

C²RBench: A Chinese Complex Reasoning Benchmark for Large Language Models

Junru Wu*, Tianhao Shen*, Linxi Su, Deyi Xiong[†]

College of Intelligence and Computing, Tianjin University, Tianjin, China
{wujunru, thshen, sulinxi, dyxiong}@tju.edu.cn

Abstract

Large language models (LLMs) have achieved remarkable progress in autonomous reasoning, evolving from basic text processing to sophisticated multimodal reasoning, a critical capability for general-purpose AI assistants. However, existing benchmarks usually fail to adequately capture the intricate multi-step reasoning demands inherent in real-world scenarios. To bridge this gap, we propose **C²RBench**: a **C**hinese **C**omplex **R**easoning **B**enchmark for evaluating multi-step, multimodal advanced reasoning capability of LLMs. C²RBench comprises 1,115 carefully curated Chinese tasks, which are organized into eight domain-specific subsets, each meticulously designed to mirror real-world challenges. This hierarchical benchmark features three difficulty tiers based on the number of reasoning steps required (average 8.44 steps per task), significantly exceeding existing benchmarks in cognitive complexity. Extensive evaluations of 20 LLMs (including DeepSeek-R1) and 24 multimodal large language models (MLLMs) on C²RBench reveal critical performance gaps: GPT-4.1¹ achieves only 52.11% accuracy, indicating substantial room for improvement. The dataset and evaluation code² are publicly available.

1 Introduction

Recent advances in reasoning capabilities of large language models (LLMs), such as o1-preview³ and DeepSeek-R1 (DeepSeek-AI et al., 2025) have significantly enhanced their ability to engage in more extended and deliberate reasoning processes, allowing them to solve increasingly complex problems compared to earlier LLMs (Yao et al., 2023). These reasoning capabilities allow models to perform multi-step reasoning, make better decisions,

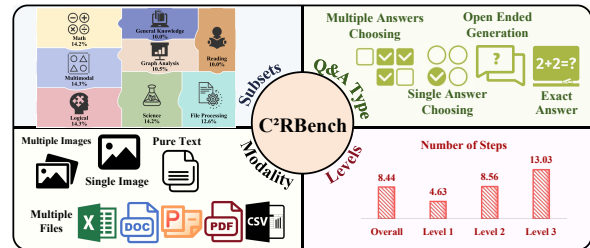


Figure 1: Illustration of the features of C²RBench.

and adapt more effectively to diverse tasks (Chen et al., 2024; Wang et al., 2024).

Reasoning has become a critical factor in evaluating the intelligence and generalization of LLMs. However, existing reasoning benchmarks often fall short of capturing the requirements of modern reasoning language models (Liu et al., 2025; Huang and Chang, 2023). First, many current benchmarks often focus on a limited set of reasoning tasks with a small number of reasoning steps. For instance, GSM8K (Cobbe et al., 2021), a purely textual benchmark, tests basic mathematical operations requiring only 2 to 8 reasoning steps. Second, benchmarks like DRCD (Shao et al., 2019) and C-Eval (Huang et al., 2023) are not well aligned with real-world scenarios and fail to capture the complexity and contextual ambiguity of authentic reasoning problems. Third, most reasoning benchmarks are limited to English and unimodal tasks, lacking support for other languages like Chinese or multimodal formats that combine text, images, tables, or structured files (Li et al., 2024; Huang et al., 2024b).

To mitigate these challenges, we propose the **Chinese Complex Reasoning Benchmark (C²RBench)**, a novel benchmark designed to evaluate LLM reasoning across multiple dimensions, including multi-task, multi-step, and multimodal reasoning in Chinese. As illustrated in Figure 1, it features broader and more reasoning steps than most existing benchmarks. Specifically, C²RBench

*Equal contribution

[†]Corresponding author

¹<https://openai.com/index/gpt-4-1/>

²<https://github.com/tjunlp-lab/C2RBench>

³<https://openai.com/o1/>

emphasizes multi-step reasoning, with an average of 8.44 reasoning steps per task. The tasks are organized into three levels of reasoning complexity, with level-3 tasks requiring an average of 13.03 reasoning steps. Additionally, C²RBench supports multimodal reasoning, encompassing text, images, tables, and files. The tasks in C²RBench are drawn from real-world scenarios, such as file processing, graph analysis, and scientific reasoning, in order to simulate the evaluation of LLM reasoning in practical applications. Although similar to the GAIA benchmark (Mialon et al., 2023), C²RBench is larger (1,115 vs. 466 tasks) and more aligned with real-world scenarios, with a focus on Chinese rather than English.

The main contributions of C²RBench are summarized as follows:

1. We present C²RBench, the first large-scale Chinese benchmark that jointly evaluates multi-task, multi-step, and multimodal reasoning, addressing critical gaps in current LLM evaluation.
2. C²RBench benefits from its high cognitive complexity and strong alignment with real-world reasoning tasks. It consists of 1,115 tasks across 8 subsets and 3 complexity levels, with an average of 4.63 reasoning steps per level-1 task, 8.56 per level-2 task and 13.03 per level-3 task.
3. We conduct extensive evaluations of 20 LLMs and 24 multimodal large language models (MLLMs) on C²RBench. Even state-of-the-art models such as GPT-4.1 achieve only 52.11% accuracy, compared to 89.75% for humans, illustrating the formidable challenge posed by C²RBench to advanced LLMs.

2 Related Work

As LLMs continue to advance, their ability to tackle complex reasoning tasks has become a key focus in both research and application (Guo et al., 2023; Wei et al., 2023; DeepSeek-AI et al., 2025). Existing benchmarks for evaluating reasoning capabilities can be broadly categorized into *task-specific* and *comprehensive* benchmarks.

Task Specific Reasoning Benchmarks. These benchmarks are tailored to assess a model’s reasoning ability in a particular domain. For instance, HelLaSwag (Zellers et al., 2019) and SIQA (Sap et al., 2019) evaluate commonsense reasoning; MathVista

(Lu et al., 2024), GSM8K (Cobbe et al., 2021) and BackMATH (Zhang and Xiong, 2025) focus on mathematical reasoning; ARC (Clark et al., 2018) targets scientific reasoning; CodeXGLUE (Lu et al., 2021) and ProBench (Yang et al., 2025) benchmark code reasoning and generation tasks; ToolBench (Qin et al., 2023) and CToolEval (Guo et al., 2024) evaluate the tool using capability of models; and RoleEval (Shen et al., 2023) aims to evaluate reasoning over knowledge required for role-playing. While these datasets provide valuable insights into specific reasoning capabilities, they fall short in evaluating general-purpose reasoning across diverse settings.

Comprehensive Reasoning Benchmarks. To systematically evaluate broader reasoning abilities, several comprehensive benchmarks have been introduced. MMMU (Yue et al., 2024a,b) spans multiple domains—ranging from science to humanities—using primarily multiple-choice questions. Similarly, C-Eval (Huang et al., 2023), M3KE (Liu et al., 2023a) and OpenEval (Liu et al., 2024) provide large-scale Chinese evaluation sets emphasizing advanced knowledge and reasoning. Other multimodal benchmarks such as MMBench (Liu et al., 2023b) and GaoKao-MM (Zong and Qiu, 2024) focus on visual reasoning in multiple-choice formats. GAIA (Mialon et al., 2023) offers open-ended tasks covering web search, code generation, multimodal understanding, and document processing.

Despite their contributions, existing benchmarks exhibit several limitations: (i) a lack of real-world contextual complexity, (ii) limited support for multi-step reasoning annotations, and (iii) insufficient coverage of the Chinese language and multimodal reasoning settings (Li et al., 2024; Huang et al., 2023). These constraints hinder their applicability for evaluating the capabilities of modern LLMs in practical, multilingual, and cognitively demanding scenarios.

To address these challenges, we propose C²RBench, a Chinese benchmark designed for multi-task, multi-modal and multi-step reasoning scenarios in real-world applications. Detailed comparison of our dataset with previous datasets is presented in Table 1.

3 Dataset Curation

We follow the pipeline illustrated in Figure 2 to curate C²RBench, aiming to evaluate the multi-step reasoning abilities of LLMs in real-world Chinese

Datasets	MS Anno?	Size	Lang	Domain	Task Type
MathVista	no	5.14K	en	Math	Open&MC
MMMU	a few	12K	en	Real-world	Open&MC
GAIA	yes	0.5K	en	Real-world	Open
GaoKao-MM	no	0.6K	zh	Exams	MC
C-Eval	no	14K	zh	Exams, Web	MC
DRCDC	no	30K	zh	Reading	Open
C ² RBench	yes	1.1K	zh	Real-world	Open&MC

Table 1: Comparison of C²RBench with existing reasoning benchmarks. The column “MS Anno?” indicates whether the dataset provides multi-step reasoning annotations. MC denotes Multi-choice QA, while open refers to open-ended QA.

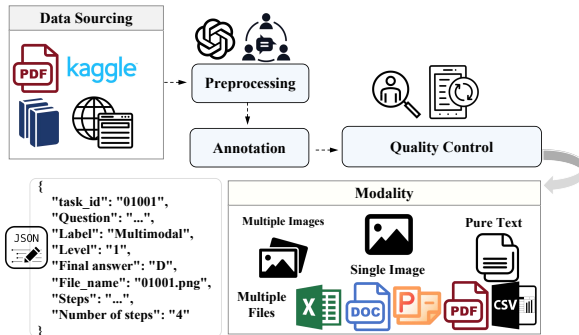


Figure 2: The dataset curation pipeline of C²RBench.

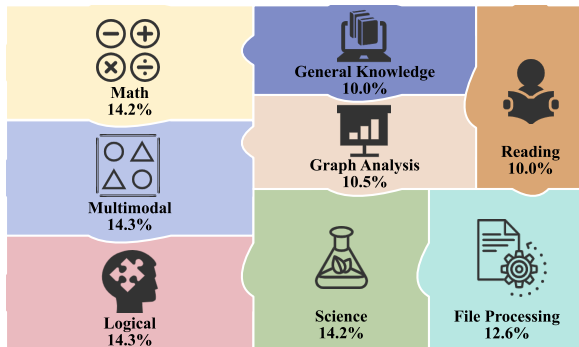


Figure 3: Proportion of different subsets in C²RBench.

scenarios.

3.1 Data Sources and Preprocessing

To ensure both authenticity and diversity, we collect samples from four main sources: (1) real-world materials such as civil service exams and classical literature; (2) authoritative databases including those from the China National Bureau of Statistics; (3) educational resources such as secondary and tertiary-level academic exam questions; and (4) structured datasets from open platforms like Kaggle. Additionally, we incorporate high-quality GPT-generated samples verified by human annotators (see §A.1), forming a hybrid corpus of authentic

and synthetic data.

