

LogicOCR: Do Your Large Multimodal Models Excel at Logical Reasoning on Text-Rich Images?

Maoyuan Ye¹, Jing Zhang^{1✉}, Juhua Liu^{1✉}, Bo Du¹, Dacheng Tao²

¹ School of Computer Science, Wuhan University, China

² College of Computing & Data Science, Nanyang Technological University, Singapore
{yemaoyuan, liuhuhua, dubo}@whu.edu.cn, {jingzhang.cv, dacheng.tao}@gmail.com

Abstract

Recent advances in Large Multimodal Models (LMMs) have significantly improved their reasoning and Optical Character Recognition (OCR) capabilities. However, their performance on complex logical reasoning tasks involving text-rich images remains underexplored. To bridge this gap, we introduce LogicOCR, a benchmark comprising 1,100 multiple-choice questions designed to evaluate LMMs' logical reasoning abilities on text-rich images, while minimizing reliance on domain-specific knowledge (*e.g.*, mathematics). We construct LogicOCR by curating a text corpus from the Chinese National Civil Servant Examination and develop a scalable, automated pipeline to convert it into multimodal samples. First, we design prompt templates to steer GPT-Image-1 to generate images with diverse backgrounds, interleaved text-illustration layouts, and varied fonts, ensuring contextual relevance and visual realism. Then, the generated images are manually verified, with low-quality examples discarded. We evaluate a range of representative open-source and proprietary LMMs under both Chain-of-Thought (CoT) and direct-answer settings. Our multi-dimensional analysis reveals key insights, such as the impact of test-time scaling, input modality differences, and sensitivity to visual-text orientation. Notably, LMMs still lag in multimodal reasoning compared to text-only inputs, indicating that they have not fully bridged visual reading with reasoning. We hope LogicOCR will serve as a valuable resource for advancing multimodal reasoning research. The dataset is available at [LogicOCR](#).

1 Introduction

Recent advances in Large Multimodal Models (LMMs) [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11] have drawn significant attention to their logical reasoning and Optical Character Recognition (OCR) capabilities. While these skills are crucial for real-world applications, LMMs' ability to perform complex logical reasoning on text-rich images remains underexplored. Table 1 compares representative benchmarks for multimodal reasoning and OCR. Existing multimodal reasoning datasets [12, 13, 14, 15] often require extensive mathematical or scientific knowledge, making it difficult to isolate pure reasoning ability from domain expertise. In contrast, most OCR-related benchmarks [16, 17, 18, 19, 20, 21, 22, 23, 24, 25] lack complexity, potentially overstating LMM progress in integrating reading, understanding, and reasoning. Although CharXiv [26] and OCRBench v2 [27] include reasoning subsets, these are either narrowly focused on chart interpretation or rely on specialized knowledge.

To address the gap in evaluating logical reasoning on text-rich images, we introduce LogicOCR, a benchmark designed to assess LMMs' ability to perform complex reasoning with minimal reliance on domain-specific knowledge. LogicOCR comprises 1,100 multiple-choice questions, spanning

✉: Corresponding author.

five reasoning types: categorical, sufficient conditional, necessary conditional, disjunctive, and conjunctive (following the previous definition [28, 29, 30]). We build the benchmark from a curated corpus of pure logical reasoning questions sourced from the National Civil Servant Examination of China, rendering them into diverse, realistic images. To achieve this, we develop an automated, scalable pipeline. To be specific, we design prompt templates to steer GPT-Image-1 [31] to generate text-rich images with varied layouts (*e.g.*, interleaved text and illustrations, backgrounds), fonts (handwritten and standard). These prompts ensure visual elements are closely aligned with the question content, making the images more natural and contextually rich. Background descriptions are generated using Qwen2.5-14B [32] based on the question context and integrated into the prompts. All generated images undergo manual quality control to remove low-quality samples.

We evaluate a range of state-of-the-art LMMs, including open-source models like Qwen2.5-VL [9] and InternVL3 [10], as well as proprietary models such as Gemini-2.5-Pro [33] and o4-mini [34], under both direct answering and Chain-of-Thought (CoT) settings. Through multi-dimensional analysis, we reveal some key findings: ① **CoT does not consistently improve accuracy**—most models fail to reason better step-by-step, suggesting flaws in their reasoning paths. ② **Test-time scaling is effective in improving performance**, though its efficiency remains limited. ③ **Reasoning over text-rich images remains a bottleneck**, suggesting that LMMs still struggle to fully bridge visual reading with reasoning. ④ **OCR robustness is still a major weakness**—minor perturbations like image rotation can reduce accuracy to near-random levels.

In summary, our main contributions are three-fold: 1) We present LogicOCR, a benchmark that evaluates LMMs’ logical reasoning on text-rich images with minimal reliance on domain knowledge. 2) We propose a scalable, automated pipeline using GPT-Image-1 to convert text corpora into diverse, high-quality multimodal samples. 3) We provide a comprehensive evaluation of LMMs and highlight key insights to guide future improvements in multimodal reasoning.

2 Related Work

Reasoning-Capable LMMs. LMMs have advanced from high-resolution image perception [35, 36, 37, 38, 39, 40, 41, 42, 4, 7, 3] to increasingly sophisticated multimodal reasoning [9, 10, 43, 34, 11, 44]. High-resolution input enables LMMs to read with human-like accuracy, while test-time scaling [43, 45] extends their capabilities to more complex tasks. However, a key question remains: have LMMs truly integrated visual reading and reasoning? This work also explores this question through our proposed LogicOCR benchmark.

Multimodal Reasoning Benchmarks. Several benchmarks have been developed to evaluate the multimodal reasoning abilities of LMMs. Some focus on both scientific knowledge and reasoning [13, 46], while others emphasize mathematical, computational, or visual puzzle tasks [12, 14, 15, 47, 48]. Most demand extensive domain knowledge but lack samples with dense visual-text content, limiting their assessment of reasoning over rich contextual information.

OCR-Related Benchmarks. Various benchmarks [16, 20, 21, 25, 18] have been introduced to assess LMMs’ capabilities in text recognition, key information extraction, document parsing, and visual question answering. However, they often neglect reasoning complexity, and some, like DocVQA [17],

Benchmarks	Data Size	Complex Reasoning	Knwl. Free	Dense Text	Diverse Topics	Image Type
<i>Multimodal Reasoning</i>						
MathVista [12]	6.1K	✓	✗	✗	✓	Real
MMMU [13]	10.5K	✓	✗	✗	✓	Real
MATH-Vision [15]	3.0K	✓	✗	✗	✓	Real
MathVerse [14]	2.6K	✓	✗	✗	✓	Real
<i>OCR-Related</i>						
TextVQA [16]	5.7K	✗	✓	✗	✓	Real
DocVQA [17]	5.2K	✗	✓	✓	✗	Real
InfoVQA [18]	3.3K	✗	✓	✓	✓	Real
ChartQA [19]	2.5K	✗	✓	✗	✗	Real
OCRBench [20]	1.0K	✗	✓	✓	✓	Real
CC-OCR [24]	7.1K	✗	✓	✓	✓	Real
<i>Multimodal Reasoning & OCR-Related</i>						
CharXiv [26]	1.3K	✓	✓	✗	✗	Real
OCRBench v2 [27]	2.2K	✓	✗	✗	✓	Real
LogicOCR (ours)	1.1K	✓	✓	✓	✓	Generated

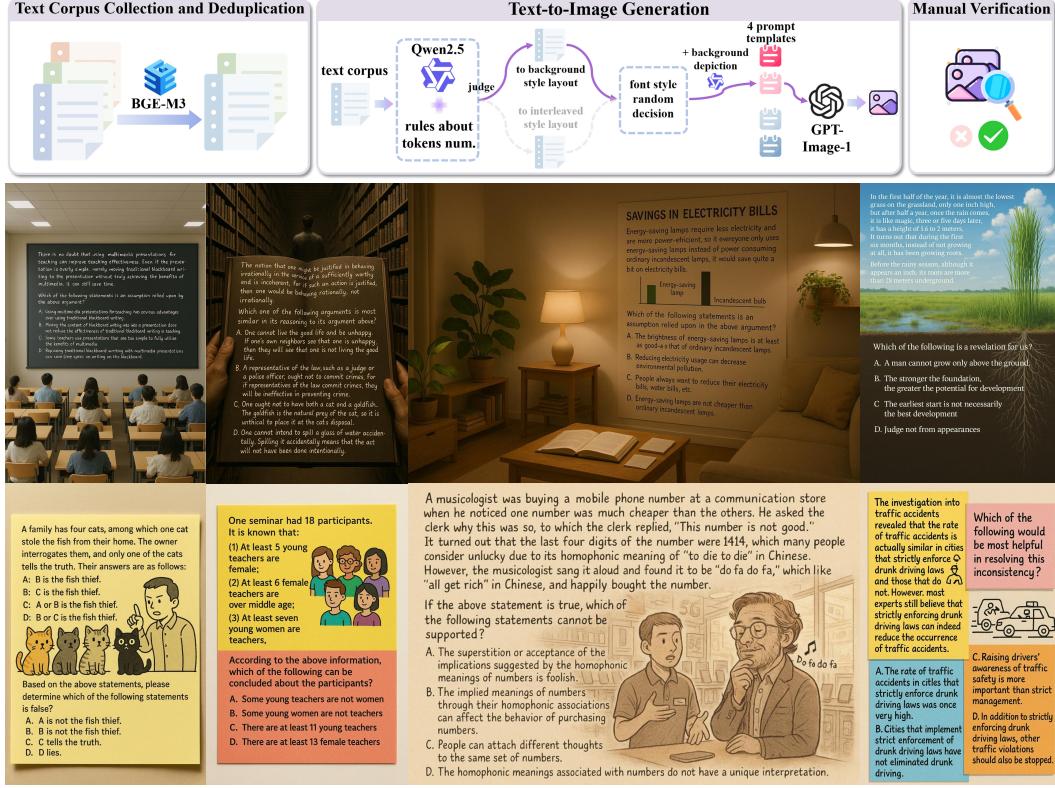


Figure 1: Illustration of the LogicOCR data construction process and sample images, showcasing background-style and text-illustration interleaved layouts from top to bottom.

quickly reach performance saturation. Recent chart-focused benchmarks [49, 50, 26, 19] introduce numerical reasoning but are limited to chart data. In contrast, our LogicOCR benchmark evaluates complex logical reasoning in diverse, text-rich images while reducing reliance on domain knowledge and complex numerical computation.

3 LogicOCR

3.1 Data Collection and Construction

To reduce the impact of domain-specific knowledge in reasoning questions, we collect a text corpus from the National Civil Servant Examination of China, which emphasizes pure logical reasoning. An automated pipeline is then employed to convert the corpus into diverse images, as shown in Fig. 1.

Text Corpus Collection and Deduplication. We reuse the text corpus from the test sets of LogiQA [29] and LogiQA2.0 [30], both derived from the National Civil Servant Examination of China. For LogiQA samples containing mixed Chinese and English, we translate the original Chinese text into English. Each sample consists of three components: *context*, *question*, and *options*, with the context providing all necessary information to answer the question. Since multiple questions may share the same context, we perform deduplication using the context, retaining only one sample per unique context. We generate embeddings for each context using BGE-M3 [51] and apply cosine similarity to identify and remove duplicates.

