasibility of chatgpt in healthcare: an analysis of multiple clinical and research scenarios. *Journal of Medical Systems*, 47(1):33.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.

Xuxin Cheng, Bowen Cao, Qichen Ye, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. 2023a. Ml-lmcl: Mutual learning and large-margin contrastive learning for improving asr robustness in spoken language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6492–6505.

Xuxin Cheng, Qianqian Dong, Fengpeng Yue, Tom Ko, Mingxuan Wang, and Yuexian Zou. 2023b. M 3 st: Mix at three levels for speech translation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Xuxin Cheng, Zhihong Zhu, Bowen Cao, Qichen Ye, and Yuexian Zou. 2023c. Mrrl: Modifying the reference via reinforcement learning for non-autoregressive joint multiple intent detection and slot filling. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10495–10505.

Xuxin Cheng, Zhihong Zhu, Wanshi Xu, Yaowei Li, Hongxiang Li, and Yuexian Zou. 2023d. Accelerating multiple intent detection and slot filling via targeted knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8900–8910.

Joseph Chervenak, Harry Lieman, Miranda Blanco-Breindel, and Sangita Jindal. 2023. The promise and peril of using a large language model to obtain clinical information: Chatgpt performs strongly as a fertility counseling tool with limitations. *Fertility and Sterility*.

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*.

Edward de Bono. 1999. Lateral thinking. *Current Biology*, 9.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165:113679.

Thilo Hagendorff and Sarah Fabi. 2023. Human-like intuitive behavior and reasoning biases emerged in language models–and disappeared in gpt-4. *arXiv preprint arXiv:2306.07622*.

Tianyong Hao, Xinxin Li, Yulan He, Fu Lee Wang, and Yingying Qu. 2022. Recent progress in leveraging deep learning methods for question answering. *Neural Computing and Applications*, pages 1–19.

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*.

James S Hernandez and MBBS Prathibha Varkey. 2008. Vertical versus lateral thinking. *Physician executive*, 34(3):26.

Taufiq Hidayat, Endang Susilaningsih, and Kurniawan dan Cepi. 2018. The effectiveness of enrichment test instruments design to measure students' creative thinking skills and problem-solving. *Thinking Skills and Creativity*.

Shulin Huang, Shirong Ma, Yinghui Li, Li Yangning, Shiyang Lin, Haitao Zheng, and Ying Shen. 2022. Towards attribute-entangled controllable text generation: A pilot study of blessing generation. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 235–247, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng

Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Xinzhe Li, Ming Liu, Shang Gao, and Wray Buntine. 2023a. A survey on out-of-distribution evaluation of neural nlp models. *arXiv preprint arXiv:2306.15261*.

Yangning Li, Tingwei Lu, Yinghui Li, Tianyu Yu, Shulin Huang, Hai-Tao Zheng, Rui Zhang, and Jun Yuan. 2023b. MESED: A multi-modal entity set expansion dataset with fine-grained semantic classes and hard negative entities. *CoRR*, abs/2307.14878.

Yangning Li, Shirong Ma, Xiaobin Wang, Shen Huang, Chengyue Jiang, Hai-Tao Zheng, Pengjun Xie, Fei Huang, and Yong Jiang. 2023c. Ecomgpt: Instruction-tuning large language models with chain-of-task tasks for e-commerce. *CoRR*, abs/2308.06966.

Yinghui Li, Haojing Huang, Shirong Ma, Yong Jiang, Yangning Li, Feng Zhou, Hai-Tao Zheng, and Qingyu Zhou. 2023d. On the (in)effectiveness of large language models for chinese text correction. *CoRR*, abs/2307.09007.

Yinghui Li, Shang Qin, Jingheng Ye, Shirong Ma, Yangning Li, Libo Qin, Xuming Hu, Wenhao Jiang, Hai-Tao Zheng, and Philip S Yu. 2024a. Rethinking the roles of large language models in chinese grammatical error correction. *arXiv preprint arXiv:2402.11420*.