All collected data undergo rigorous preprocessing, adhering to three principles: format normalization with semantic preservation, consistent annotation practices via standardized workflows, and comprehensive annotator training. We further implement a multi-stage quality control mechanism involving progress audits, random sample reviews, and error correction cycles to maintain annotation consistency and data integrity.

3.2 Annotation

Each task is designed to meet the following criteria:

- **Deterministic Answers:** Questions must yield unique, unambiguous answers with concise solutions.
- **Multi-Step Reasoning:** All tasks require at least three reasoning steps, with clearly delineated intermediate inferences.
- **Non-Retrievability:** Problems must not be solvable via direct keyword search, ensuring genuine reasoning is required.
- **Synthetic Data Validation:** GPT-generated samples are subject to rigorous manual validation, including checks for semantic alignment and logical consistency.
- **Contamination Prevention:** We confirm that all questions cannot be answered via simple web queries to avoid data leakage.

To enforce these standards, we implement a dual-layer verification framework comprising automated rule-based checks and expert manual review. This ensures that all instances meet the necessary complexity thresholds and remain grounded in real-world reasoning scenarios.

3.3 Quality Control

Post-annotation, we apply a rigorous cross-verification protocol to ensure logical coherence, answer correctness, and supporting evidence relevance. Quality assurance is conducted in two iterative phases:

- **Deduplication:** Near-duplicate samples are identified using TF-IDF with cosine similarity and then manually validated before removal.
- **Dynamic Difficulty Calibration:** Question difficulty is dynamically adjusted based on solver feedback. Tasks exceeding 11 reasoning steps are automatically escalated to Level 3 difficulty, followed by human verification.

This comprehensive quality control framework guarantees both difficulty progression and content diversity while preserving pedagogical validity.

3.4 Task Construction Manual

For detailed information on how tasks are constructed, annotated, and verified, please refer to the Task Construction Manual in the Appendix A.2.

3.5 Statistics

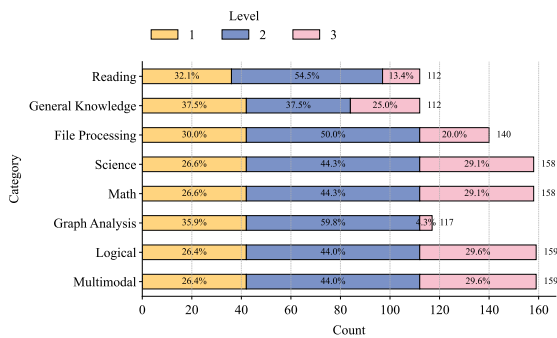


Figure 4: Proportion of different difficulty levels in each subset.

C²RBench contains 1,115 manually curated instances across eight subsets: Math, Multimodal, Logical, General Knowledge, Graph Analysis, Reading, Science, and File Processing. Figure 3 illustrates their distribution.

To enable fine-grained analysis, we categorize tasks into three difficulty levels based on the number of annotated reasoning steps:

- **Level 1 (3–6 steps):** Basic reasoning, testing fundamental capabilities of LLMs.

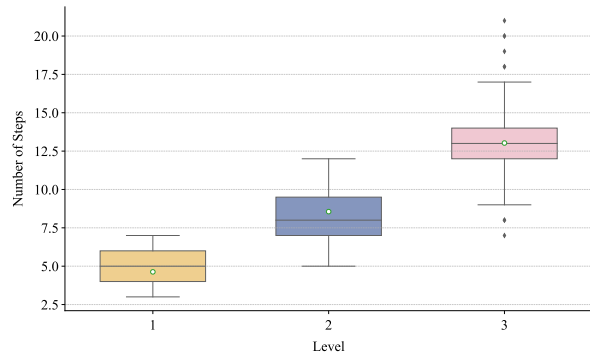


Figure 5: Distribution of annotated reasoning steps in C²RBench.

- **Level 2 (7–11 steps):** Intermediate reasoning, probing multi-hop inference capacity.
- **Level 3 (≥ 12 steps):** Advanced reasoning, evaluating long-horizon inference and information synthesis.

The overall distribution of difficulty levels and reasoning steps is shown in Figure 4 and 5, providing a comprehensive overview of C²RBench structure and complexity. For more details, please refer to Appendix A.4.

4 Experiments

We conducted extensive experiments with 20 LLMs and 24 MLLMs on C²RBench to evaluate its effectiveness and utility for assessing the reasoning capabilities of LLMs and MLLMs. All models used for evaluation are summarized in Table 2.

4.1 Evaluation Settings

We evaluate a diverse set of models on C²RBench, encompassing both LLMs and MLLMs, including closed-source and open-source variants. All evaluations are conducted in a zero-shot setting to measure the models’ intrinsic reasoning capabilities without fine-tuning or few-shot demonstrations.

Accuracy (ACC) is adopted as the primary evaluation metric. Given the dataset’s mixture of open-ended and multiple-choice questions, rule-based matching proves insufficient for reliable evaluation. To address this, we use **GPT-4o** as an automatic evaluator. To verify its reliability, we perform human evaluation involving three trained annotators. In each of the 10 rounds, 100 model responses are randomly sampled and compared. GPT-4o achieves over 98% agreement with human judgments, validating its effectiveness as an evaluation proxy. The evaluation pipeline is illustrated in Figure 6.

Models	Open Source?	Model Size	Multimodal?	Reasoning?
DeepSeek-R1 (DeepSeek-AI et al., 2025)	Yes	671B	No	Yes
DeepSeek-V2.5-1210 (DeepSeek-AI, 2024)	Yes	236B	No	No
DeepSeek-V3 (DeepSeek-AI et al., 2024)	Yes	671B	No	No
Doubao-1.5-vision-pro-32k ⁴	No	undisclosed	Yes	No
Doubao-1.5-thinking-pro-250415 ⁵	No	undisclosed	Yes	Yes
Doubao-pro-32k ⁴	No	undisclosed	No	No
Doubao-vision-pro-32k ⁴	No	undisclosed	Yes	No
Ernie-4.0-Turbo-8k ⁶	No	undisclosed	No	No
Ernie-x1-turbo-32k ⁷	No	undisclosed	No	Yes
Gemini 1.5 Flash ⁸	No	undisclosed	Yes	No
Gemini 1.5 Pro ⁸	No	undisclosed	Yes	No
GLM-4-Long (GLM et al., 2024)	No	undisclosed	Yes	No
GLM-4v-plus ⁹	No	undisclosed	Yes	No
GLM-Z1-AirX ¹⁰	No	undisclosed	No	Yes
GLM-Zero-preview ⁹	No	undisclosed	No	Yes
GPT-3.5-Turbo ¹¹	No	undisclosed	No	No
GPT-4o ¹¹	No	undisclosed	Yes	No
GPT-4o-mini ¹¹	No	undisclosed	Yes	Yes
GPT-4-Turbo ¹¹	No	undisclosed	No	No
GPT-4.5-Preview ¹¹	No	undisclosed	Yes	No
GPT-4.1 ¹	No	undisclosed	Yes	No
o1-mini ¹¹	No	undisclosed	No	Yes
o1-preview ³	No	undisclosed	No	Yes
Pixtral-12B (Agrawal et al., 2024)	Yes	12B	Yes	Yes
QwQ-32B-Preview ¹²	Yes	32B	No	Yes
hunyuan-turbo-latest ¹³	No	undisclosed	No	No
hunyuan-turbo-vision ¹³	No	undisclosed	Yes	No
Kimi-thinking-preview ¹⁴	No	undisclosed	Yes	Yes
Llama-3.2-Vision (Patterson et al., 2022)	Yes	11B, 72B	Yes	No
LLaVA-OneVision (Patterson et al., 2022)	Yes	7B, 72B	Yes	No
Moonshot-v1-32k ¹⁴	No	undisclosed	No	No
Moonshot-v1-32k-vision-preview ¹⁴	No	undisclosed	Yes	No
Qwen2-VL ¹⁵	Yes	2B, 7B, 72B	Yes	No
Qwen3-235B-A22B ¹⁶	Yes	235B	No	Yes
Qwen-VL-max (Bai et al., 2023)	No	undisclosed	Yes	No
SenseChat-5-1202 ¹⁷	No	undisclosed	No	No
Spark4.0 Ultra ¹⁸	No	undisclosed	No	No
SparkX1 ¹⁸	No	undisclosed	No	Yes
Yi-Lightning (Wake et al., 2025)	No	undisclosed	No	No
Yi-Vision-V2 ¹⁹	No	undisclosed	Yes	No

Table 2: Models evaluated on the C²RBench.

All experiments are conducted on machines equipped with eight NVIDIA A6000 GPUs (48GB memory each). Wherever possible, we use the official APIs released by the model developers for consistency and fairness.

To ensure reproducibility and control randomness, we standardize the decoding parameters as follows: we set temperature = 0 whenever supported. For models requiring a non-zero temperature (e.g., LLaMA), we use temperature = 0.001 and disable sampling by setting do_sample = False. Note that o1-preview only supports temperature = 1, and DeepSeek-R1 does not respond to temperature adjustments.

⁴<https://www.volcengine.com/product/doubao>
⁵<https://www.volcengine.com/docs/82379/1536428>
⁶<https://ai.baidu.com/ai-doc/WENXINWORKSHOP/am3ih7x209a>
⁷<https://cloud.baidu.com/doc/WENXINWORKSHOP/s/hm90w209a>
⁸<https://deepmind.google/technologies/gemini/>
⁹<https://www.zhipuai.cn>
¹⁰<https://open.bigmodel.cn/dev/api/Reasoning-models/glm-z1>
¹¹<https://openai.com>
¹²<https://huggingface.co/QwQ/QwQ-32B-Preview>
¹³<https://hunyuan.tencent.com>
¹⁴<https://platform.moonshot.cn>
¹⁵<https://qwenlm.github.io/blog/qwen2-vl/>
¹⁶<https://huggingface.co/Qwen/Qwen3-235B-A22B>
¹⁷<https://platform.sensnova.cn>
¹⁸<https://www.xfyun.cn>
¹⁹<https://platform.lingyiwanwu.com>

Category	Average Accuracy (%)
Overall	89.75
File Processing	100.00
General Knowledge	90.90
Graph Analysis	100.00
Logical	93.80
Math	100.00
Multimodal	56.20
Reading	97.00
Science	86.80

Table 3: Human performance across different subsets..

4.2 Tested Models

The 44 tested models are summarized in Table 2.

MLLMs. We selected 24 MLLMs, including both open-source and closed-source models. Among these, two are specifically optimized for reasoning. For each model family, we selected the most recent and high-performing version available at the time of evaluation. Although the official documentation for Llama-3.2-Vision lists support for only eight languages (excluding Chinese), our experiments show that it can handle Chinese reasonably well, albeit with occasional code-switching.