Table 2: LogicOCR statistics. Each sample may cover multiple reasoning categories.

Statistics	Value
Total Questions	1,100
- LogiQA [29] corpus	330
- LogiQA2.0 [30] corpus	770
Reasoning Categories	
- categorical	50.5%
- sufficient conditional	86.3%
- necessary conditional	37.5%
- disjunctive	21.6%
- conjunctive	66.3%
Answer Distribution	
- A/B/C/D	24.0%/23.3%/27.7%/25.0%
Image Characteristics	
- background style layout	41.5%
- interleaved style layout	58.5%
- non-handwritten style font	53.2%
- handwritten style font	46.8%
- average words in image	139.2

Text-to-Image Generation. We develop an automated pipeline that converts the collected text corpus into images using Qwen2.5 [32] and GPT-Image-1 [31]. To enhance image diversity, generation success, and visual-text readability, we introduce four prompt templates that account for layout and font style when instructing GPT-Image-1.

Firstly, we design two layout styles: (1) full-frame background scenes with contrastive visual-text in the foreground, and (2) interleaved text and illustrations on randomly colored paper. To determine suitability for the background style, we use Qwen2.5-14B for evaluation and assign this layout to samples exceeding 250 tokens. The background scenes and illustrations are contextually aligned with the questions to enhance visual relevance and naturalness. To support this, we add tailored instructions into the prompt templates. For samples requiring background scenes, Qwen2.5-14B also generates short contextual descriptions, which are inserted into designated placeholders in the templates for text-to-image generation. For the interleaved style, GPT-Image-1 effectively produces suitable illustrations based on our prompt templates, so no additional image descriptions are needed.

Secondly, we define font style as either handwritten or non-handwritten, sampled randomly without specifying exact fonts. Combined with the two layout types, this yields four prompt templates, detailed in Appendix A. For GPT-Image-1 generation, we set the quality to “high” and the resolution to either 1536×1024 or 1024×1536 .

Motivation for Using GPT-Image-1 in Image Generation. Because generated images and their source text are like modality twins, we can directly compare performance across input types by feeding plain-text questions and multimodal problems into the same LMM. This enables us to pinpoint multimodal reasoning bottlenecks in LogicOCR. GPT-Image-1 offers strong instruction-following and high-fidelity visual-text rendering, presenting a promising alternative to prior synthetic methods [52, 53]. Unlike those approaches, it produces more natural, vivid images without requiring complex rule design. Notably, GPT-Image-1 can intelligently adapt text placement, color, and wrapping to fit the background and surrounding elements.

Manual Verification. Generated images are saved in JPEG format and resized to a maximum dimension of 1024 pixels to reduce the computation overhead of LMMs during evaluation. Each image undergoes a manual verification step to ensure text readability. Images lacking correct answer options or impairing judgment are discarded. The overall generation success rate is about 73%.

Prompt Templates for Evaluating LMMs. We evaluate LMMs under both CoT and direct answering settings, with prompts incorporating images and instructions. The templates used are detailed in Appendix B, including those for text-only input evaluation.

3.2 Statistics

The key statistics of LogicOCR are presented in Tab. 2. Based on the definition of corpus sources [29, 30], five logical reasoning categories are considered: categorical, sufficient conditional, necessary conditional, disjunctive, and conjunctive. For LogiQA2.0 [30], we directly use the annotated reasoning type for each sample. In contrast, since LogiQA [29] lacks such annotations, we classify its samples using o4-mini [34] with few-shot examples. Note that a single sample may correspond to multiple reasoning types. Regarding image layout, 41.5% follow a background-style format, while the rest adopt an interleaved layout. A word cloud illustrating background scenes is shown in Fig. 2. On average, each image contains over 130 words, exceeding the text density of other OCR benchmarks such as HierText[54], which averages 103.8 words per image.

3.3 Evaluation Protocol

Accuracy is used as the evaluation metric. Each LMM completion is parsed to extract the selected option letter, and a binary score is assigned based on the ground-truth option.



Figure 2: Word cloud of background scenarios.

Table 3: Evaluation results on LogicOCR under CoT and direct answering settings. Questions may involve multiple reasoning types, with 17.4%, 26.1%, 35.0%, and 21.5% containing 1, 2, 3, and more than 3 reasoning types, respectively. ‘♠’ denotes document-oriented LMMs.

Model	Reasoning Types				Average	Average
	1	2	3	>3	(CoT)	(Direct)
<i>Proprietary LMMs</i>						
GPT-4o [3]	64.6	67.2	60.5	61.0	63.1	60.3
o4-mini [34]	79.7	77.0	76.4	77.1	77.3	—
Claude-3.7-Sonnet [55]	75.0	73.5	68.1	73.3	71.8	59.1
Gemini-2.5-Pro [33]	82.3	81.7	80.6	78.4	80.7	—
<i>Open-Source LMMs</i>						
TextMonkey [56] ♠	12.0	15.3	14.8	22.5	16.1	24.0
DocOwl2 [37] ♠	13.0	22.6	16.9	22.9	19.0	16.2
MiniMonkey [57]	30.7	31.0	28.3	30.9	30.0	30.6
LLaVA-OV-7B [58]	35.9	31.7	33.2	32.6	33.2	34.3
DeepSeek-VL2 [7]	37.0	36.6	35.8	39.8	37.1	41.3
NVILA-8B [2]	37.0	40.4	39.0	46.2	40.6	41.5
Kimi-VL-A3B-Instruct [11]	48.4	49.8	48.8	51.3	49.6	54.1
Ovis2-4B [59]	41.1	41.8	35.8	41.1	39.5	45.0
Ovis2-8B [59]	50.0	50.5	45.5	47.0	47.9	49.2
Ovis2-16B [59]	57.3	57.5	53.8	50.4	54.6	52.6
Ovis2-34B [59]	61.5	56.1	56.9	55.9	57.3	59.8
InternVL3-2B [10]	38.0	37.6	33.0	32.2	34.9	37.7
InternVL3-8B [10]	58.9	50.9	48.1	46.6	50.4	50.6
InternVL3-14B [10]	64.1	61.3	60.5	58.5	60.9	61.6
InternVL3-38B [10]	64.6	60.6	59.7	61.9	61.3	61.7
Qwen2-VL-2B [4]	38.5	39.7	36.9	36.9	37.9	40.6
Qwen2-VL-7B [4]	43.2	47.0	46.8	41.9	45.2	50.7
Qwen2.5-VL-3B [9]	47.4	47.4	42.9	43.2	44.9	49.6
Qwen2.5-VL-7B [9]	58.9	50.2	49.6	49.6	51.4	52.6
Qwen2.5-VL-32B [9]	67.2	63.4	66.8	63.1	65.2	64.2
Qwen2.5-VL-72B [9]	70.3	64.8	62.9	67.4	65.6	67.2

4 Experiments

4.1 Setup

We select a variety of LMMs for evaluation, including both open-source and proprietary models. We also incorporate the models optimized for multimodal reasoning in the evaluation. We report both the performance under CoT and direct answering (only answer the option’s letter) settings.

Proprietary Models. We select several cutting-edge proprietary models, including GPT-4o [3], o4-mini [34], Claude-3.7-Sonnet [55], and Gemini-2.5-Pro [33].

Open-Source Models. For open-source LMMs, we select a various of candidates including TextMonkey [56], DocOwl2 [37], MiniMonkey [57], LLaVA-OV [58], DeepSeek-VL2 [7], NVILA [2], Kimi-VL [11], Ovis2 series [59], InternVL3 series [10], Qwen2-VL series [4], Qwen2.5-VL series [9], and also QvQ-72B-Preview [44].

For o4-mini, Claude-3.7-Sonnet, Gemini-2.5-Pro, and QvQ-72B-Preview, the max output token length is set to 8,192, while 2,048 is used for other models. The experiments are conducted on Nvidia A800 GPUs.

4.2 Main Results

The quantitative accuracy of LMMs is shown in Tab. 3. Overall, LogicOCR presents a significant challenge to most state-of-the-art LMMs. Under the CoT setting, Gemini-2.5-Pro achieves the highest accuracy of 80.7% among both proprietary and open-source models. Recent open-source models have made notable strides in multimodal reasoning, with Qwen2.5-VL-32B outperforming

GPT-4o. However, comparing the CoT results of Qwen2.5-VL-32B and Qwen2.5-VL-72B, we see no substantial improvement with the increase in parameter size.

Most LMMs show no improvement with CoT on LogicOCR. We find that only large-scale LMMs, like Ovis2-16B, Qwen2.5-VL-32B, and proprietary models, benefit from CoT. In contrast, smaller models such as Ovis2-4B, InternVL3-2B, and Qwen2.5-VL-3B experience a notable performance drop when using CoT. This suggests that smaller LMMs lack the reliable multimodal reasoning capabilities needed for CoT, which may introduce unnecessary complexity, reducing their accuracy. For larger models, such as Ovis2-34B and InternVL3-38B, CoT still does not outperform direct answers, indicating that these models struggle to identify the appropriate reasoning path for logical tasks. Previous research [60] shows that CoT is more effective for math, symbolic, and algorithmic tasks, while its utility may be limited for tasks like commonsense reasoning. Our LogicOCR benchmark presents a challenging multimodal reasoning scenario for LMMs.

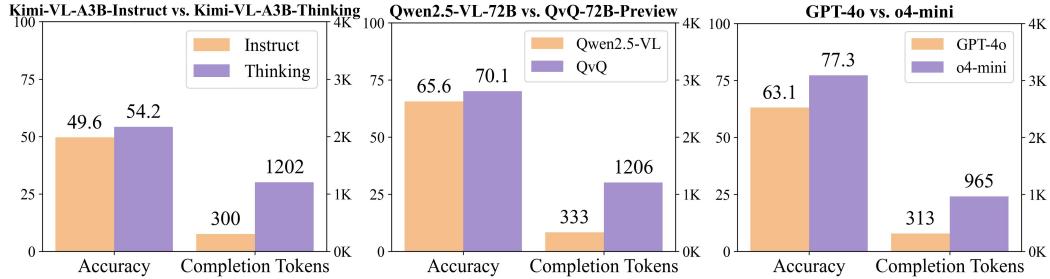


Figure 3: Comparison of average accuracy and output length (completion tokens) between general LMMs and their reasoning-enhanced counterparts. Overall, test-time scaling improves accuracy. Notably, o4-mini achieves much higher accuracy using fewer tokens than QvQ-72B-Preview.

4.3 Analysis

4.3.1 Is Test-Time Scaling Beneficial for LogicOCR?

Recent models have explored test-time scaling to enhance multimodal reasoning. Our experiments confirm that **test-time scaling significantly improves performance on LogicOCR, though the efficiency of open-source LMMs still leaves room for improvement.**

Specifically, for open-source models, we select Kimi-VL-A3B-Thinking [11] and QvQ-72B-Preview [44], and compare them with their general counterparts, *i.e.*, Kimi-VL-A3B-Instruct and Qwen2.5-VL-72B. For proprietary models, we assess o4-mini [34]. Figure 3 presents the average accuracy under the CoT setting and the corresponding completion token usage. Kimi-VL-A3B-Thinking achieves a 4.6-point accuracy gain over its base version, Kimi-VL-A3B-Instruct, but requires 4× more completion tokens. Similarly, QvQ-72B-Preview improves accuracy by 4.5 points at a comparable token cost, *i.e.*, 4× more completion tokens than Qwen2.5-VL-72B. It is noteworthy that o4-mini attains 7.2 points higher accuracy than QvQ-72B-Preview while using only 80% of its token length, suggesting potential redundancy in QvQ-72B-Preview’s reasoning. These findings indicate that future reasoning-oriented LMMs should aim for more concise reasoning strategies to enhance efficiency without compromising accuracy. In terms of data and training, curating variable-length CoT data according to task difficulty is essential [61]. For reinforcement learning, applying length penalty on responses while ensuring the answer correctness [62, 63] is a natural and useful solution. During inference, token-level and structured sampling strategies [64, 65] with higher efficiency and lower performance decline are worth further investigation.