Yinghui Li, Zishan Xu, Shaoshen Chen, Haojing Huang, Yangning Li, Yong Jiang, Zhongli Li, Qingyu Zhou, Hai-Tao Zheng, and Ying Shen. 2023e. Towards real-world writing assistance: A chinese character checking benchmark with faked and misspelled characters. *CoRR*, abs/2311.11268.

Yinghui Li, Qingyu Zhou, Yangning Li, Zhongli Li, Ruiyang Liu, Rongyi Sun, Zizhen Wang, Chao Li, Yunbo Cao, and Hai-Tao Zheng. 2022. The past mistake is the future wisdom: Error-driven contrastive probability optimization for chinese spell checking. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3202–3213. Association for Computational Linguistics.

Yinghui Li, Qingyu Zhou, Yuanzhen Luo, Shirong Ma, Yangning Li, Hai-Tao Zheng, Xuming Hu, and Philip S. Yu. 2024b. When llms meet cunning questions: A fallacy understanding benchmark for large language models. *CoRR*, abs/2402.11100.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ruiyang Liu, Yinghui Li, Linmi Tao, Dun Liang, and Hai-Tao Zheng. 2022. Are we ready for a new paradigm shift? A survey on visual deep MLP. *Patterns*, 3(7):100520.

Shirong Ma, Yinghui Li, Rongyi Sun, Qingyu Zhou, Shulin Huang, Ding Zhang, Li Yangning, Ruiyang Liu, Zhongli Li, Yunbo Cao, Haitao Zheng, and Ying Shen. 2022. Linguistic rules-based corpus generation for native Chinese grammatical error correction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 576–589, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xuefei Ning, Zinan Lin, Zixuan Zhou, Huazhong Yang, and Yu Wang. 2023. Skeleton-of-thought: Large language models can do parallel decoding. *arXiv preprint arXiv:2307.15337*.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Sam Bowman. 2021. Bbq: A hand-built bias benchmark for question answering. In *Findings*.

Héctor Raúl Santa María Relaiza, Doris Fuster-Guillén, Yolvi Ocaña-Fernández, Patricia Edith Guillén Aparicio, and Freddy Antonio Ochoa Tataje. 2021. Cognitive processes and creative lateral thinking in students of the naval school of peru. *Neuroquantology*, 19:10–17.

Sandra W Russ. 1988. Primary process thinking, divergent thinking, and coping in children. *Journal of Personality Assessment*, 52(3):539–548.

Sandra Walker Russ. 2013. *Affect and creativity: The role of affect and play in the creative process*. Routledge.

Together Computer SambaNova Systems. 2023. Bloomchat: a new open multilingual chat llm.

Bilgehan Sel, Ahmad Al-Tawaha, Vanshaj Khattar, Lu Wang, Ruoxi Jia, and Ming Jin. 2023. Algorithm of thoughts: Enhancing exploration of ideas in large language models. *arXiv preprint arXiv:2308.10379*.

Paul Sloane and Des MacHale. 1994. *Great lateral thinking puzzles*. Sterling Publishing Company, Inc.

A. S. Syahrin, Dawud, Heri Suwignyo, and Endah Tri Priyatni. 2019. Creative thinking patterns in student's scientific works. *Eurasian Journal of Educational Research*.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities. https://github.com/InternLM/InternLM.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Kuan Chen Tsai. 2012. Play, imagination, and creativity: A brief literature review. *Journal of Education and learning*, 1(2):15–20.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

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.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Jingheng Ye, Yinghui Li, Qingyu Zhou, Yangning Li, Shirong Ma, Hai-Tao Zheng, and Ying Shen. 2023. CLEME: debiasing multi-reference evaluation for grammatical error correction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6174–6189. Association for Computational Linguistics.

Tianyu Yu, Chengyue Jiang, Chao Lou, Shen Huang, Xiaobin Wang, Wei Liu, Jiong Cai, Yangning Li, Yinghui Li, Kewei Tu, Hai-Tao Zheng, Ningyu Zhang, Pengjun Xie, Fei Huang, and Yong Jiang. 2023. Seqgpt: An out-of-the-box large language model for open domain sequence understanding. *CoRR*, abs/2308.10529.