Text-only LLMs. We included 20 text-only LLMs, 9 of which are explicitly designed for reasoning, such as DeepSeek-R1 and o1-preview.

4.3 Human Performance Evaluation

To benchmark human performance on C²RBench, we conducted a stratified sampling of 10% of the questions, ensuring balanced coverage across all subsets and difficulty levels. Eleven participants, all with at least undergraduate-level education, were recruited through targeted selection to ensure relevant domain expertise. Each question was independently answered by 2–3 individuals.

Human participants achieved an overall accuracy of 89.75%, significantly outperforming current LLMs. Table 3 presents accuracy by category. Additional details on the sampled questions and participant backgrounds are provided in Appendix A.7.

4.4 Results

Overall Performance: Human participants significantly outperformed all evaluated models, achieving an overall accuracy of **89.75%**. In contrast, the best-performing models, **Doubao-1.5-thinking-pro** and **GPT-4.1**, only reached **65.74%**

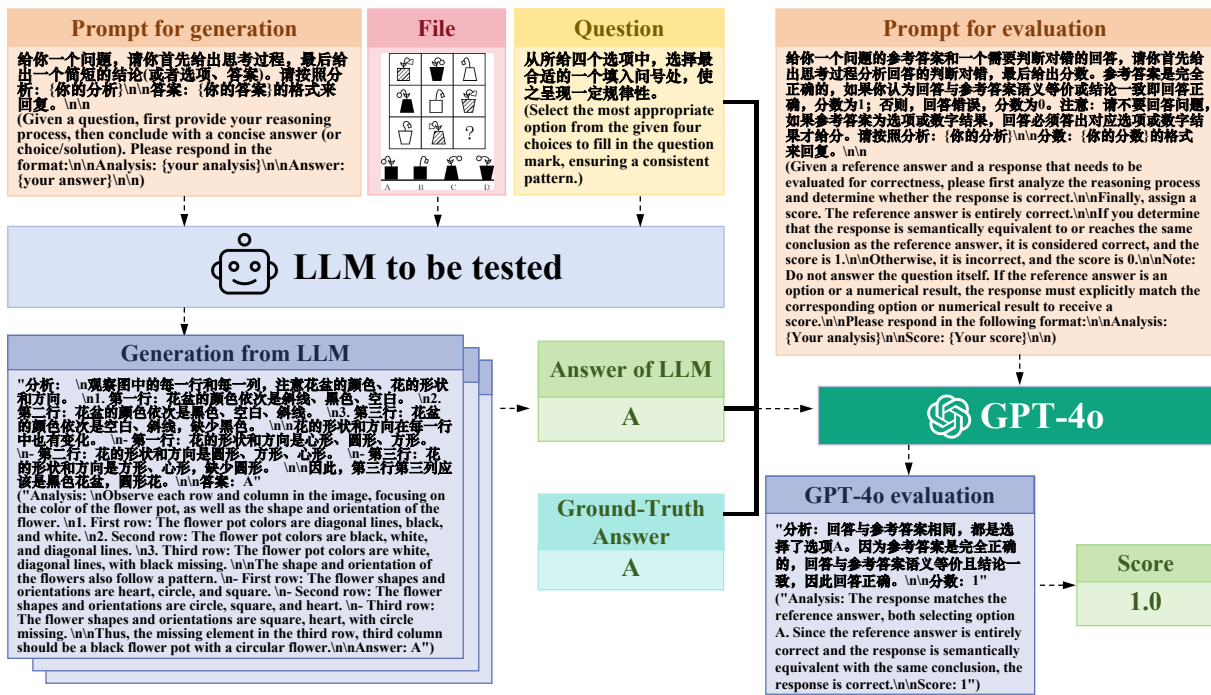


Figure 6: The pipeline for task evaluation in C²RBench is illustrated using an example. LLMs represent all the models to be evaluated (see Table 2). The input to the LLM to be tested consists of the prompt for generation, file, and question. The answer of the LLM is derived by extracting its generation. GPT-4o is used as the evaluation model. The input to GPT-4o includes the prompt for evaluation, the answer of the LLM, and the ground truth answer. Finally, the score for the evaluated reasoning task is derived by extracting GPT-4o evaluation results.

and 52.11%, respectively (see Figure 7 and 8). This substantial performance gap underscores the current limitations of LLMs in approximating general-purpose human reasoning capabilities.

Disparity between Open-source and Closed-source Models: The best-performing open-source MLLM (e.g., **Qwen2-VL-72B-Instruct**) achieves an overall accuracy of only 37.76%, which lags far behind leading closed-source models such as **GPT-4.1**. Despite this gap, open-source models demonstrate strong potential on specific reasoning tasks. Notably, **Deepseek-R1** attains 76.10% accuracy on the logical reasoning subset, outperforming closed-source models like **o1-preview** in this category.

Model Performance across Different Subsets: Figure 7 shows the performance of MLLMs across the eight subsets of C²RBench. Current MLLMs exhibit notable weaknesses in Math reasoning and Multimodal reasoning. Additionally, as shown in Figure 10, text-only LLMs tend to perform poorly on tasks in the General Knowledge subset of C²RBench, further indicating a limitation in handling certain types of reasoning tasks.

Model Performance across Different Levels: Figure 8 and 10 illustrate that as task complexity increases, the accuracy of all models tends to decrease. Taking GPT-4.1 as an example, its accuracy drops from 62.73% at Level 1 to 54.30% at Level 2, and further decreases to 34.35% at Level 3. This trend is consistent across all models, highlighting the difficulties large language models encounter when dealing with more complex reasoning tasks.

MLLMs Performance on Multimodal Reasoning: Figure 9 presents the performance of MLLMs on the multimodal subset. Open-source models, such as Qwen2-VL-7B-Instruct, attain an accuracy of 25.79%, slightly outperforming closed-source models like GPT-4o, which achieves an accuracy of 25.16%. Despite this, the overall performance on the multimodal subset remains low, revealing substantial room for improvement in both multimodal and multi-step reasoning tasks.

Insights of Logical Reasoning: Even the state-of-the-art models, such as o1-preview and DeepSeek-R1, achieve an accuracy of only 72.33% and 76.10% (see Figure 10). These results underscore the significant challenges posed by C²RBench in advanced logical reasoning tasks.

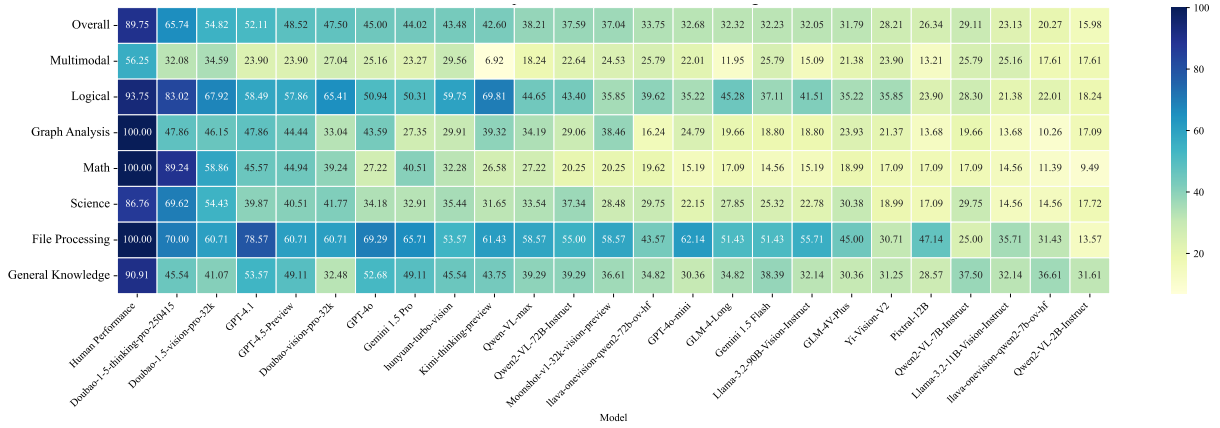


Figure 7: MLLMs' performance across different subsets.

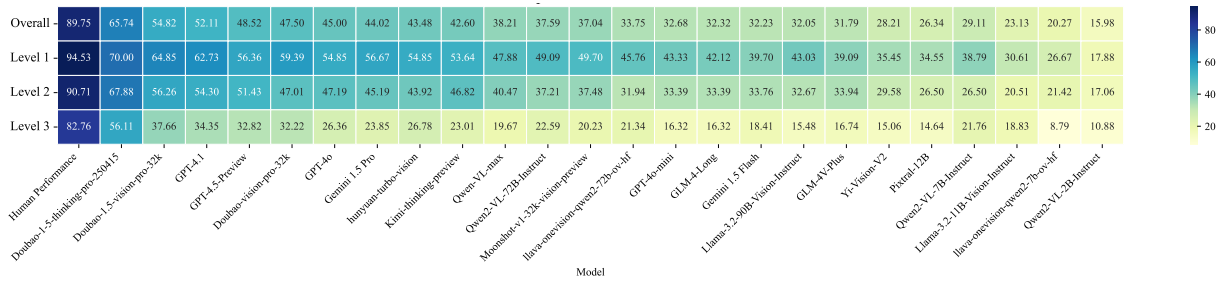


Figure 8: MLLMs' performance across different levels.

Further investigation of DeepSeek-R1's performance via web interface testing resulted in a 74.48% accuracy on the logical reasoning subset. A detailed analysis of inference times (Table 4) reveals a positive correlation between reasoning time and task difficulty. Specifically, incorrect responses consistently require longer processing times (mean = 285.83s) compared to correct solutions (mean = 100.69s), in line with previous findings (Huang et al., 2024a). This pattern suggests that persistent reasoning loops contribute to inefficiencies in LLMs, emphasizing the importance of addressing unproductive computational cycles in future model development.

Additionally, an in-depth error analysis and case study (see Appendix A.5) were conducted to further explore the underlying factors affecting the model's logical reasoning performance.

4.5 Error Analysis

To systematically assess the limitations in model performance, we conducted an extensive error typology analysis across four state-of-the-art LLMs.

The responses from each model were categorized into six distinct error classes through a multi-stage annotation process. (1) **Perception & Input**: Failures in processing multimodal inputs. (2)

Times (s)	All	Correct	Wrong
Overall	147.26	100.69	285.83
Level 1	83.57	63.88	167.25
Level 2	132.49	91.98	281.00
Level 3	226.19	158.37	345.88

Table 4: Average reasoning time per task for DeepSeek-R1 on the logical reasoning subset.