4.3.2 Have LMMs Effectively Bridged Visual Reading and Reasoning?

To investigate this question, we evaluate three state-of-the-art LMMs: Ovis2 [59], InternVL3 [10], and Qwen2.5-VL [9]. Each model is tested with both text-only and multimodal inputs to assess whether their reasoning over visual content matches their language-only reasoning capabilities. In addition, we test LMMs in text-only setting with their OCR results as input, instead of providing ground-truth question corpora.

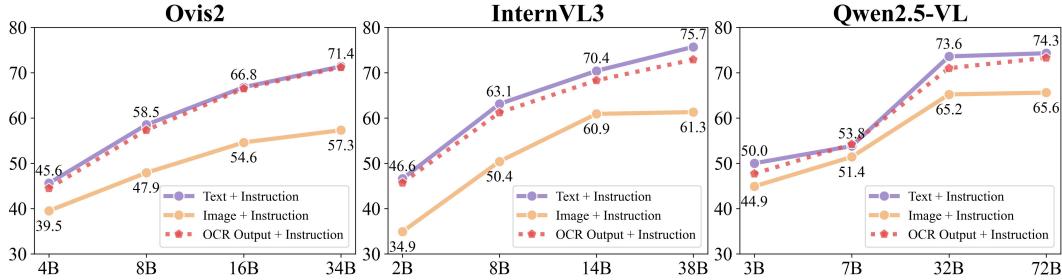


Figure 4: Impact of input modalities on LMMs under the CoT setting. These LMMs use Qwen2.5 [32] as the backbone. ‘Text + Instruction’ denotes text-only input, where the question is provided in the text. ‘Image + Instruction’ refers to multimodal input, with the question embedded in the image. ‘OCR Output + Instruction’ represents feeding LMM’s OCR-extracted text to themselves instead of ground-truth text. While this two-step strategy (prompt given in Appendix D) yields higher accuracy than direct multimodal input, it incurs significant inference overhead and contradicts the objective of end-to-end multimodal reasoning from raw visual inputs without task-specific priors.

Table 4: OCR performance of LMMs on LogicOCR. The used prompt is shown in Appendix C. The collected text corpus serves as ground truth for OCR. Results show strong OCR capabilities across models. The strong performance also highlights the high visual-text fidelity of GPT-Image-1.

Model	Edit Distance↓	F1-score↑	Precision↑	Recall↑	BLUE↑	METEOR↑
Ovis2-4B [59]	0.052	94.8	95.3	94.5	90.0	94.1
Ovis2-8B [59]	0.023	97.4	98.1	96.8	94.1	96.6
Ovis2-16B [59]	<u>0.020</u>	<u>97.6</u>	<u>98.2</u>	<u>97.1</u>	<u>94.5</u>	<u>96.9</u>
Ovis2-34B [59]	0.020	97.8	98.3	97.3	94.9	97.1
InternVL3-2B [10]	0.048	95.6	97.2	94.6	90.4	94.3
InternVL3-8B [10]	0.026	96.8	97.4	96.4	92.9	96.2
InternVL3-14B [10]	<u>0.021</u>	<u>96.9</u>	<u>97.5</u>	<u>96.5</u>	<u>93.3</u>	<u>96.3</u>
InternVL3-38B [10]	0.020	97.1	97.6	96.6	93.6	96.5
Qwen2.5-VL-3B [9]	0.041	96.5	98.1	95.5	92.1	94.9
Qwen2.5-VL-7B [9]	<u>0.032</u>	<u>97.0</u>	<u>98.0</u>	96.2	92.9	95.6
Qwen2.5-VL-32B [9]	0.034	96.8	96.6	97.0	93.4	96.6
Qwen2.5-VL-72B [9]	0.028	97.1	97.8	96.5	93.2	96.0

As shown in Fig. 4, all models exhibit a noticeable drop in reasoning performance when processing multimodal inputs. For instance, the Ovis2 series achieves, on average, 10.8 points higher accuracy with text-only input. InternVL3 shows a larger average gap of 12.1 across scales. In contrast, Qwen2.5-VL demonstrates better alignment between modalities, with an average gap of just 6.2, and only 2.4 for the Qwen2.5-VL-7B variant. This stronger modal consistency likely stems from more extensive vision-language pre-training [9]. Notably, the performance gap does not appear to result from deficiencies in visual recognition, as state-of-the-art models perform well in the OCR task (see Tab. 4). Moreover, in the text-only setting, using either ground-truth question context or OCR predictions from LMMs yields similar accuracy (see purple and red dashed lines in Fig. 4), suggesting that current LMMs have yet to fully integrate visual reading with high-level reasoning.

Moreover, a closer examination of model scaling reveals that both input modalities, text and multimodal, benefit from increased model capacity, primarily due to enhanced reasoning capabilities. For example, as model size increases from 8B to 38B parameters, InternVL3 improves multiple-choice accuracy by 12.6 points with text input and 10.9 points with multimodal input, while OCR performance remains largely unchanged (Tab. 4). However, when scaling from ~14B to ~30B parameters for Ovis2 and InternVL3, accuracy gains under multimodal input diminish, whereas improvements with text-only input remain notable. This suggests that aligning vision and language representations for multimodal reasoning remains a significant challenge at larger scales.

These findings highlight two key insights: First, **LMMs still fall short of fully integrating visual reading and reasoning**. Second, while vision-language alignment suffices for perception tasks like OCR, **it remains inadequate for more complex reasoning, especially as model size grows**. Therefore, achieving thorough vision-language alignment is vital for advancing multimodal reasoning in the future. While some works [66, 67] explore explicit alignment between visual and language tokens through contrastive learning and text mask segmentation, there remains an opportunity to

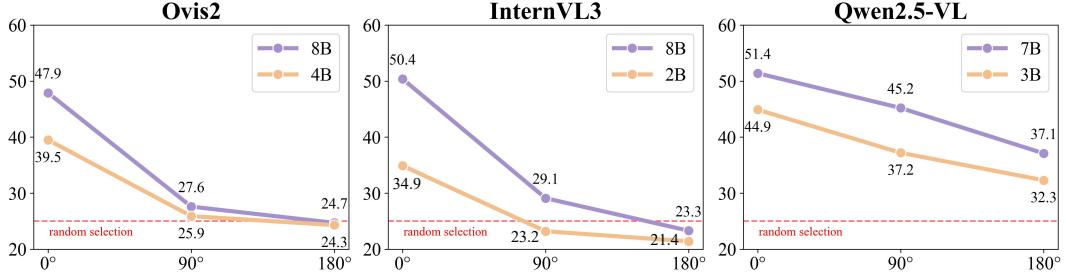


Figure 5: Impact of visual-text orientation on LMM performance. Images were rotated 90° and 180° clockwise to alter text orientation. Results show that state-of-the-art LMMs are sensitive to such changes, e.g., Ovis2 and InternVL2 accuracy drops to near-random levels, while Qwen2.5-VL demonstrates greater robustness.

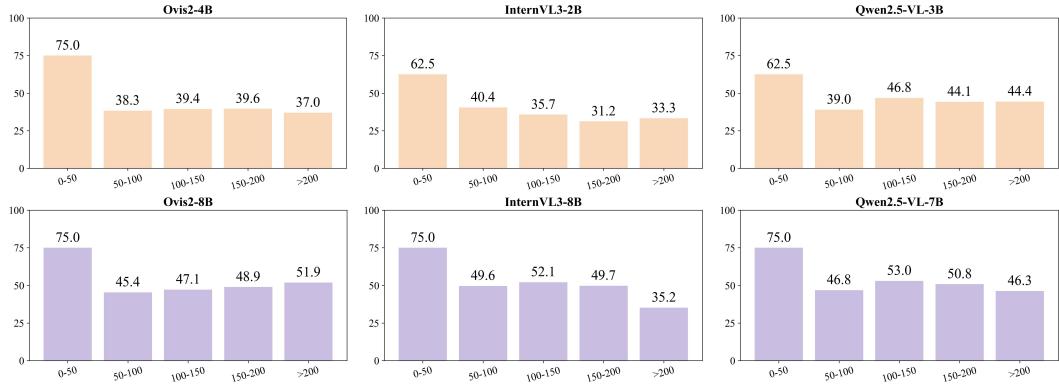


Figure 6: Impact of visual-text density on accuracy. Only 0.7% of images contain fewer than 50 words, 12.8% have 50–100 words, 49.2% contain 100–150 words, 32.4% have 150–200 words, and 4.9% include more than 200 words.

design scalable and tailored training objectives and datasets for reasoning tasks. In addition, perhaps we can explore an alignment feedback (e.g., performance gap between different modalities) to boost the self-improvement on vision-language integration.

4.3.3 Towards Robust LMMs Against Visual-Text Orientation Variations

We observe that recent state-of-the-art LMMs exhibit sensitivity to visual-text orientation. As illustrated in Fig. 5, rotating input images sharply degrades performance, e.g., Ovis2-4B and InternVL3-2B accuracy drops by 13.6 and 11.7 points, respectively, when rotated 90°. Larger models fare worse: Ovis2-8B and InternVL3-8B show declines of 20.3 and 21.3 points. In contrast, the Qwen2.5-VL series demonstrates greater robustness, maintaining stable performance even with 180° rotations.

In summary, **the perception robustness of LMMs across different visual-text orientations needs further improvement**. Qwen2.5-VL outperforms Ovis2 and InternVL3 in this evaluation, likely due to data augmentation strategies during training, such as image rotation, which is commonly employed to enhance the robustness of OCR and visual-text parsing models [68, 69, 70, 71]. To better improve LMMs’ robustness on varied visual-text orientations, introducing rotation-equivariant [72] image encoder may be promising.

4.3.4 The Impact of Visual-Text Density on Reasoning Performance

To investigate how visual-text density affects LMM performance, we categorize images based on their text density and report the accuracy for each group. As shown in Fig. 6, LMMs perform well on images with fewer than 50 words, achieving higher accuracy. However, their performance on the samples with high text density is not satisfactory. InternVL3-2B’s accuracy declines as text density rises. InternVL3-8B and Qwen2.5-VL-7B show similar trends, with both models reaching peak accuracy on images containing 100 to 150 words. In conclusion, **LMMs cannot achieve satisfactory reasoning performance on images with high text density**. To enhance the reasoning

performance on text-dense images, leveraging large models to automatically construct multi-hop reasoning data [73] based on existing easy samples is one of the feasible solutions.