Knowledge & Comprehension: Gaps in domain-specific knowledge. (3) **Logic & Reasoning**: Breakdowns in multi-step reasoning chains. (4) **Task Execution**: Procedural errors in problem-solving. (5) **Refusal & Inability**: Unwarranted non-responses. (6) **Generation & Output**: Errors in formatting or linguistic generation. Note: Individual problems may exhibit multiple error types simultaneously. Full definitions and prompt for error analysis are provided in Appendix A.3.

Table 5 highlights two critical needs for LLMs in Chinese reasoning: improved cross-modal understanding to resolve perceptual-input mismatches, and more efforts dedicated to multi-step reasoning. The persistent prevalence of logical errors (exceeding 65% across all LLMs) particularly underscores the urgency of addressing these challenges.

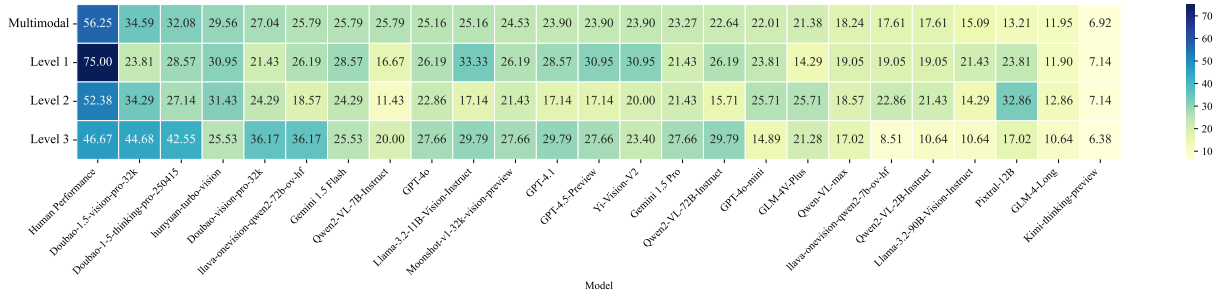


Figure 9: Results of MLLMs on the multimodal reasoning subset.

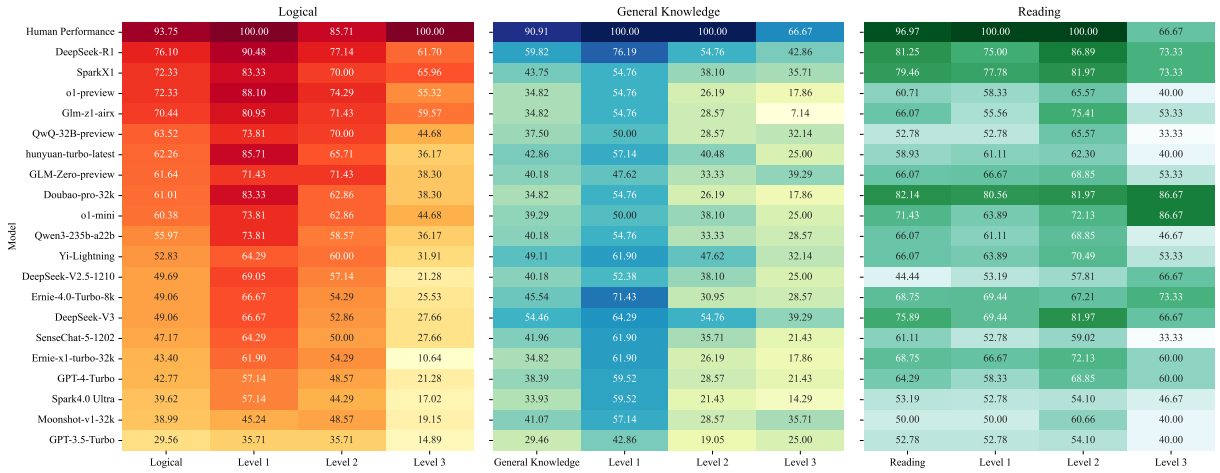


Figure 10: Results of text-only LLMs on the Logical & Graph Analysis & Reading reasoning subset.

Error Type	Deepseek-R1	o1-preview	GPT-4o	Qwen2-VL-72B
Perception & Input	0.00%	0.00%	4.42%	3.17%
Knowledge & Comprehension	13.16%	18.92%	52.21%	52.88%
Logic & Reasoning	97.37%	97.73%	69.39%	65.56%
Task Execution	10.53%	0.00%	9.82%	13.40%
Refusal & Inability	0.00%	0.00%	1.96%	7.06%
Generation & Output	0.00%	2.27%	1.31%	2.74%

Table 5: Error type distribution across models. Please note that one output of LLM may involve multiple types of errors (e.g., both Knowledge & Comprehension and Logic & Reasoning), resulting in column totals exceeding 100%. It is important to note that DeepSeek-R1 and o1-preview were evaluated exclusively on the logical subset, while GPT-4o and Qwen2-VL-72B were assessed on the full C²Rbench dataset.

4.6 Case Study

To further analyze the strengths and limitations of different MLLMs, we present two representative case studies.

Figure 21 illustrates a multimodal reasoning case that requires precise alignment between visual content and textual understanding. From the model outputs, it is evident that both Qwen2-VL-72B and LLaMA-3.2-90B-Vision struggle with fine-grained visual recognition and spatial reasoning, leading to incorrect conclusions. These errors suggest that de-

spite their overall capabilities, current MLLMs still face challenges in extracting detailed visual cues and integrating them effectively into the reasoning process.

Figure 22 presents a logical reasoning scenario that highlights differences in reasoning chains among the evaluated models. These variations reflect the disparity in multi-step reasoning abilities and indicate that some models rely more heavily on surface-level patterns rather than deep inference.

More comprehensive discussions and additional examples can be found in Section A.5, where we further dissect model behaviors across a range of challenging scenarios.

4.7 Analysis of Reasoning Models

We analyzed token usage across tasks to evaluate multi-step reasoning efficiency, comparing **Doubao-1.5-Thinking-Pro** and **Kimi-Thinking-Preview** with direct-answer model **GPT-4o**, aiming to identify *overthinking*—excessive reasoning with little gain.

For Doubao and Kimi, we computed the following token categories: **Reasoning Tokens**, **Answer Tokens**, and **Total Tokens**, while for GPT-4o we

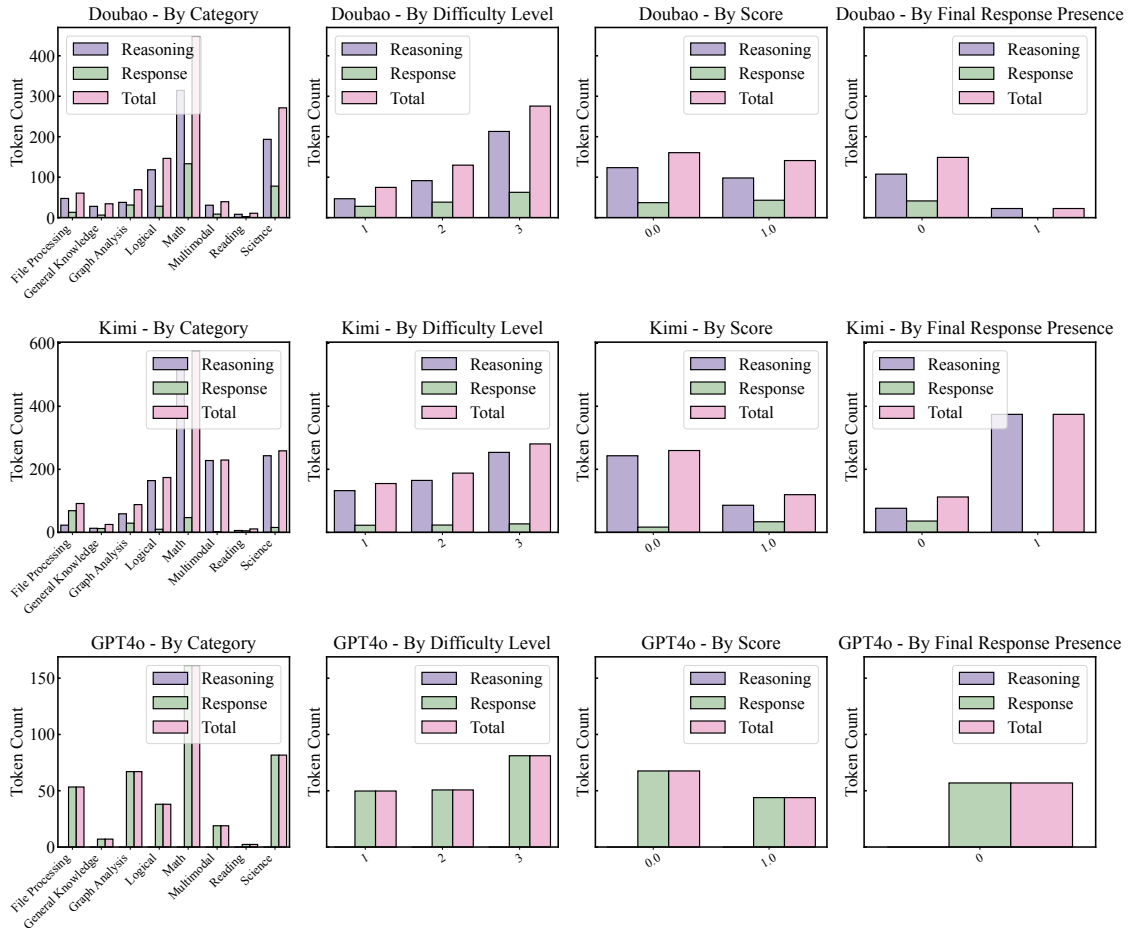


Figure 11: Overthinking analysis across task category, level, score, and final answer presence. Bars represent token counts for reasoning, answer, and total. In the “Final Response Presence” plots, **1 indicates absence of a final answer**.

only counted the final response tokens.

Figure 11 illustrates a common phenomenon in reasoning models: while token usage increases in more complex tasks, higher reasoning token consumption does not necessarily improve accuracy and may even result in answer absence due to token exhaustion.

The phenomenon of overthinking becomes evident with increasing complexity in tasks. Both Doubao and Kimi show excessive reasoning steps without significant improvements in accuracy, often leading to token exhaustion before reaching a conclusion. In contrast, GPT-4o’s direct approach, despite potentially using fewer reasoning tokens, achieves more consistent results.

To address these issues, future advancements should focus on adaptive control mechanisms for reasoning depth, budget-aware planning for token usage, trace-aware evaluations to balance efficiency with accuracy, and dead-end detection to avoid unproductive reasoning loops. For a more detailed

analysis of token consumption, see Appendix A.6.

5 Conclusion

In this paper, we have presented C²RBench, a novel benchmark designed to evaluate the multi-step reasoning capabilities of LLMs in Chinese. C²RBench consists of 1,115 tasks organized into eight distinct subsets. These tasks are categorized into three difficulty levels based on the number of steps involved. The key highlights of C²RBench are its real-world scenarios, extensive task coverage, and its focus on Chinese. Our core contribution lies in providing a comprehensive and challenging benchmark for assessing LLMs’ performance on these complex, multi-step tasks in Chinese. Experimental evaluations on 24 MLLMs and 20 text-only LLMs indicate that C²RBench remains challenging for state-of-the-art models. By systematically assessing complex reasoning across eight subsets, C²RBench establishes a rigorous and comprehensive benchmark.

Limitation

Although C²RBench is comprehensive, like any benchmark, it has its limitations. The manual curation process may introduce biases, and its focus on high-difficulty, multi-step reasoning problems may not fully capture the capabilities required for LLMs. However, we believe strong benchmark performance is essential for evaluating advanced LLMs. The reasoning challenges posed by C²RBench are evident in the performance of 22 MLLMs and 20 text-only LLMs. To balance complexity and practicality, the dataset includes multiple-choice, single-choice, calculation-based tasks, and concise open-ended questions.

Acknowledgments

The present research was supported by the National Key Research and Development Program of China (Grant No. 2024YFE0203000). We would like to thank the anonymous reviewers for their insightful comments.

References

- Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. 2024. *Pixtral 12B*. *Preprint*, arXiv:2410.07073.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. *Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond*. *arXiv preprint arXiv:2308.12966*.
- Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024. *M³CoT: A Novel Benchmark for Multi-Domain Multi-step Multi-modal Chain-of-Thought*. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8199–8221, Bangkok, Thailand. Association for Computational Linguistics.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. *Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge*. *ArXiv*, abs/1803.05457.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. *Training Verifiers to Solve Math Word Problems*. *Preprint*, arXiv:2110.14168.
- DeepSeek-AI. 2024. *DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model*. *Preprint*, arXiv:2405.04434.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song,

- Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. [DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning](#). *Preprint*, arXiv:2501.12948.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yudian Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2024. [DeepSeek-V3 Technical Report](#). *Preprint*, arXiv:2412.19437.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. [ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools](#). *Preprint*, arXiv:2406.12793.
- Zishan Guo, Yufei Huang, and Deyi Xiong. 2024. [CToolEval: A Chinese benchmark for LLM-powered agent evaluation in real-world API interactions](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15711–15724, Bangkok, Thailand. Association for Computational Linguistics.
- Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiakuan Li, Bojian Xiong, Deyi Xiong, et al. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*.
- Jie Huang and Kevin Chen-Chuan Chang. 2023. [Towards Reasoning in Large Language Models: A Survey](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.
- Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024a. [Large Language Models Cannot Self-Correct Reasoning Yet](#). In *The Twelfth International Conference on Learning Representations*.
- Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. [C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models](#). In *Advances in Neural Information Processing Systems*.
- Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, Yikai Zhang, Yuqing Yang, Ting Wu, Binjie Wang, Shichao Sun, Yang Xiao, Yiyuan Li, Fan Zhou, Steffi Chern, Yiwei Qin, Yan Ma, Jiadi Su, Yixiu Liu, Yuxiang Zheng, Shaoting Zhang, Dahua Lin, Yu Qiao, and Pengfei Liu. 2024b. [OlympicArena: Benchmarking Multi-discipline Cognitive Reasoning for Superintelligent AI](#). *arXiv preprint arXiv:2406.12753*.
- Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, Ying Tai, Wankou Yang, Yabiao Wang, and Chengjie Wang. 2024. [A Survey on Benchmarks of Multimodal Large Language Models](#). *Preprint*, arXiv:2408.08632.

- Chuang Liu, Renren Jin, Yuqi Ren, Linhao Yu, Tianyu Dong, Xiaohan Peng, Shuting Zhang, Jianxiang Peng, Peiyi Zhang, Qingqing Lyu, et al. 2023a. M3ke: A massive multi-level multi-subject knowledge evaluation benchmark for chinese large language models. *arXiv preprint arXiv:2305.10263*.
- Chuang Liu, Linhao Yu, Jiakuan Li, Renren Jin, Yufei Huang, Ling Shi, Junhui Zhang, Xinmeng Ji, Tingting Cui, Liutao Liutao, et al. 2024. Openeval: Benchmarking chinese llms across capability, alignment and safety. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 190–210.
- Hanmeng Liu, Zhizhang Fu, Mengru Ding, Ruoxi Ning, Chaoli Zhang, Xiaozhang Liu, and Yue Zhang. 2025. [Logical Reasoning in Large Language Models: A Survey](#). *Preprint*, arXiv:2502.09100.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023b. MMBench: Is Your Multi-modal Model an All-around Player? *arXiv:2307.06281*.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. In *International Conference on Learning Representations (ICLR)*.
- Shuai Lu, Daya Guo, Shuo Ren, Junjie Huang, Alexey Svyatkovskiy, Ambrosio Blanco, Colin Clement, Dawn Drain, Daxin Jiang, Duyu Tang, et al. 2021. Codexglue: A machine learning benchmark dataset for code understanding and generation. *arXiv preprint arXiv:2102.04664*.
- Grégoire Mialon, Clémentine Fourier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. [GAIA: a benchmark for General AI Assistants](#). *Preprint*, arXiv:2311.12983.
- David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluís-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2022. [The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink](#). *Preprint*, arXiv:2204.05149.
- Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. 2023. [Toollm: Facilitating large language models to master 16000+ real-world apis](#). *arXiv preprint arXiv:2307.16789*.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense Reasoning about Social Interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473.
- Chih Chieh Shao, Trois Liu, Yuting Lai, Yiyang Tseng, and Sam Tsai. 2019. [DRCD: a Chinese Machine Reading Comprehension Dataset](#). *Preprint*, arXiv:1806.00920.
- Tianhao Shen, Sun Li, Quan Tu, and Deyi Xiong. 2023. Roleeval: A bilingual role evaluation benchmark for large language models. *arXiv preprint arXiv:2312.16132*.
- Alan Wake, Bei Chen, C. X. Lv, Chao Li, Chengen Huang, Chenglin Cai, Chujie Zheng, Daniel Cooper, Fan Zhou, Feng Hu, Ge Zhang, Guoyin Wang, Heng Ji, Howard Qiu, Jiangcheng Zhu, Jun Tian, Katherine Su, Lihuan Zhang, Liying Li, Ming Song, Mou Li, Peng Liu, Qicheng Hu, Shawn Wang, Shijun Zhou, Shiming Yang, Shiyong Li, Tianhang Zhu, Wen Xie, Wenhao Huang, Xiang He, Xiaobo Chen, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Yanpeng Li, Yongke Zhao, Yongzhen Luo, Yuchi Xu, Yuxuan Sha, Zhaodong Yan, Zhiyuan Liu, Zirui Zhang, and Zonghong Dai. 2025. [Yi-Lightning Technical Report](#). *Preprint*, arXiv:2412.01253.
- Yulong Wang, Tianhao Shen, Lifeng Liu, and Jian Xie. 2024. Sibyl: Simple yet effective agent framework for complex real-world reasoning. *arXiv preprint arXiv:2407.10718*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-Thought Prompting Elicits Reasoning in Large Language Models](#). *Preprint*, arXiv:2201.11903.
- Lei Yang, Renren Jin, Ling Shi, Jianxiang Peng, Yue Chen, and Deyi Xiong. 2025. Probench: Benchmarking large language models in competitive programming. *arXiv preprint arXiv:2502.20868*.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. [Tree of Thoughts: Deliberate Problem Solving with Large Language Models](#). *Preprint*, arXiv:2305.10601.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhao Chen. 2024a. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of CVPR*.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhao Chen, and Graham Neubig. 2024b. MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark. *arXiv preprint arXiv:2409.02813*.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a

Machine Really Finish Your Sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Shaowei Zhang and Deyi Xiong. 2025. **BackMATH: Towards backward reasoning for solving math problems step by step**. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 466–482, Abu Dhabi, UAE. Association for Computational Linguistics.

Yi Zong and Xipeng Qiu. 2024. **GAOKAO-MM: A Chinese Human-Level Benchmark for Multimodal Models Evaluation**. *Preprint*, arXiv:2402.15745.

A Appendix

A.1 Prompts for Dataset Curation

In Figure 12, we present a complete prompt example required to construct a Graph Analysis task using ChatGPT. The generated content is manually reviewed each time to ensure its relevance and usability.

A.2 Task Construction Manual

Each instance in C²RBench follows a standardized schema with the following fields:

- **task_id**: Unique identifier (category code + serial number).
- **Question**: Problem statement in Chinese.
- **Label**: Task category (Math, Multimodal, Logical, General Knowledge, Graph Analysis, Reading, Science, File Processing).
- **Level**: Difficulty level (1–3), determined by the number of reasoning steps.
- **Final answer**: A concise final answer.
- **File_name**: Optional multimodal inputs (e.g., images, PDFs, Excel files).
- **Annotator Metadata**: Includes the full reasoning chain (Steps) and number of steps.

A.2.1 Category Composition

The dataset includes 140 instances in each of the eight task categories, with each category designed to pose distinct domain-specific reasoning challenges. The categories are: **Multimodal**, **Logical**, **Graph Analysis**, **Math**, **Science**, **File**, **General Knowledge**, and **Reading**.

Each category is designed to address specific reasoning challenges:

- **Multimodal**: Includes graphical pattern completion and audio-based reasoning tasks.
- **Logical**: Abstract logic puzzles and reasoning tasks.
- **Graph Analysis**: Reasoning over related visual plots or statistical charts.
- **Math**: Symbolic and natural language math word problems (LaTeX encoded).
- **Science**: Physics, chemistry, and biology problems grounded in educational material.