4.3.5 Error Analysis

We analyze the error types of Qwen2.5-VL-72B during reasoning by categorizing them into six major types: 1) conceptual error, 2) logistic error, 3) argument structure error, 4) option analysis error, 5) information usage error, and 6) image reading error. These categories encompass 17 specific error types in total, though not all occur in every failure case. Definitions of each error type can be found in Appendix E. To aid in our analysis, we use o4-mini [34], with the prompt template for error analysis provided in Appendix E. Our analysis is based on 191 instances where Qwen2.5-VL-72B fails, but o4-mini gives correct answers.

As shown in Fig. 7, the main error categories in the reasoning process are information usage (68.5%), option analysis (24.0%), and logistic errors (4.7%). First, information usage errors occur when the model misinterprets the context, leading to an incorrect reasoning path. It often overlooks key information or requirements, failing to double-check if all given details and conditions are considered before reaching a conclusion. Second, option analysis errors are mainly due to conditional fallacy, where the model struggles to distinguish between necessary and sufficient conditions or gets confused by affirmative and negative forms. Additionally, some choices are selected without sufficient evidence. Finally, logistic errors include unreasonable hypotheses, overgeneralizations, and reversed causal relationships. In the future, several strategies can be explored to mitigate these errors and improve overall reasoning performance. For example, integrating retrieval-augmented models may help ensure that all relevant details are retained and revisited throughout the reasoning process. In addition, during training, we can categorize training instances by error type via annotations or model feedback, then dynamically resample training data to improve under-performing categories.

5 Conclusion

In this work, we present LogicOCR, a new benchmark designed to evaluate the logical reasoning abilities of Large Multimodal Models (LMMs) on text-rich images, with minimal reliance on domain-specific knowledge. We develop a scalable pipeline that leverages GPT-Image-1 to generate visually diverse, contextually grounded images from a curated text corpus, followed by manual verification to ensure quality. Our evaluation of various LMMs under both direct-answer and Chain-of-Thought (CoT) settings reveals that most models do not benefit from CoT prompting, suggesting potential weaknesses in their reasoning processes. While these models excel at OCR, their performance on multimodal reasoning lags behind their text-only counterparts, indicating a gap between visual understanding and logical inference. We also show that LMMs are sensitive to visual-text orientation and benefit from test-time scaling, highlighting important factors affecting multimodal reasoning.

Limitations. LogicOCR currently focuses on multiple-choice questions in English. Although LogicOCR has revealed some useful insights, it can be further extended to a multilingual version. With the exciting breakthrough of multilingual visual text generation [74, 75, 76, 77, 78, 31, 79], we believe the rendering quality of complex multilingual characters can be continuously improved. We can translate existing English samples into other languages or collect raw multilingual corpora to generate diverse images. In addition, our data generation method is also applicable to open-ended reasoning tasks while further efforts are required to collect the target corpora.

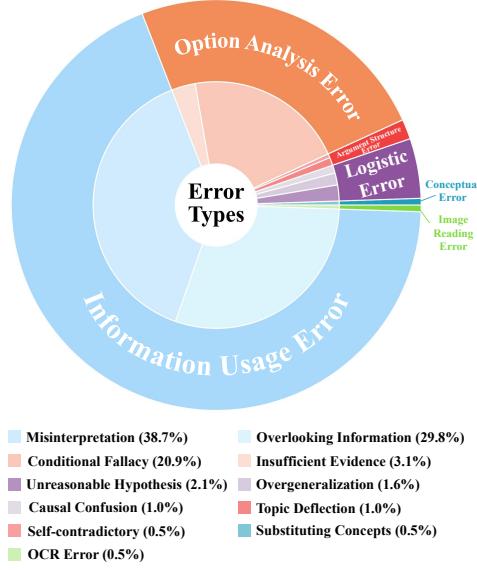


Figure 7: Error analysis for Qwen2.5-VL-72B reveals three main issues: misinterpretation, overlooking key information, and conditional fallacy.

References

- [1] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*, volume 37, pages 87310–87356, 2024.
- [2] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*, 2024.
- [3] OpenAI. Hello GPT-4o. <https://openai.com/index/gpt-4v-system-card>, 2024. Accessed: 2024-12-29.
- [4] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- [5] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.
- [6] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhushu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.
- [7] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.
- [8] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.
- [9] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [10] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.
- [11] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.
- [12] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, et al. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024.
- [13] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, pages 9556–9567, 2024.
- [14] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *ECCV*, pages 169–186, 2025.
- [15] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *NeurIPS*, 37:95095–95169, 2024.
- [16] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, et al. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019.
- [17] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, pages 2200–2209, 2021.
- [18] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, pages 1697–1706, 2022.

- [19] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of ACL*, pages 2263–2279, 2022.
- [20] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, et al. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024.
- [21] Chenglong Liu, Haoran Wei, Jinyue Chen, Lingyu Kong, Zheng Ge, Zining Zhu, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Focus anywhere for fine-grained multi-page document understanding. *arXiv preprint arXiv:2405.14295*, 2024.
- [22] Shuo Zhang, Biao Yang, Zhang Li, Zhiyin Ma, Yuliang Liu, and Xiang Bai. Exploring the capabilities of large multimodal models on dense text. In *ICDAR*, pages 281–298, 2024.
- [23] Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, et al. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*, 2024.
- [24] Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, et al. Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy. *arXiv preprint arXiv:2412.02210*, 2024.
- [25] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024.
- [26] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *NeurIPS*, 37:113569–113697, 2024.
- [27] Ling Fu, Biao Yang, Zhebin Kuang, Jiajun Song, Yuzhe Li, et al. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *arXiv preprint arXiv:2501.00321*, 2024.
- [28] Patrick J. Hurley. A concise introduction to logic. *Nelson Education*, 2014.
- [29] Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In *IJCAI*, 2020.
- [30] Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. Logiqa 2.0—an improved dataset for logical reasoning in natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2947–2962, 2023.
- [31] OpenAI. Introducing our latest image generation model in the API. <https://openai.com/index/image-generation-api/>, 2025. Accessed: 2025-04-23.
- [32] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- [33] DeepMind. Gemini 2.5 Pro. <https://deepmind.google/technologies/gemini/pro/>, 2025.
- [34] OpenAI. OpenAI o3 and o4-mini System Card. <https://openai.com/index/o3-o4-mini-system-card/>, 2025. Accessed: 2025-04-16.
- [35] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024.
- [36] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, et al. Monkey: Image resolution and text label are important things for large multi-modal models. In *CVPR*, pages 26763–26773, 2024.
- [37] Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, et al. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. *arXiv preprint arXiv:2409.03420*, 2024.
- [38] Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. *arXiv preprint arXiv:2403.03003*, 2024.
- [39] Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, et al. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*, 2024.

- [40] Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *ECCV*, pages 390–406, 2024.
- [41] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, et al. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *NeurIPS*, 37:42566–42592, 2024.
- [42] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024.
- [43] OpenAI. OpenAI o1 System Card. <https://openai.com/index/openai-o1-system-card/>, 2024. Accessed: 2024-12-05.
- [44] Qwen Team. Qvq: To see the world with wisdom, December 2024.
- [45] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [46] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.
- [47] Huanqia Cai, Yijun Yang, and Winston Hu. Mm-iq: Benchmarking human-like abstraction and reasoning in multimodal models. *arXiv preprint arXiv:2502.00698*, 2025.
- [48] Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge. *arXiv preprint arXiv:2504.10342*, 2025.
- [49] Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, et al. Chartx & chartvilm: A versatile benchmark and foundation model for complicated chart reasoning. *arXiv preprint arXiv:2402.12185*, 2024.
- [50] Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. Chartbench: A benchmark for complex visual reasoning in charts. *arXiv preprint arXiv:2312.15915*, 2023.
- [51] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024.
- [52] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *CVPR*, pages 2315–2324, 2016.
- [53] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *ECCV*, pages 498–517. Springer, 2022.
- [54] Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, and Michalis Raptis. Towards end-to-end unified scene text detection and layout analysis. In *CVPR*, pages 1049–1059, 2022.
- [55] Anthropic. Claude 3.7 Sonnet and Claude Code. <https://www.anthropic.com/news/claude-3-7-sonnet>, 2025. Accessed: 2025-02-25.
- [56] Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, et al. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024.
- [57] Mingxin Huang, Yuliang Liu, Dingkang Liang, Lianwen Jin, and Xiang Bai. Mini-monkey: Multi-scale adaptive cropping for multimodal large language models. In *ICLR*, 2025.
- [58] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [59] Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv:2405.20797*, 2024.

- [60] Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint arXiv:2409.12183*, 2024.
- [61] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, et al. Seed1.5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.
- [62] Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in llms. *arXiv preprint arXiv:2502.03373*, 2025.
- [63] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.
- [64] Sicheng Feng, Gongfan Fang, Xinyin Ma, and Xinchao Wang. Efficient reasoning models: A survey. *arXiv preprint arXiv:2504.10903*, 2025.
- [65] Yifu Ding, Wentao Jiang, Shunyu Liu, Yongcheng Jing, Jinyang Guo, Yingjie Wang, Jing Zhang, Zengmao Wang, Ziwei Liu, Bo Du, et al. Dynamic parallel tree search for efficient llm reasoning. *arXiv preprint arXiv:2502.16235*, 2025.
- [66] Zining Wang, Tongkun Guan, Pei Fu, Chen Duan, Qianyi Jiang, Zhentao Guo, Shan Guo, Junfeng Luo, Wei Shen, and Xiaokang Yang. Marten: Visual question answering with mask generation for multi-modal document understanding. In *CVPR*, 2025.
- [67] Tongkun Guan, Zining Wang, Pei Fu, Zhengtao Guo, Wei Shen, Kai Zhou, Tiezhu Yue, Chen Duan, Hao Sun, Qianyi Jiang, et al. A token-level text image foundation model for document understanding. *arXiv preprint arXiv:2503.02304*, 2025.
- [68] Jianqiang Wan, Sibo Song, Wenwen Yu, Yuliang Liu, Wenqing Cheng, Fei Huang, Xiang Bai, Cong Yao, and Zhibo Yang. Omniparser: A unified framework for text spotting key information extraction and table recognition. In *CVPR*, pages 15641–15653, 2024.
- [69] Wenwen Yu, Zhibo Yang, Jianqiang Wan, Sibo Song, Jun Tang, Wenqing Cheng, Yuliang Liu, and Xiang Bai. Omniparser v2: Structured-points-of-thought for unified visual text parsing and its generality to multimodal large language models. *arXiv preprint arXiv:2502.16161*, 2025.
- [70] Mingxin Huang, Jiaxin Zhang, Dezheng Peng, Hao Lu, Can Huang, Yuliang Liu, Xiang Bai, and Lianwen Jin. Estextspotter: Towards better scene text spotting with explicit synergy in transformer. In *ICCV*, pages 19495–19505, 2023.
- [71] Maoyuan Ye, Jing Zhang, Shanshan Zhao, Juhua Liu, Tongliang Liu, Bo Du, and Dacheng Tao. Deep solo: Let transformer decoder with explicit points solo for text spotting. In *CVPR*, pages 19348–19357, 2023.
- [72] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: A rotation-equivariant detector for aerial object detection. In *CVPR*, pages 2786–2795, 2021.
- [73] Ryota Tanaka, Taichi Iki, Taku Hasegawa, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. Vdocrag: Retrieval-augmented generation over visually-rich documents. In *CVPR*, 2025.
- [74] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. In *NeurIPS*, volume 36, pages 9353–9387, 2023.
- [75] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. In *ECCV*, pages 386–402, 2024.
- [76] Yukang Yang, Dongnan Gui, Yuhui Yuan, Weicong Liang, Haisong Ding, Han Hu, and Kai Chen. Glyphcontrol: glyph conditional control for visual text generation. In *NeurIPS*, volume 36, pages 44050–44066, 2023.
- [77] Zeyu Liu, Weicong Liang, Yiming Zhao, Bohan Chen, Lin Liang, Lijuan Wang, Ji Li, and Yuhui Yuan. Glyph-byt5-v2: A strong aesthetic baseline for accurate multilingual visual text rendering. *arXiv preprint arXiv:2406.10208*, 2024.
- [78] Zeyu Liu, Weicong Liang, Zhanhao Liang, Chong Luo, Ji Li, Gao Huang, and Yuhui Yuan. Glyph-byt5: A customized text encoder for accurate visual text rendering. In *ECCV*, pages 361–377, 2024.
- [79] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. In *ICLR*, 2024.