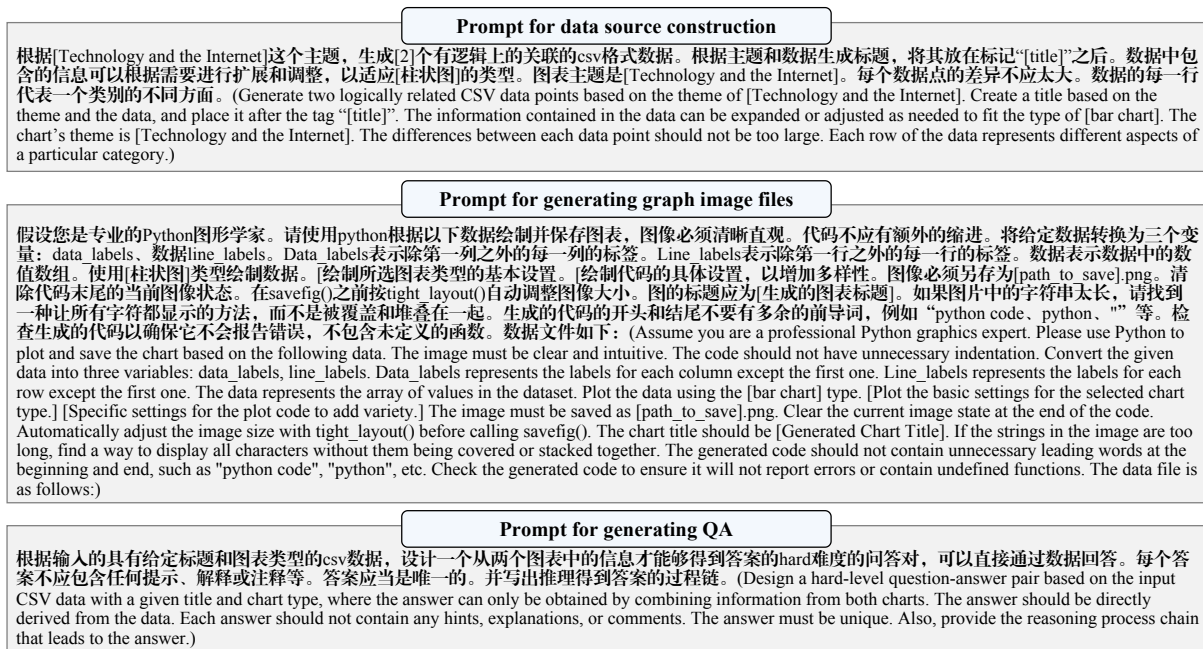


Figure 12: Example prompts for constructing a Graph Analysis task using ChatGPT.

- **File Processing:** Multi-file document understanding (e.g., cross-referencing PDF and Excel).
- **General Knowledge:** Tasks for general knowledge reasoning.
- **Reading:** Reasoning tasks based on long Chinese passages.

A.2.2 Task Construction Guidelines

The construction of each task strictly follows the guidelines below:

- **Reasoning Chain:** Each question must include necessary reasoning steps. Each reasoning step must be logically required, concise, and context-dependent.
- **Non-Redundancy:** Each step of the task must be non-redundant, ensuring the reasoning process is minimal and effective.
- **Non-Retrievability:** Tasks are designed to prevent solving via search engine queries, avoiding shortcut solutions.
- **Multi-Step Reasoning:** All tasks must involve at least three reasoning steps (average reasoning steps: 8.44 steps, with up to 12.03 steps for Level 3 tasks).

A.2.3 Quality Assurance and Annotator Training

All annotators underwent detailed training using a handbook, and regular reviews were conducted. The training included:

- Ensuring task format consistency, annotation uniformity, and data quality.
- Regular inter-annotator agreement checks to ensure annotation quality.
- Secondary review for all Level 3 tasks.
- Verification of attachment integrity and accessibility.

A total of 373 annotation hours were logged, with an average time of 20 minutes per instance. Additionally, random sampling was used for quality checks to ensure data reliability.

A.3 Prompt for Error Analysis

The GPT-4o-based error analysis prompt is as follows:

You will be provided with: 1. A reference answer that follows standard reasoning steps. 2. A model-generated response containing errors.

Your task is to analyze the errors, attribute them to their root causes, and classify them into an appropriate error type and subtype based on the following categories:

1. **Perception & Input Errors** -
 - Perceptual Error: Misinterprets visual or structured input (e.g., tables, charts).
 - Input Parsing Error: Mishandles input format, structure, or symbols (e.g., JSON, XML).
2. **Knowledge & Comprehension Errors** -
 - Lack of Knowledge: Lacks necessary domain knowledge.
 - Factual Error: Provides incorrect or hallucinated facts.
 - Misinterpretation of Question Intent: Misunderstands task requirements.
3. **Logical & Reasoning Errors** -
 - Logical Reasoning Error: Produces inconsistent or invalid reasoning.
 - Inductive Reasoning Error: Incorrectly generalizes from specific cases.
 - Deductive Reasoning Error: Misapplies general rules to specific instances.
 - Long-range Dependency Error: Fails to maintain key context across reasoning steps.
 - Hypothesis Generation Error: Assumes unsupported premises.
 - Causal Reasoning Error: Misidentifies cause-effect relationships.
 - Analytical Reasoning Error: Struggles with multi-step analysis (e.g., mathematical proofs).
4. **Task Execution Errors** -
 - Arithmetic Error: Makes numerical miscalculations.
 - Constraint Violation Error: Ignores explicit constraints (e.g., word limits, formatting).
 - Planning & Omission Error: Skips necessary reasoning steps or lacks structural coherence.
5. **Refusal & Inability Errors** -
 - Overly Conservative Refusal: Unjustified refusal despite sufficient information.
 - Failure to Answer: States inability to respond when reasoning is possible.
 - Model Breakdown: Produces irrelevant,

incoherent, or nonsensical output.

6. **Generation & Output Errors** -
 - Grammar & Syntax Error: Contains grammatical or syntactical mistakes.
 - Disorganized Response: Presents information in a confusing manner.
 - Hallucination Error: Generates non-existent or fabricated content.
 - Redundancy & Repetition Error: Repeats content excessively.
 - Ambiguous Response: Provides vague or unclear answers.
- Response Format:**
- Analysis: {Detailed analysis of the errors, including their root causes}
 - Error Type: {Main error category}
 - Error Subtype: {Specific error subtype }

A.4 Dataset Details

We divide the dataset into eight subsets. Below, we introduce the data source and specific content of each subset and provide one example image for illustration.

A.4.1 Multimodal

This subset comes from graphical reasoning questions in civil service exams. It mainly tests the model's ability to handle and reason over multimodal information. An example is shown in Figure 13.



Figure 13: An example from the Multimodal subset.

A.4.2 Logical

This subset includes logical reasoning questions from major entrance exams and is used to evaluate the model's reasoning and problem-solving abilities. See Figure 14.

```

"task_id": "02159",
"Question": "在一天的24小时之中，时针、分针、秒针完全重合在一起有几次，分别是什么时间？",
"Label": "Logical",
"Level": "3",
"Final answer": "2次，12点和0点，(Twice, 12 o'clock and 0 o'clock)",
"File_name": "02159.png",
"Steps": "1. 问题描述：在一天的24小时之中，时针、分针、秒针完全重合在一起有几次，分别是什么时间？\n2. 确定角速度：\n设时针的角速度为  $\omega_1$ ，分针的角速度为  $\omega_2$ ，秒针的角速度为  $\omega_3$ 。则有  $\omega_1 = \frac{2\pi}{12}$ ， $\omega_2 = \frac{2\pi}{1}$ ， $\omega_3 = \frac{2\pi}{60}$ 。分针和秒针重合的条件： $\omega_2 t = \omega_3 t + 2k\pi$ ， $k \in \mathbb{Z}$ 。解得  $t = \frac{60k}{59}$ 。时针和分针重合的条件： $\omega_1 t = \omega_2 t + 2m\pi$ ， $m \in \mathbb{Z}$ 。解得  $t = \frac{12m}{11}$ 。综合以上两个条件，得到  $t = \frac{60k}{59} = \frac{12m}{11}$ 。整理得  $11k = 59m$ 。由于11和59互质，所以  $k = 59n$ ， $m = 11n$ ， $n \in \mathbb{Z}$ 。代入  $t = \frac{60k}{59}$ ，得到  $t = 12n$ 。即在24小时内，时针、分针、秒针完全重合只有两次：12点和0点。",
"Number of steps": "12"

```

Figure 14: An example from the Logical subset.

A.4.3 Graph Analysis

This subset is sourced from national statistics, government data, and academic databases. It includes a variety of chart types (e.g., bar charts, line charts, pie charts) and covers multiple domains such as business, education, environment, and transportation. See Figure 15.



Figure 15: An example from the Graph Analysis subset.

A.4.4 Math

This subset includes math problems from Chinese high school and university entrance exams, covering algebraic reasoning, natural language math problems, and spatial geometry. An example is shown in Figure 16.

A.4.5 Science

This subset includes tasks sourced from publicly available Chinese textbook materials. It covers a wide range of disciplines, including biology, physics, chemistry, engineering, and economics. See Figure 17.

A.4.6 File Processing

This subset comes from structured data provided by governments, international organizations, research institutions, and Kaggle competitions. It involves

```

"task_id": "04012",
"Question": "已知函数  $f(x) = e^x - a x^2 + 2$ ,  $g(x) = x^2 + \cos x$ . 当  $a = 1$  时, 判断函数  $f(x)$  的单调性.",
"Label": "Math",
"Level": "1",
"Final answer": "单调递增",
"File_name": "04012.png",
"Steps": "1. 当  $a = 1$  时,  $f(x) = e^x - x^2 + 2$ , 所以  $f'(x) = e^x - 2x$ . 令  $m(x) = e^x - 2x$ , 所以  $m'(x) = e^x - 2$ . 当  $x < \ln 2$  时,  $m'(x) < 0$ ; 当  $x > \ln 2$  时,  $m'(x) > 0$ , 所以  $m(x) \geq m(\ln 2) = 2 - 2\ln 2 > 0$ . 因此,  $f'(x) > 0$ . 从而函数  $f(x)$  在  $(-\infty, +\infty)$  上单调递增.",
"Number of steps": "6"

```

Figure 16: An example from the Math subset.



Figure 17: An example from the Science subset.

demographics, economics, health data, etc. An example is shown in Figure 18.

A.4.7 General Knowledge

This subset is designed to test multi-step reasoning ability beyond math and science. Some of the content was co-annotated with the help of ChatGPT. See Figure 19.

A.4.8 Reading Comprehension

This subset contains data from over 20 books, including detective novels (e.g., Sherlock Holmes), thematic novels, and academic papers, to evaluate reading comprehension and reasoning skills. See Figure 20.