A Prompt Templates for Text-to-Image Generation with GPT-Image-1

The four prompt templates for instructing GPT-Image-1 are provided here. Note that {context}, {question}, and {options} are place-holders for the context, question, and options parts of each sample, respectively. And {depiction} is a place-holder for the background depiction generated by Qwen2.5.

Interleaved Style Layout & Non-Handwritten Style Font

Generate an image about random color paper with smallest font-size. Firstly, it is written about context information: "{context}"
An illustration figure or scene described by the above context is shown.
Then, the image displays a question: "{question}"
Finally, four choices are written on the image: "{options}"
Do not summarize the visual text content given above.

Interleaved Style Layout & Handwritten Style Font

Generate an image about random color paper with smallest font-size. Firstly, it is written about context information in handwritten style: "{context}"
An illustration figure or scene described by the above context is shown.
Then, the image displays a question in handwritten style: "{question}"
Finally, four choices in handwritten style are written on the image: "{options}"
Do not summarize the visual text content given above.

Background Style Layout & Non-Handwritten Style Font

Generate an image with smallest font-size. {depiction} Some text paragraphs with contrastive color are shown. Specifically, firstly, it is written about context information: "{context}"
Then, the image displays a question: "{question}"
Finally, four choices are written on the image: "{options}"
Do not summarize the visual text content given above.

Background Style Layout & Handwritten Style Font

Generate an image with smallest font-size. {depiction} Some text paragraphs with handwritten style and contrastive color are shown. Specifically, firstly, it is written about context information: "{context}"
Then, the image displays a question: "{question}"
Finally, four choices are written on the image: "{options}"
Do not summarize the visual text content given above.

B Prompt Templates for Evaluating LMMs

Prompt for Multimodal Input under CoT setting

Solve the multiple-choice question in image and then answer with one option letter. The last line of your response should be of the following format: 'Answer: LETTER' where LETTER is one of options. Think step by step before answering.

Prompt for Multimodal Input under Direct Answering Setting

Solve the multiple-choice question in image. Directly answer the question with one option letter without explanation.

Prompt for Pure Text Input under CoT setting

Question: {context} {question}
Options:
{options}

Solve the multiple-choice question and then answer with one option letter. The last line of your response should be of the following format: 'Answer: LETTER' where LETTER is one of options. Think step by step before answering.

Prompt for Pure Text Input under Direct Answering Setting

Question: {context} {question}

Options:

{options}

Directly answer the question with one option letter without explanation.

C Prompt Template for Evaluating OCR Performance on LogicOCR

The OCR Prompt for LMMs

Please recognize the text paragraphs in the image. Do not add explanation. Do not answer the question in this image.

D Prompt Template for Evaluating LMMs with Their OCR Results as Input

In this evaluation, the image is not provided to LMMs. The OCR results are achieved in advance by using the prompt in Appendix C and then inserted in the place-holder '{OCR results}', resulting in a two-step solution. Note that state-of-the-art LMMs have been equipped with strong OCR performance in order to eliminate the need for an explicit OCR step. However, we observe that end-to-end multimodal reasoning underperforms this time-consuming two-step strategy on LogicOCR, suggesting that state-of-the-art LMMs have yet to fully bridge visual reading with reasoning. This two-step strategy is actually impractical for real-world applications due to its substantial inference latency and the restricted assumption that an image can be represented by sole OCR results. Through this evaluation, we want to reveal one of the key bottlenecks in multimodal reasoning, rather than encouraging this two-step strategy.

Prompt for Pure Text Input under CoT Setting.

{OCR results}

Solve the above multiple-choice question and then answer with one option letter. The last line of your response should be of the following format: 'Answer: LETTER' where LETTER is one of options. Think step by step before answering.

E The Prompt Template for Error Analysis with o4-mini

In the prompt, {corpus}, {solution}, and {response} are place-holders for the text corpus of each question, the correct option's letter, and the response of evaluated model.

Prompt for Error Analysis

Here is a multiple-choice question written on image:

{corpus}

The correct choice is {solution}. After reading the image, one AI model got a wrong answer process:
{response}

Please carefully analyze why the process is wrong and choose one most appropriate error from the following 17 types. Answer the error type of the following format "error type: chosen-error", such as "error type: 6. overgeneralization".

about conceptual error

1. substituting concepts: Failure to maintain conceptual consistency within the same reasoning.

2. improper juxtaposition: Confusing classification standards.

3. circular definition: The defining term directly or indirectly includes the defined term, such as "an optimist is an optimistic person."

about logistic error

4. unreasonable hypothesis: Adding subjective assumptions or reverse reasoning (e.g., from "smokers are in poor health" to "non-smokers are in good health").

5. exaggeration: Expressing possibility as certainty, such as exaggerating "may be extinct" to "already extinct".

6. overgeneralization: Using local phenomena to infer the whole, such as using "14 percent people like Peking Opera" to infer "general lack of traditional culture".
7. causal confusion: Reversing or imposing causal relationships, such as mistaking "low immune system causes psychological problems" for "psychological problems lower immunity."
- ## about argument structure error
8. topic deflection: Deviating from the original discussion focus.
9. self-contradictory: Affirming contradictory propositions at the same time, such as "entangled at all times" and "temporarily put aside".
10. equivocating: There is no clear statement on right and wrong issues, such as "neither comprehensive nor one-sided".
11. circular argument: The argument relies on the premise itself, such as "lying is treasonous, therefore you are a traitor."
- ## about option analysis error
12. overstatement: It exceeds the reasonable scope of the question, such as inferring "may" as "certain".
13. conditional fallacy: Can't distinguish between necessary and sufficient conditions, or confuse affirmative and negative forms.
14. insufficient evidence: The options lack support from the question or the information is one-sided.
- ## about information usage error
15. misinterpretation: The key words in the question (such as "most different") are not captured.
16. overlooking information: Omission of key information in the material leads to misjudgment.
- ## about image reading error
17. OCR error: The wrong optical character recognition results from the image affect the reasoning process.

F Prompt Reference for Determining Layout Style

Prompt for Choosing Layout Style

Imagine you are designing an image. Given the context information between <context> and </context> tags. Is it more suitable to draw a full frame background scene or an illustration figure? Answer "background" or "illustration".

<context>{context}</context>

G Prompt Reference for Background Depiction

Prompt for Background Depiction

You are designing the background scene for an image based on the context information given between <context> and </context> tags, please generate a short paragraph of caption in pure English for the background scene.

<context>{context}</context>

H Broader Impacts

This work contributes the LogicOCR benchmark that highlights critical limitations in current LMMs with respect to complex logical reasoning over text-rich images. Our evaluation of various LMMs under both direct-answer and CoT settings reveals that most models do not benefit from CoT prompting, suggesting potential shortcomings in their reasoning processes. While state-of-the-art LMMs excel at OCR, their performance on multimodal reasoning lags behind their text-only counterparts, indicating a gap between visual reading and logical reasoning. Additionally, our analysis shows that LMMs' performance is sensitive to visual-text orientation. Identifying these key flaws is crucial, especially in document understanding tasks and high-stakes applications such as medical and healthcare domains, where erroneous reasoning can lead to significant consequences. By revealing these limitations, LogicOCR provides a valuable source for guiding future improvements in multimodal reasoning. As an evaluation benchmark, there is no direct negative societal impacts.

I Datasheet for LogicOCR

I.1 Motivation

1. For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

A1: Existing multimodal reasoning datasets often require extensive mathematical or scientific knowledge, making it difficult to isolate pure reasoning ability from domain expertise. In contrast, most OCR-related benchmarks lack complexity and some of them only focus on narrow topics. LogicOCR was created to evaluate LMMs' complex logical reasoning capabilities on text-rich images while minimizing reliance on domain knowledge.

2. Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

A2: This dataset is created by the authors of this paper.

3. Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

A3: N/A.

I.2 Composition

1. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances(e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

A1: Each multimodal sample in LogicOCR consists of one image and a templated instruction. Each image contains a multiple-choice question. The images have diverse layouts (background style and text-illustration interleaved) and font style (handwritten and standard).

2. How many instances are there in total (of each type, if appropriate)?

A2: LogicOCR comprises 1,100 instances.

3. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

A3: The collected text corpora source from the Chinese National Civil Servant Examination, *i.e.*, the larger set. Although the latest questions in Chinese National Civil Servant Examination are not included, LogicOCR is still representative, covering categorical, sufficient conditional, necessary conditional, disjunctive, and conjunctive reasoning types.

4. What data does each instance consist of? “Raw” data (e.g., unprocessed text or images)or features? In either case, please provide a description.

A4: Each instance contains an image and a templated instruction. A multiple-choice question is rendered on each image. The question corpora source from the Chinese National Civil Servant Examination. GPT-Image-1 is leveraged to generate diverse images based on the curated corpora.

5. Is there a label or target associated with each instance? If so, please provide a description.

A5: Yes, each instance contains exactly one target as the true answer corresponding to the question.

6. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

A6: Very few images miss the last option's content. The missing option is cut out by GPT-Image-1 during text-to-image generation. We manually verify that the missing option is not the correct answer and does not affect problem-solving.

7. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

A7: No, there is no explicit relationship between each instance.

8. Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

A8: Yes, we recommend using all instances of LogicOCR for testing.

9. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

A9: The text corpora are sourced from the Chinese National Civil Servant Examination and have been curated by LogiQA [29] and LogiQA2.0 [30]. We reuse the corpora from LogiQA and LogiQA2.0, then further conduct deduplication. There are no redundancies in LogicOCR.

10. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

A10: The dataset is self-contained. The official archival versions of the complete dataset can be downloaded from GitHub. There are no fees. The data should be used in non-commercial scenarios following CC BY-NC-SA 4.0 license.

11. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

A11: No.

12. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

A12: No.

I.3 Collection Process

1. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

A1: The data associated with each instance is directly observable.

2. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

A2: The text corpora are reused from LogiQA and LogiQA2.0. GPT-Image-1 API is used from transfer the text corpora in to diverse images. The images are manually verified.

3. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

A3: Cosine similarity of the sentence embeddings is used for text corpora deduplication. Then, the generated images based on the text corpora are filtered according to subjective quality assessment.

4. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

A4: Only the authors of this paper.

5. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

A5: The timeframe for the generated images is from April to May 2025.

I.4 Preprocessing/cleaning/labeling

1. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

A1: For individual samples with mixed languages in the collected text corpora, we retranslate the original Chinese corpus into English with Qwen2.5 [32]. The text corpus of each sample can be divided into three parts, including *context*, *question*, and *options*. The context part provides all the information required to solve the question. Some different questions share the same context corpus. We use the context part of each sample for further deduplication. We first convert each context corpus into one embedding using BGE-M3 [51] and then calculate the cosine similarity for deduplication.

2. Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.

A2: The raw text corpora can be found in [LogiQA](#) and [LogiQA2.0](#)

3. Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

A3: Link for [Qwen2.5](#) and [BGE-M3](#).

I.5 Uses

1. Has the dataset been used for any tasks already? If so, please provide a description.

A1: No.

2. Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

A2: No.

3. What (other) tasks could the dataset be used for?

A3: It can be used for testing and evaluating the LMMs’ pure logical reasoning performance on text-rich images, while minimizing the reliance on domain knowledge.

4. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

A4: No.

5. Are there tasks for which the dataset should not be used? If so, please provide a description.

A5: The dataset is a test set. It should not be used for training.

I.6 Distribution

1. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

A1: Yes.

2. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

A2: It will be publicly available at [LogicOCR](#).

3. When will the dataset be distributed?

A3: The dataset will be distributed in May 2025.

4. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

A4: It will be distributed under the [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License](#).

5. Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

A5: No.

6. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

A6: No.

I.7 Maintenance

1. Who will be supporting/hosting/maintaining the dataset?

A1: The authors.

2. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

A2: They can be contacted via email available on the paper.

3. Is there an erratum? If so, please provide a link or other access point.

A3: No.

4. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

A4: Yes. It will be updated by the authors of the paper and released on GitHub.

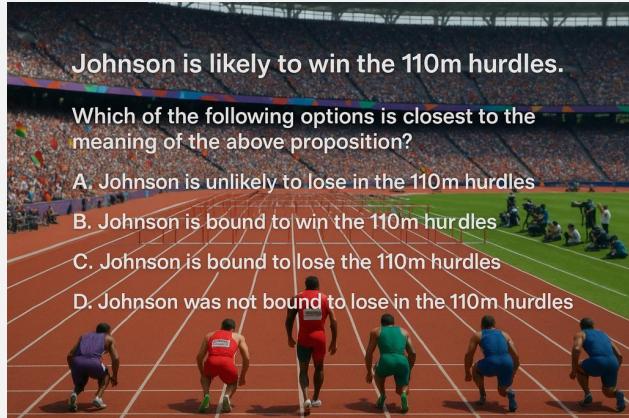
5. Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

A5: Yes. We will also provide the download link of the older versions.

6. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

A6: Others can contribute through GitHub. The contributions will be verified.

J Case Study



Johnson is likely to win the 110m hurdles.

Which of the following options is closest to the meaning of the above proposition?

- A. Johnson is unlikely to lose in the 110m hurdles
- B. Johnson is bound to win the 110m hurdles
- C. Johnson is bound to lose the 110m hurdles
- D. Johnson was not bound to lose in the 110m hurdles

Layout: background style

Font style: standard

Reasoning types:

- 1. sufficient conditional

Solution: D



Qwen2.5-VL-72B:

To determine which option is closest to the meaning of the proposition "Johnson is likely to win the 110m hurdles," let's analyze each option:

A. Johnson is unlikely to lose in the 110m hurdles.

- This option suggests that it is improbable for Johnson to lose, which aligns closely with the idea that he is likely to win.

B. Johnson is bound to win the 110m hurdles.

- This option implies certainty, which is stronger than just being likely. The original statement does not guarantee a win, only a high probability.

C. Johnson is bound to lose the 110m hurdles.

- This option contradicts the original statement, as it suggests certainty of losing rather than likelihood of winning.

D. Johnson was not bound to lose in the 110m hurdles.

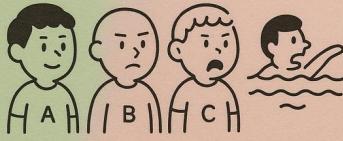
- This option is somewhat ambiguous but generally means that losing was not certain. While this is related, it doesn't directly convey the likelihood of winning as strongly as option A.

After evaluating all options, the one that most closely matches the meaning of the original statement is:

Answer: A 

Figure 8: A failure case of Qwen2.5-VL-72B on the question involving one reasoning type.

Among the three of A, B, and C, only one can swim.
 A said: I can, B said: I can't, and C said: A can't.



If only one of these three sentences is true, then the one who can swim is

A. A
 B. B
 C. C
 D. Can't judge



Qwen2.5-VL-72B:

To solve this problem, we need to determine who can swim based on the given statements and the condition that only one of the three sentences is true.

Let's analyze each possibility:

1. **Assume A can swim:**

- A said: "I can" (True)
- B said: "I can't" (True)
- C said: "A can't" (False)

This results in two true statements, which contradicts the condition that only one statement is true. Therefore, A cannot be the one who can swim.

2. **Assume B can swim:**

- A said: "I can" (False)
- B said: "I can't" (False)
- C said: "A can't" (True)

This results in one true statement, which satisfies the condition. Therefore, B could be the one who can swim.

3. **Assume C can swim:**

- A said: "I can" (False)
- B said: "I can't" (True)
- C said: "A can't" (True)

This results in two true statements, which contradicts the condition that only one statement is true. Therefore, C cannot be the one who can swim.

From the analysis, the only consistent scenario is when B can swim.

Answer: B

Figure 9: A correct case of Qwen2.5-VL-72B on the question involving one reasoning type.



Qwen2.5-VL-72B:

This summer, Lao Li's three children took part in three sports competitions held in Hangzhou, Kunming and Harbin. Men's gymnastics; pole vault and marathon. What is known is that the eldest did not go to Harbin to participate in the marathon; the third did not participate in the events held in Hangzhou; and the child who won the national championship in gymnastics did not go to Kunming, the winner of the marathon was not his third.



Which of the following is the national champion of men's gymnastics?

- A. The eldest
- B. The second
- C. The third
- D. Can't judge

Layout: interleaved style

Font style: handwritten

Reasoning types:

- 1. sufficient conditional
- 2. conjunctive

Solution: A

To solve this problem, let's analyze the given information step by step:

1. **Eldest did not go to Harbin for the marathon:**
 - This means the eldest participated in either Hangzhou or Kunming.
2. **Third did not participate in the events held in Hangzhou:**
 - This means the third participated in either Kunming or Harbin.
3. **Child who won the national championship in gymnastics did not go to Kunming:**
 - This means the gymnastics champion participated in either Hangzhou or Harbin.
4. **Winner of the marathon was not his third:**
 - This means the third child did not win the marathon.

Let's summarize the possible scenarios based on these clues:

- If the eldest went to Hangzhou, he could have participated in gymnastics or pole vault.
- If the eldest went to Kunming, he could have participated in gymnastics or pole vault.
- The third cannot be in Hangzhou, so he must be in Kunming or Harbin.
- The gymnastics champion cannot be in Kunming, so he must be in Hangzhou or Harbin.
- The third cannot be the marathon winner, so he must be in gymnastics or pole vault.

Given these constraints, let's deduce further:

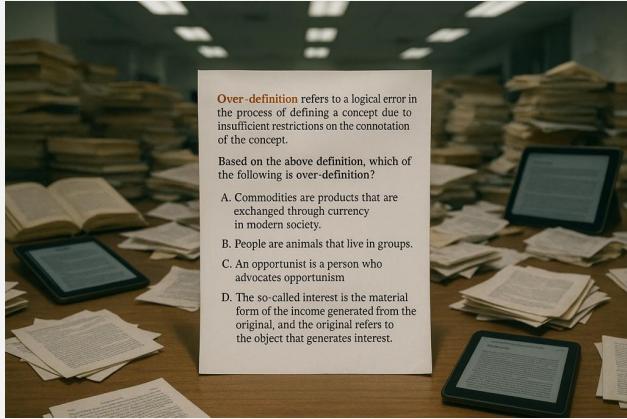
- If the third is in Kunming, he cannot be the gymnastics champion (since the champion is not in Kunming).
- If the third is in Harbin, he cannot be the marathon winner (since the winner is not the third).

Considering all these points, the only consistent scenario is that the second child must be the gymnastics champion because:

- The eldest cannot be in Kunming for gymnastics.
- The third cannot be in Kunming for gymnastics.
- Therefore, the second must be the gymnastics champion.

Answer: B

Figure 10: A failure case of Qwen2.5-VL-72B on the question involving two reasoning types.



Layout: background style

Font style: standard

Reasoning types:

1. sufficient conditional
2. conjunctive

Solution: B



Qwen2.5-VL-72B:

To determine which option represents an over-definition, we need to identify the one that provides insufficient restrictions on the connotation of the concept being defined.

Let's analyze each option:

- A. "Commodities are products that are exchanged through currency in modern society."
- This definition restricts commodities to products exchanged through currency in modern society, which seems specific enough and does not appear to be overly broad.
- B. "People are animals that live in groups."
- This definition is too broad because it includes all animals that live in groups, not just humans. It lacks sufficient restriction to specifically define people.
- C. "An opportunist is a person who advocates opportunism."
- This definition is circular and does not provide additional information or restriction beyond the term itself. However, it doesn't necessarily broaden the concept excessively.
- D. "The so-called interest is the material form of the income generated from the original, and the original refers to the object that generates interest."
- This definition is somewhat redundant but does not appear to be overly broad in its restrictions.

Based on the analysis, option B is the best example of an over-definition because it defines "people" too broadly as any animals that live in groups, without sufficient restriction to specifically define humans.

Answer: B ✓

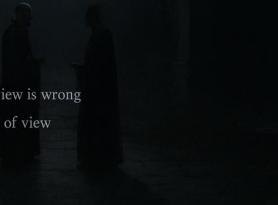
Figure 11: A correct case of Qwen2.5-VL-72B on the question involving two reasoning types.

Wu Mazi said to Mozi: I can't love both. I love my family more than the people in my hometown, my parents more than my family, and I love myself more than my parents. This is because I am close to me, if you hit me, I will hurt others, but it won't hurt me, so I will only kill others for my benefit. Mozi asked: will you tell others this righteousness? Wu Mazi replied: why should I hide? I will tell others.

Mozi said: since you are like this, if someone likes your idea, then this person will kill you for his own benefit. If someone doesn't like your idea, he will also kill you, because he thinks you are a person who spreads ominous words

It can be inferred from the above:

- A. Wu Mazi can only go far away
- B. Wu Mazi must admit that her point of view is wrong
- C. Wu Mazi must change his original point of view
- D. Wu Mazi will be killed




Qwen2.5-VL-72B:

To solve this multiple-choice question, let's analyze the given text and the options provided:

1. ****Understanding the Context**:**

- Wu Mazi expresses his belief that he loves himself more than his parents, his parents more than his family, and his family more than the people in his hometown.
- He believes in acting for his own benefit, even if it means hurting others.
- Mozi warns him that if he shares this belief, people who like or dislike his idea will both pose a threat to him.