A.5 Case Study

Figure 21 presents a multimodal case from C²R Bench, showcasing responses from four models: two high-performing closed-source models,


```

"task_id": "06138",
"Question": "2020年一季度, 奥迪品牌在美国市场最畅销的车型, 和在韩国市场最畅销的车型, 谁的销售增长速度更快? ",
("In the first quarter of 2020, which model of Audi brand had a faster sales growth rate: the best-selling model in the U.S. market or the best-selling model in the South Korean market?")
"Label": "File Processing",
"Level": "3",
"Final answer": "在韩国市场最畅销的车型Audi A4", ("The best-selling model in the Korean market, Audi A4.")
"File_name": "06138a.xlsx, 06138b.xlsx",
"Steps": "1. 打开文件06138a.xlsx, 06138b.xlsx, 分别检查文件中的表格和数据结构。 \n2. 在06138a.xlsx文件中, 定位到2020年1月、2020年2月和2020年3月, 韩国市场奥迪品牌各车型的销量数据。 \n3. 按照车型合并并计算2020年1月、2020年2月和2020年3月的销量数据, 得到奥迪品牌各车型2020年一季度在韩国市场的销量数据。 \n4. 对上述数据进行排序, 确定2020年一季度奥迪品牌在韩国市场最畅销的车型为Audi A4 (1,094 units)。 \n5. 在06138b.xlsx文件中, 定位到2020年1月、2020年2月和2020年3月, 美国市场奥迪品牌各车型的销量数据。 \n6. 按照车型合并并计算2020年1月、2020年2月和2020年3月的销量数据, 得到奥迪品牌各车型2020年一季度在美国市场的销量数据。 \n7. 对上述数据进行排序, 确定2020年一季度奥迪品牌在美国市场最畅销的车型为Audi Q5 (9,959 units)。 \n8. 在06138a.xlsx文件中, 定位到2019年1月、2019年2月和2019年3月, Audi A4在韩国市场的销量数据分别为188辆、113辆、163辆。 \n9. 合并计算上述数据, 得到Audi A4车型2019年一季度在韩国市场的销量为64辆。 \n10. 计算Audi A4的同比增长率: 1094/64-1=135.78%。 \n11. 在06138b.xlsx文件中, 定位到2019年1月、2019年2月和2019年3月, Audi Q5在美国市场的销量数据分别为4541辆、4888辆、5302辆。 \n12. 合并计算上述数据, 得到Audi Q5车型2019年一季度在美国市场的销量为14731辆。 \n13. 计算Audi A4的同比增长率: 9959/14731-1=-32.29%。 \n14. 得出结论: 2020年一季度, 奥迪品牌在韩国市场最畅销的车型Audi A4的销售增长速度更快。
("1. Open the files 06138a.xlsx and 06138b.xlsx, and inspect the tables and data structures in each file. \n2. In the 06138a.xlsx file, locate the sales data for Audi models in the Korean market for January 2020, February 2020, and March 2020. \n3. Merge and calculate the sales data for Audi models in January 2020, February 2020, and March 2020, to obtain the sales data for Audi models in the Korean market during the first quarter of 2020. \n4. Sort the above data to determine that the best-selling Audi model in the Korean market during the first quarter of 2020 was the Audi A4 (1,094 units). \n5. In the 06138b.xlsx file, locate the sales data for Audi models in the U.S. market during the first quarter of 2020. \n6. Merge and calculate the sales data for Audi models in January 2020, February 2020, and March 2020, to obtain the sales data for Audi models in the U.S. market during the first quarter of 2020. \n7. Sort the above data to determine that the best-selling Audi model in the U.S. market during the first quarter of 2020 was the Audi Q5 (9,959 units). \n8. In the 06138a.xlsx file, locate the sales data for the Audi A4 in the Korean market for January 2019, February 2019, and March 2019, which were 188 units, 113 units, and 163 units, respectively. \n9. Merge and calculate the above data to obtain the sales data for the Audi A4 model in the Korean market during the first quarter of 2019, which was 64 units. \n10. Calculate the year-over-year growth rate for the Audi A4: 1094/64 - 1 = 135.78%. \n11. In the 06138b.xlsx file, locate the sales data for the Audi Q5 in the U.S. market for January 2019, February 2019, and March 2019, which were 4,541 units, 4,888 units, and 5,302 units, respectively. \n12. Merge and calculate the above data to obtain the sales data for the Audi Q5 model in the U.S. market during the first quarter of 2019, which was 14,731 units. \n13. Calculate the year-over-year growth rate for the Audi Q5: 9959/14731 - 1 = -32.29%. \n14. Conclusion: In the first quarter of 2020, the sales growth rate of the best-selling Audi model in the Korean market, the Audi A4, was faster than that of the best-selling Audi model in the U.S. market, the Audi Q5.")
"Number of steps": "14"

```

Figure 18: An example from the File Processing subset.

```

"task_id": "07100",
"Question": "2019年诺贝尔物理学奖的得主中, 谁曾在瑞士大学任职过的人, 获得诺贝尔奖时多大年纪? ", ("The winner of the 2019 Nobel Prize in Physics who had never worked at a Swiss university, how old were they when they won the Nobel Prize?")
"Label": "General Knowledge",
"Level": "2",
"Final answer": "84",
"File_name": "",
"Steps": "1. 在网上搜索诺贝尔物理学奖官网。 \n2. 单击诺贝尔奖官方网站或权威新闻网站, 查看2019年诺贝尔物理学奖得主。 \n3. 查找并确认获奖者的名字 (詹姆斯·皮布尔斯、米歇尔·麦耶和迪迪埃·奎洛兹)。 \n4. 返回搜索引擎, 依次搜索“迪迪埃·奎洛兹”、“米歇尔·麦耶”和“迪迪埃·奎洛兹”的百科页面。 \n5. 在个人简介中查找每个人与瑞士大学的关系。 \n6. 经过对比, 可以知道詹姆斯·皮布尔斯曾在瑞士大学任职过。 \n7. 确认詹姆斯·皮布尔斯的出生日期为1935年4月25日。 \n8. 则他获得2019年诺贝尔奖时已经84岁高龄。
", ("1. Search for the official Nobel Prize website on the internet. \n2. Click on the official Nobel Prize website or an authoritative news website to view the page for the 2019 Nobel Prize in Physics. \n3. Find and confirm the names of the laureates (James Peebles, Michel Mayor, and Didier Queloz). \n4. Return to the search engine and search for the Wikipedia pages of "Didier Queloz," "Michel Mayor," and "Didier Queloz" in sequence. \n5. Look for the connection of each person with Swiss universities in their biographies. \n6. After comparison, it is evident that James Peebles has never worked at a Swiss university. \n7. Confirm that James Peebles was born on April 25, 1935. \n8. Therefore, he was 84 years old when he received the 2019 Nobel Prize.")
"Number of steps": "8"

```

Figure 19: An example from the General Knowledge subset.

Doubao-1.5-Vision-Pro-32k and GPT-4o; and two high-performing open-source models, Qwen2-VL-72B and Llama-3.2-90B-Vision. This multimodal reasoning example effectively facilitates a comparative analysis of the performance differences among existing MLLMs.

Doubao-1.5-Vision-Pro-32k and GPT-4o provided both accurate analyses and correct answers. In contrast, Qwen2-VL-72B produced an accurate answer but with errors in its analysis process, while Llama-3.2-90B-Vision failed in both its analysis and the final answer.

Figure 22 illustrates a logical reasoning example from C²RBench, showcasing responses from four models: one high-performing closed-source large reasoning model, o1-preview; one high-performing

```

"task_id": "08036",
"Question": "她是否坚定地不接受丈夫人的提议? ", ("Did Yuanyang firmly refuse to accept Madam King's proposal?")
"Label": "Reading",
"Level": "2",
"Final answer": "是", ("Yes")
"File_name": "08036a.docx, 08036b.docx",
"Steps": "1. 了解那夫人的提议: 那夫人提议让曾曾去老太太那里成为大老多的姨婆, 描述了这个位置的优越性和尊贵性。 \n2. 分析曾曾的初步反应: 起初, 曾曾在那夫人面前只是低头不语, 显露出她的尴尬和不情愿。 \n3. 那夫人的劝说: 那夫人解释为曾曾, 提到她的性格优点和这个提议对她的好处, 试图说服曾曾。 再告知其家人做主。 \n4. 分析曾曾的态度: 曾曾知道还会有人提议这件事, 因此躲了出去。 \n5. 注意曾曾的进一步反应: 当平儿和婆人建议曾曾以其他理由拒绝大老多时, 曾曾对这些建议表示愤怒和不满, 进一步表明她坚定的不接受接受。 \n6. 理解曾曾的坚决表态: 曾曾明确表示, 她宁愿保持清白, 不接受那夫人的提议, 即使面临其他困难情况, 如成为寡妇或死。 \n7. 总结曾曾的坚定立场: 曾曾的态度没有改变, 她清楚地表达了自已宁愿不接受那夫人的提议, 也不愿接受大老多的任何提议, 显示出她的坚定立场。
", ("1. Understanding Madam King's Proposal: Madam King suggested that Yuanyang follow her to the Old Lady's place to become the eldest master's concubine, describing the advantages and nobility of this position. \n2. Analyzing Yuanyang's Initial Reaction: At first, Yuanyang merely lowered her head and remained silent in front of Madam King, showing her embarrassment and reluctance. \n3. Madam King's Persuasion: Madam King continued to persuade Yuanyang, mentioning her personality traits and the benefits of this proposal, attempting to convince Yuanyang and informing her that she would seek permission from her family. \n4. Analyzing Yuanyang's Attitude: Yuanyang knew that others would also propose this matter, so she hid away. \n5. Noticing Yuanyang's Further Reaction: When Ping'er and Xiren suggested that Yuanyang refuse the eldest master for other reasons, Yuanyang expressed anger and dissatisfaction with these suggestions, further indicating her firm refusal to accept. \n6. Understanding Yuanyang's Firm Statement: Yuanyang clearly stated that she would rather remain pure and not accept Madam King's proposal, even if faced with extreme situations, such as becoming a nun or dying. \n7. Summarizing Yuanyang's Firm Stance: Yuanyang's attitude remained unchanged. She clearly expressed that she would rather not accept Madam King's proposal, nor would she accept any marriage proposal from the eldest master, demonstrating her steadfast position.")
"Number of steps": "7"

```

Figure 20: An example from the Reading Comprehension subset.

open-source large reasoning model, DeepSeek-R1; Hunyuan-turbo-latest, and GPT-4o.

For the logical (see Figure 21) and multimodal (see Figure 22) subsets, we show responses sampled from four models, with green text representing correct answers and red text representing incorrect ones.

A.6 Overthinking Analysis Setting

In this appendix, we present the detailed metrics for token consumption by the models under various evaluation conditions.

Token Breakdown:

- **Reasoning Tokens:** The number of tokens spent on intermediate reasoning steps.
- **Answer Tokens:** The tokens used to generate the final response.
- **Total Tokens:** The combined total of reasoning and answer tokens.

We track these values for each model (Doubao-1.5-Thinking-Pro, Kimi-Thinking-Preview, and GPT-4o) across different task types, highlighting inefficiencies in reasoning and areas for improvement.

A.7 Human Evaluation

A.7.1 Task Selection Strategy

We adopted a stratified random sampling strategy to ensure balanced and representative task coverage across the C²RBench. Specifically, we sampled 10% of tasks from each combination of difficulty level (1–3) and category label, resulting in a total of 113 tasks. This approach guarantees proportional representation across task types and complexity levels.



Figure 21: An example from the **Multimodal** subset. Correct analyses are highlighted in green, while errors are marked in red.

A.7.2 Task Distribution

Table 6 summarizes the detailed distribution of the 113 sampled tasks, including task ID, category label, difficulty level (1–3), and average human score.

A.7.3 Participant Demographics

A total of ten human evaluators were recruited for the human assessment, each assigned to specific subsets based on their background and expertise:

- **Multimodal:** This subset consists of visual reasoning questions inspired by Chinese civil service examinations. Three participants with prior experience in civil service exam preparation were selected. Two held undergraduate degrees in Chinese Language and Literature, and one held a degree in Artificial Intelligence.
- **Logical, Graph Analysis, File Processing, General Knowledge:** These subsets involve general reasoning tasks requiring basic operational knowledge of computer tools, without demanding domain-specific expertise. Two participants evaluated these subsets: one was a master’s student in Electronic Information Engineering, and the other was a bachelor’s graduate in Artificial Intelligence.
- **Math:** This subset includes mathematics problems at the high school to university

level. Two participants, both from mathematics backgrounds, were involved: one was a Ph.D. candidate and the other a master’s student in Mathematics.

- **Science:** This subset comprises physics, biology, and chemistry questions at the level of the Chinese National College Entrance Examination (Gaokao). Two participants evaluated this subset: one was an undergraduate student majoring in Architecture with a science background from Gaokao, and the other was a master’s student in Physics.
- **Reading:** This subset contains reading comprehension tasks based on classical and modern literature, both Chinese and foreign. Three undergraduate students majoring in Chinese Language and Literature were recruited to evaluate this subset.

<p>"Question": "在一天的24小时之中，时钟的时针、分针和秒针完全重合在一起有几次，分别是什么时间？" ("In a 24-hour day, how many times do the hour, minute, and second hands of a clock coincide completely, and at what times?") "Label": "Logical", "Level": "3", "Final answer": "Twice, at 12:00 and 00:00.", "File name": "", "Steps": "1. 问题陈述... \n2. 确定角速度... \n3. 分针与时针重合的计算... \n4. 时间换算... \n5. 秒针重合验证... \n6. 三针重合条件: \n三针重合的唯一时间是在正12点和0点. \n7. 数学证明: ... \n8. 计算重合时间: ... \n9. 最小公倍数: ... 最小公倍数为 $11 * 719$. \n重合时间间隔为 $11 * 719 * \Delta$ (即12小时). \n10. 首次重合: \n三针在 $t = 0$ 时重合. \n11. 下一次重合: \n下一次重合时间: $t = 12$ 小时. \n12. 结论: \n在24小时内，时针、分针和秒针完全重合的时间只有两次: 12点和0点." ("1. Problem statement... \n2. Determine angular velocities ... \n3. Compute the coincidence of the minute and hour hands ... \n4. Convert time units ... \n5. Verify second-hand alignment ... \n6. Condition for complete overlap... \n - The only times when all three hands coincide are exactly at 12:00 and 00:00. \n7. Mathematical proof: ... \n8. Compute the coincidence time: ... \n9. Least common multiple: ... The coincidence interval is $11 * 719 * \Delta$ (i.e., 12 hours). \n. First coincidence: \n - The three hands coincide at $t = 0$. \n11. Next coincidence: \n - The next coincidence occurs at $t = 12$ hours. \n12. Conclusion: \n - In a 24-hour period, the hour, minute, and second hands coincide completely only twice: at 12:00 and 00:00.",)</p>			
<p>o1-preview 分析: \n时钟的时针、分针和秒针完全重合的时刻，是当它们同时指向12的位置。... 尽管它们的速度不同，但每12小时都会同时回到12的位置。因此，在24小时内，它们会在凌晨12:00:00和中午12:00:00各重合一次。 \n\n答案: \n一共2次，分别在凌晨12:00:00和中午12:00:00时刻。 (Analysis: \nThe moment when the hour, minute, and second hands of a clock completely coincide occurs when all three hands point to the 12 o'clock position simultaneously. ... Despite their differing speeds, all three hands align at the 12 o'clock position every 12 hours. Therefore, within a 24-hour period, they coincide twice: once at 12:00:00 AM and again at 12:00:00 PM. \n\nAnswer: \nA total of two times, specifically at 12:00:00 AM and 12:00:00 PM.)</p>	<p>DeepSeek-R1 分析: 在24小时内，时针和分针每12小时重合11次，但加入秒针后，只有当秒针也同时指向同一位置时才会三针重合。由于分针和时针的重合时间点并非整秒数（如约1小时5分27秒），此时秒针位置无法与分针、时针同步。只有12点整和0点整时，三个指针完全重合。24小时内，0点和12点各出现一次，共两次。 \n\n答案: 两次，分别是0时0分0秒和12时0分0秒。 (Analysis: Within 24 hours, the hour hand and minute hand overlap 11 times every 12 hours. However, with the addition of the second hand, the three hands only coincide when the second hand also points to the same position. Since the overlap time of the minute hand and hour hand is not an exact second (for example, around 1 hour, 5 minutes, and 27 seconds), the second hand cannot synchronize with the positions of the minute and hour hands at this moment. The three hands only coincide exactly at 12:00:00 and 0:00:00. Therefore, within 24 hours, this happens twice—once at 0:00:00 and once at 12:00:00. \n\nAnswer: Twice, at 0:00:00 and 12:00:00.)</p>	<p>Hunyuan-turbo-latest 分析: ... 要使得时针、分针和秒针完全重合，需要在分针和时针重合的时刻，秒针也恰好在12点位置。... 一天24小时中，时针、分针和秒针完全重合在一起只有两次，分别是0点0分0秒和12点0分0秒。 \n\n答案: 两次，分别是0点0分0秒和12点0分0秒。 (Analysis: ... For all three hands to coincide, the second hand must also be at the 12 o'clock position at the exact moment when the minute and hour hands overlap. ... it is determined that within a 24-hour period, the hour, minute, and second hands align perfectly only twice: at 12:00:00 AM and 12:00:00 PM. \n\nAnswer: Twice, at 12:00:00 AM and 12:00:00 PM.)</p>	<p>GPT-4o 分析: ... 因为时针每12小时转一圈，分针每小时转一圈，秒针每分钟转一圈，所以它们在每个整点时刻重合。一天有24小时，因此它们会重合24次。 \n\n答案: 24次，分别是每个整点时刻(例如: 0:00:00, 1:00:00, 2:00:00, ..., 23:00:00) (Since the hour hand completes one full revolution every 12 hours, the minute hand every hour, and the second hand every minute, they align precisely at the beginning of each hour. Given that there are 24 hours in a day, this alignment occurs 24 times. \n\nAnswer: 24 times, occurring at the start of each hour (e.g., 0:00:00, 1:00:00, 2:00:00, ..., 23:00:00).)</p>

Figure 22: An example from the Logical subset. Correct analyses are highlighted in green, while errors are marked in red.

Table 6: Detailed distribution of the sampled tasks used for human evaluation.

Task ID	Category	Level	Avg. Score
01007	Multimodal	1	1
01016	Multimodal	2	0.33
01018	Multimodal	2	0.33
01031	Multimodal	3	0.67
01039	Multimodal	3	0.67
01042	Multimodal	2	0.33
01046	Multimodal	2	0.33
01057	Multimodal	3	1
01075	Multimodal	3	1
01106	Multimodal	2	0.33
01107	Multimodal	2	0.33
01118	Multimodal	2	0.67
01133	Multimodal	2	0.67
01142	Multimodal	3	0.33
01143	Multimodal	3	0.33
01145	Multimodal	3	0.33
02003	Logical	3	1
02007	Logical	1	1
02009	Logical	1	1
02015	Logical	1	1
02033	Logical	1	1
02042	Logical	2	0
02043	Logical	2	0
02066	Logical	3	1
02073	Logical	3	1
02081	Logical	2	1
02097	Logical	2	1
02108	Logical	2	1
02130	Logical	2	1
02142	Logical	3	1
02156	Logical	3	1
02159	Logical	3	1
03005	Graph Analysis	1	1
03014	Graph Analysis	1	1
03037	Graph Analysis	1	1
03042	Graph Analysis	2	1
03073	Graph Analysis	2	1
03076	Graph Analysis	1	1
03079	Graph Analysis	1	1
03083	Graph Analysis	2	1
03090	Graph Analysis	2	1
03100	Graph Analysis	1	1
03101	Graph Analysis	1	1
03116	Graph Analysis	1	1
04002	Math	1	1
04011	Math	1	1
04012	Math	1	1
04017	Math	3	1
04032	Math	3	1
04035	Math	1	1
04045	Math	1	1
04054	Math	2	1
04082	Math	2	1
04085	Math	3	1
04096	Math	3	1
04114	Math	1	1
04124	Math	1	1
04135	Math	2	1
04137	Math	2	1
04153	Math	3	1

Task ID	Category	Level	Avg. Score
05003	Science	2	1
05007	Science	3	1
05014	Science	3	1
05026	Science	1	0.75
05037	Science	3	1
05043	Science	1	0.5
05047	Science	2	1
05049	Science	1	1
05060	Science	3	1
05069	Science	3	1
05091	Science	1	1
05096	Science	1	1
05108	Science	3	1
05124	Science	3	1
05129	Science	3	1
05149	Science	3	1
05155	Science	3	1
06010	File Processing	2	1
06023	File Processing	2	1
06031	File Processing	2	1
06041	File Processing	2	1
06046	File Processing	1	1
06047	File Processing	1	1
06050	File Processing	3	1
06056	File Processing	3	1
06075	File Processing	1	1
06094	File Processing	1	1
06102	File Processing	2	1
06112	File Processing	2	1
06138	File Processing	3	1
06140	File Processing	3	1
07022	General Knowledge	1	1
07028	General Knowledge	1	1
07029	General Knowledge	1	1
07035	General Knowledge	3	0
07045	General Knowledge	2	1
07057	General Knowledge	2	1
07059	General Knowledge	2	1
07076	General Knowledge	2	1
07090	General Knowledge	2	1
07100	General Knowledge	2	1
07102	General Knowledge	2	1
08010	Reading	1	1
08018	Reading	1	1
08027	Reading	1	1
08033	Reading	1	1
08038	Reading	2	1
08041	Reading	2	1
08063	Reading	2	1
08064	Reading	2	1
08075	Reading	2	1
08109	Reading	2	1
08110	Reading	2	1