2. ****Analyzing the Options**:**

- **A. Wu Mazi can only go far away**: This option suggests that Wu Mazi must leave to avoid danger. However, the text does not explicitly state that going away is the only solution.
- **B. Wu Mazi must admit that her point of view is wrong**: The text does not indicate that Wu Mazi must admit his viewpoint is incorrect. It only discusses the potential dangers of sharing his beliefs.
- **C. Wu Mazi must change his original point of view**: There is no indication in the text that Wu Mazi needs to change his beliefs. The focus is on the consequences of sharing them.
- **D. Wu Mazi will be killed**: The text implies that Wu Mazi will face danger from others regardless of whether they agree or disagree with him. This suggests a high likelihood of being harmed or killed.

3. ****Conclusion**:**

- Based on the analysis, the most accurate inference from the text is that Wu Mazi will be killed due to the dangers posed by both those who agree and disagree with his beliefs.

Answer: D

Layout: background style

Font style: standard

Reasoning types:

1. categorical
2. sufficient conditional
3. conjunctive

Solution: B

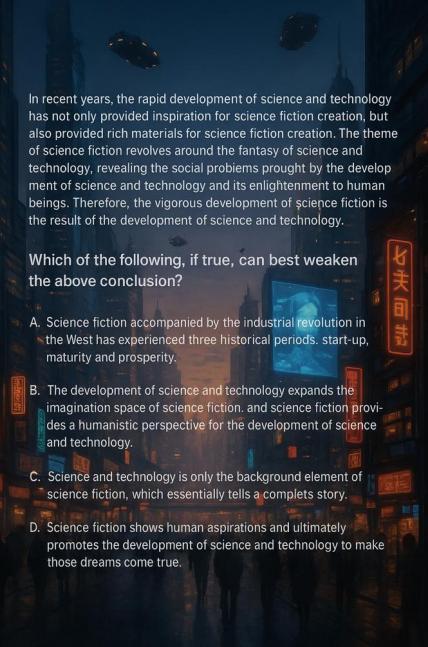
Figure 12: A failure case of Qwen2.5-VL-72B on the question involving three reasoning types.



Qwen2.5-VL-72B:

To solve this problem, we need to identify which option best weakens the conclusion that the vigorous development of science fiction is primarily due to the development of science and technology.

Let's analyze each option:



In recent years, the rapid development of science and technology has not only provided inspiration for science fiction creation, but also provided rich materials for science fiction creation. The theme of science fiction revolves around the fantasy of science and technology, revealing the social problems brought by the development of science and technology and its enlightenment to human beings. Therefore, the vigorous development of science fiction is the result of the development of science and technology.

Which of the following, if true, can best weaken the above conclusion?

- A. Science fiction accompanied by the industrial revolution in the West has experienced three historical periods: start-up, maturity and prosperity.
- B. The development of science and technology expands the imagination space of science fiction, and science fiction provides a humanistic perspective for the development of science and technology.
- C. Science and technology is only the background element of science fiction, which essentially tells a complete story.
- D. Science fiction shows human aspirations and ultimately promotes the development of science and technology to make those dreams come true.

Layout: background style

Font style: standard

Reasoning types:

1. sufficient conditional
2. necessary conditional
3. conjunctive

Solution: C

B. "The development of science and technology expands the imagination space of science fiction, and science fiction provides a humanistic perspective for the development of science and technology."

- This statement supports the conclusion by reinforcing the interplay between science/technology and science fiction, suggesting they mutually influence each other.

C. "Science and technology is only the background element of science fiction, which essentially tells a complete story."

- This statement suggests that science and technology are not central to the development of science fiction but rather serve as a backdrop. If true, it would weaken the conclusion that the development of science fiction is primarily driven by advancements in science and technology.

D. "Science fiction shows human aspirations and ultimately promotes the development of science and technology to make those dreams come true."

- This statement supports the conclusion by suggesting that science fiction influences the direction of scientific and technological progress.

Given the analysis, Option C is the one that most effectively weakens the conclusion by downplaying the role of science and technology in the development of science fiction.

Answer: C

Figure 13: A correct case of Qwen2.5-VL-72B on the question involving three reasoning types.

We-media refers to the way and instant transmission methods that an ordinary citizen or organization that can access the Internet at any time, any place, and in any way, connect with the global knowledge system through modern digital technology, and provide or share their true views and their own news.

According to the above definition, which of the following acts does not involve We-media?

- Professor Zhang posted his courseware on the internet for students to download.
- Li discussed the food safety issue with netizens on Weibo.
- Xiao Zhang emailed his Chinese style crossing the road to netizens.
- Lao Liu browses news about Spring Festival travel on the forum.



Qwen2.5-VL-72B:

To determine which act does not involve We-media, let's analyze each option based on the definition provided:

- **We-media Definition**:** It involves ordinary citizens or organizations using the Internet to connect with the global knowledge system, share true views, and provide their own news through modern digital technology.
- **Option A**:** "Professor Zhang posted his courseware on the internet for students to download."
 - This involves an individual (Professor Zhang) sharing information (courseware) via the Internet. This fits the definition of We-media as it involves providing content online.
- **Option B**:** "Li discussed the food safety issue with netizens on Weibo."
 - This involves an individual (Li) discussing a topic with others (netizens) on a social media platform (Weibo). This also fits the definition of We-media as it involves sharing views and connecting with others online.
- **Option C**:** "Xiao Zhang emailed his Chinese style crossing the road to netizens."
 - This involves sending an email, which is a direct communication method rather than a public sharing platform. Emailing does not necessarily involve connecting with the global knowledge system in the same way as posting on a forum or social media. Therefore, this may not fit the We-media definition as closely as the other options.
- **Option D**:** "Lao Liu browses news about Spring Festival travel on the forum."
 - This involves an individual (Lao Liu) accessing information (news) on a forum. While browsing news doesn't directly involve sharing content, forums are platforms where users can engage and share views, fitting the We-media context.

Based on the analysis, Option C stands out as the least fitting example of We-media because it involves private communication (email) rather than public sharing or discussion on a digital platform.

Answer: C X

Layout: background style

Font style: handwritten

Reasoning types:

- categorical
- sufficient conditional
- disjunctive
- conjunctive

Solution: D

Figure 14: A failure case of Qwen2.5-VL-72B on the question involving four reasoning types.

Professor Zhang: the prehistoric essential tools found in South America existed 13,000 years ago. Some archaeologists believe that these tools were used by people whose ancestors migrated from Siberia to Alaska. This view is untenable. In order to reach South America, these people had to trek 13,000 years ago, and 13,000-year-old wooden tools were never found between Alaska and South America.

Researcher Li: I'm afraid you've overlooked: these wooden tools are found in peat bogs, which are rarely found in North America. Wooden tools rot and dissolve in ordinary soil within a few years.

Which of the following most accurately summarizes the coping methods of researcher Li?

- It is pointed out that Professor Zhang's argument is against the facts.
- Quote authoritative research results that are at odds with Professor Zhang conclusion.
- It is pointed out that Professor Zhang misinterpreted the views of archaeologists.
- Question Professor Zhang's implicit hypothesis.

Layout: interleaved style

Font style: handwritten

Reasoning types:

- categorical
- sufficient conditional
- necessary conditional
- conjunctive

Solution: D



Qwen2.5-VL-72B:

To solve this problem, we need to carefully analyze the arguments presented by Professor Zhang and Researcher Li and determine which option best summarizes Researcher Li's coping method.

1. **Professor Zhang's Argument**:

- He claims that the prehistoric essential tools found in South America existed 13,000 years ago.
- He believes these tools were used by people whose ancestors migrated from Siberia to Alaska.
- He argues that this view is untenable because if these people had trekked 13,000 years ago, wooden tools should have been found between Alaska and South America, but they weren't.

2. **Researcher Li's Response**:

- He points out that Professor Zhang overlooked the fact that these wooden tools are found in peat bogs, which are rare in North America.
- He explains that wooden tools rot and dissolve in ordinary soil within a few years, implying that the absence of wooden tools in the region does not necessarily mean they were never there.

Now, let's evaluate each option:

A. **It is pointed out that Professor Zhang's argument is against the facts.**

- This option suggests that Researcher Li is claiming Professor Zhang's argument contradicts known facts. However, Researcher Li is not directly stating that Zhang's argument is against the facts; he is providing an explanation for why the tools might not have been found.

B. **Quote authoritative research results that are at odds with Professor Zhang conclusion.**

- There is no mention of quoting any authoritative research results in Researcher Li's response. He is simply providing an explanation based on the properties of peat bogs and soil.

C. **It is pointed out that Professor Zhang misinterpreted the views of archaeologists.**

- Researcher Li does not address whether Professor Zhang misinterpreted the views of archaeologists. His focus is on explaining why the tools might not have been found in the region.

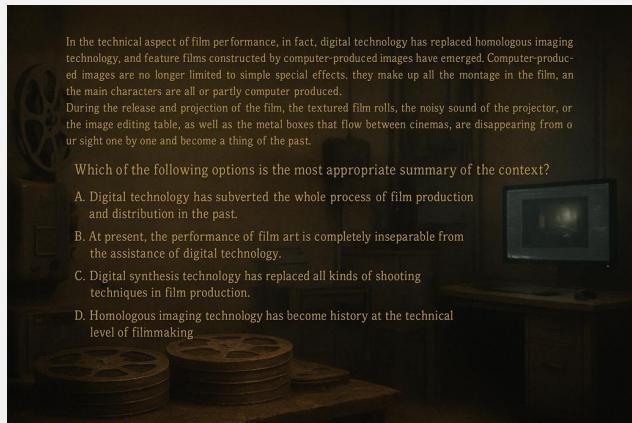
D. **Question Professor Zhang's implicit hypothesis.**

- Researcher Li questions the implicit assumption made by Professor Zhang that the absence of wooden tools in the region means they were never there. He provides an alternative explanation that the tools could have rotted away due to the type of soil.

Given this analysis, the most accurate summary of Researcher Li's coping method is that he questions Professor Zhang's implicit hypothesis.

Answer: D

Figure 15: A correct case of Qwen2.5-VL-72B on the question involving four reasoning types.



Layout: background style

Font style: standard

Reasoning types:

- categorical
- sufficient conditional
- necessary conditional
- conjunctive
- disjunctive

Solution: A



Qwen2.5-VL-72B:

To determine the most appropriate summary of the context, let's analyze each option in relation to the provided text:

A. "Digital technology has subverted the whole process of film production and distribution in the past."

- This option suggests that digital technology has completely overturned the traditional methods of film production and distribution. The text does mention the replacement of homologous imaging technology and the emergence of computer-produced images, but it doesn't explicitly state that the entire process has been subverted.

B. "At present, the performance of film art is completely inseparable from the assistance of digital technology."

- This option implies that digital technology is now an integral part of film art. The text supports this by mentioning that computer-produced images make up all the montage in films and that main characters are partly computer-produced. It also notes the disappearance of traditional film equipment, suggesting a strong reliance on digital technology.

C. "Digital synthesis technology has replaced all kinds of shooting techniques in film production."

- This option is too absolute. While the text mentions the use of computer-produced images, it doesn't claim that all shooting techniques have been replaced by digital synthesis technology.

D. "Homologous imaging technology has become history at the technical level of filmmaking."

- This option is partially correct as the text does mention that digital technology has replaced homologous imaging technology. However, it doesn't fully capture the broader impact described in the text, such as the role of digital technology in the entire filmmaking process.

Considering the analysis, option B best summarizes the context as it captures the current reliance on digital technology in film art.

Answer: B



Figure 16: A failure case of Qwen2.5-VL-72B on the question involving five reasoning types.

Ms. Zhang especially loves beauty. For many years, she likes to wear skirts in winter to show her graceful figure. Since last winter, she has felt knee pain in cold and overcast days. Later, the doctor diagnosed that she had arthritis. Therefore, Ms. Zhang believes that wearing less in cold and overcast days is the cause of arthritis.

Which of the following, if true, questions Ms. Zhang's point of view most?

- A. Some young women in Japan like to wear skirts in winter, but they don't suffer from arthritis because they wear less in cloudy and cold days
- B. Modern medical research shows that the root cause of arthritis is strain, infection or creation. Wearing less in cloudy and cold days is easy to attack
- C. Ms. Zhang's sister lives in a city with her. For many years, she also likes to wear skirts in winter, but she doesn't have arthritis
- D. Many people who wear more clothes in cloudy and cold days also get arthritis, and most of them are middle-aged and elderly people



Layout: background style

Font style: standard

Reasoning types:

1. categorical
2. sufficient conditional
3. necessary conditional
4. conjunctive
5. conjunctive

Solution: B



Qwen2.5-VL-72B:

To determine which option most questions Ms. Zhang's point of view, we need to analyze each choice and see how it challenges her belief that wearing less in cold and overcast days causes arthritis.

- A. This option suggests that some young women in Japan wear skirts in winter without suffering from arthritis because they wear less in cloudy and cold days. However, this does not directly challenge Ms. Zhang's belief; it merely provides an example of a different behavior pattern.
- B. This option states that modern medical research shows the root cause of arthritis is strain, infection, or creation, and that wearing less in cloudy and cold days makes one more susceptible. This directly contradicts Ms. Zhang's belief by providing a scientific explanation for arthritis that is unrelated to clothing choices.
- C. This option mentions that Ms. Zhang's sister also wears skirts in winter but doesn't have arthritis. While this provides a counterexample, it does not provide a broader explanation or challenge the underlying cause as effectively as option B.
- D. This option notes that many people who wear more clothes in cloudy and cold days still get arthritis, especially middle-aged and elderly people. This challenges the idea that clothing alone is the cause but does not provide a clear alternative explanation like option B.

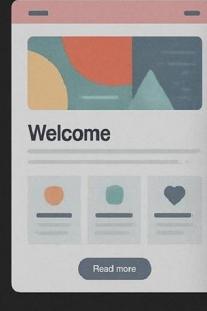
After analyzing all the options, the one that most effectively questions Ms. Zhang's point of view by providing a scientific explanation is:

Answer: B



Figure 17: A correct case of Qwen2.5-VL-72B on the question involving five reasoning types.

Man is a kind of cultural animal. Human behavior is determined not only by innate biological instincts, but also by acquired cultural and social factors. Everyone has a love of beauty. However, behind anything wonderful, there is an unbeautiful nature of origin, whether you realize it or not. Beauty is often a lie, but the truth is often ugly. Browse the web and you may be attracted by the gorgeous pages. If you look at the source code from the browser menu, you will find a mess that doesn't make any sense to you.



Which of the following is the main idea of this passage?

- Human behavior is affected by many factors such as acquired culture and society
- Human beings are a cultural animal
- Behind anything beautiful, there is an unbeautiful essence or origin
- Beauty is often a lie, and the truth is often hard to hear

Solution: C

o4-mini (792 completion tokens):

Let's break it down:

- The opening ("Man is a kind of cultural animal...") sets up how people are shaped by culture and instincts, but that is only an introduction.
- The bulk of the passage focuses on beauty's outward appeal versus its hidden, unattractive makeup ("behind anything wonderful, there is an unbeautiful nature of origin... if you look at the source code... you will find a mess").
- This leads to the key insight that outward beauty can mask an ugly reality.

Option C ("Behind anything beautiful, there is an unbeautiful essence or origin") captures that central contrast most precisely.

Answer: C



QvQ-72B-Preview (986 completion tokens):

Alright, I've got this passage to read and then answer a multiple-choice question about its main idea. Let's take it step by step.

First, the passage starts by saying that man is a kind of cultural animal. Human behavior isn't just determined by biological instincts but also by acquired cultural and social factors. So right off the bat, it's emphasizing that human behavior is influenced by a mix of things—biology and culture.

Next, it says everyone has a love of beauty. However, behind anything wonderful, there is an unbeautiful nature of origin, whether you realize it or not. Beauty is often a lie, but the truth is often ugly. This seems a bit philosophical. It's suggesting that surface beauty might not reflect the underlying reality, which can be less attractive or even ugly.

Then, it mentions browsing the web and being attracted to gorgeous pages. But if you look at the source code from the browser menu, you'll find a mess that doesn't make any sense to you. This seems like an example to illustrate the previous point. The beautiful webpage is like the beautiful surface, but the source code is the messy reality behind it.

So, overall, the passage seems to be about how human behavior is influenced by cultural and social factors, and how beauty or attractiveness can be deceptive, hiding an uglier truth.

Now, looking at the multiple-choice options:

A. Human behavior is affected by many factors such as acquired culture and society.

B. Human beings are a cultural animal.

C. Behind anything beautiful, there is an unbeautiful essence or origin.

D. Beauty is often a lie, and the truth is often hard to hear.

I need to choose which one captures the main idea of the passage.

Option A seems straightforward. It directly reflects what the passage says about human behavior being affected by cultural and social factors. That's definitely a key point in the passage.

Option B, "Human beings are a cultural animal," is also mentioned in the passage, but it's more of a starting point or a way to introduce the idea that human behavior is influenced by culture. It doesn't capture the full scope of the passage's main idea.

Option C, "Behind anything beautiful, there is an unbeautiful essence or origin," is related to the part of the passage that talks about beauty being a lie and truth being ugly. However, this seems more like a secondary point or an illustration of the main idea rather than the main idea itself.

Option D, "Beauty is often a lie, and the truth is often hard to hear," is similar to option C. It's talking on the idea that appearances can be deceptive, but it's phrased in a way that's more about beauty and truth rather than the broader concept of human behavior being influenced by cultural and social factors.

When I think about the entire passage, it seems that the main idea is indeed about how human behavior is shaped by various factors, particularly culture and society. The parts about beauty and truth seem to be more of an analogy or an example to illustrate this point rather than the central thesis.

So, between options A and B, A seems more comprehensive. It not only acknowledges that human behavior is affected by cultural and social factors but also emphasizes that there are many factors at play, including acquired ones.

Option B, "Human beings are a cultural animal," is more of a statement about humans in general, without specifying how culture affects behavior. It's a bit too broad. I can't capture the main idea of the passage, which talks about the influence of both biological instincts and cultural factors on behavior.

Therefore, I think option A is the best answer. It directly reflects the passage's main point about the multifaceted influences on human behavior, highlighting the role of acquired cultural and social factors.

Just to make sure, let me reread the passage quickly.

"Man is a kind of cultural animal. Human behavior is determined not only by innate biological instincts, but also by acquired cultural and social factors. Everyone has a love of beauty. However, behind anything wonderful, there is an unbeautiful nature of origin, whether you realize it or not. Beauty is often a lie, but the truth is often ugly. Browse the web and you may be attracted by the gorgeous pages. If you look at the source code from the browser menu, you will find a mess that doesn't make any sense to you."

Yes, the passage starts by establishing that humans are cultural animals and that their behavior is shaped by both biological instincts and cultural factors. Then it moves on to discuss beauty and truth, using the webpage and source code analogy to illustrate its point.

So, again, option A seems to capture the essence of the main idea: human behavior is affected by many factors, including acquired culture and society.

Therefore, my answer is A.

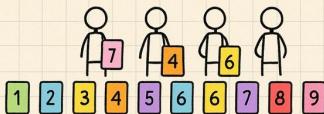
Final Answer

\boxed{A}



Figure 18: The response comparison of o4-mini and QvQ-72B-Preview (the thinking process of o4-mini cannot be achieved by user). In this case, o4-mini got correct answer with fewer output length.

There are nine cards, one to nine. A, B, C, D four people take cards, each take 2. It is known that the sum of the two cards taken by A is 10; the difference between the two cards taken by B is 1; the product of the two cards taken by C is 24; and the quotient of the two cards taken by D is 3.



Which of the following is the left card?

- A. Six.
- B. Three.
- C. Seven.
- D. Four.

Solution: D



(1958 completion tokens):

Let's denote the cards by 1–9. Four people (A, B, C, D) each pick 2 cards and one card is left over. We're given:
 A's cards sum to 10.
 B's cards differ by 1.
 C's cards multiply to 24.
 D's cards have a quotient of 3.

Step 1: List possible pairs.

A (sum 10): (1,9), (2,8), (3,7), (4,6)
 B (diff 1): (1,2), (2,3), (3,4), (4,5),
 (5,6), (6,7), (7,8), (8,9)
 C (product 24): (3,8), (4,6)
 D (quotient 3): (1,3), (2,6), (3,9)

Step 2: Try each C-pair.

Case C = (3,8).

Then D cannot use any pair containing 3 or 8 except (2,6).
 So D = (2,6). Used so far: {2,3,6,8}.
 Remaining cards: {1,4,5,7,9}.
 A must sum to 10 \Rightarrow only (1,9) is still available.
 So A = (1,9). Now used: {1,2,3,6,8,9}.
 Left for B and the spare: {4,5,7}.
 B must differ by 1 \Rightarrow B = (4,5).
 That leaves 7 as the single leftover.

Case C = (4,6).

One checks both possible D-pairs, but neither leads to a valid B from the remainders. So this case fails.

Thus the unique solution leaves card 7 unused.

Answer: C ✓



QvQ-72B-Preview (3341 completion tokens):

.....

Wait, in this scenario, C took (4,6), so 3 is not taken by C.

So, D can take (1,3) or (3,9).

If D takes (1,3), then 3 is taken.

If D takes (3,9), then 3 and 9 are taken.

Now, assign A: sum to 10.

Possible pairs:

- (1,9): 9 is taken by D.
- (2,8): both available.
- (3,7): 3 is taken by D.
- (4,6): 4 and 6 are taken by C.
- So, A can only take (2,8).

Now, assign B: difference of 1.

Possible pairs:

- (1,2): 2 is taken by A.
- (2,3): 2 is taken by A, 3 is taken by D.
- (3,4): 3 is taken by D, 4 is taken by C.
- (4,5): 4 is taken by C.
- (5,6): 6 is taken by C.
- (6,7): 6 is taken by C.
- (7,8): both available.
- (8,9): 9 is taken by D.

So, B can take (7,8).

Now, cards taken: 4,6,1,3,2,8,7,8.

Wait, 8 is taken twice, which isn't allowed.

So, this combination doesn't work.

Therefore, the only valid combination seems to be C taking (3,8), D taking (2,6), A taking (1,9), and B taking (4,5), leaving 7 as the unused card.

Therefore, the answer should be C. Seven.

Final Answer

\boxed{C} ✓

Figure 19: The response comparison of o4-mini and QvQ-72B-Preview. In this case, compared to QvQ-72B-Preview, o4-mini only uses 58.6% completion tokens to obtain the correct choice.