Big Escape Benchmark: Evaluating Human-Like Reasoning in Language Models via Real-World Escape Room Challenges

Zinan Tang[†] **Qiyao Sun**[†]
Beijing University of Post and Telecommunication
Beijing, China
tangzinan@bupt.edu.cn, 2022213402@bupt.cn

Abstract

Large Language Models (LLMs) have recently demonstrated remarkable reasoning capabilities across a wide range of tasks. While many benchmarks have been developed on specific academic subjects, coding, or constrained visual tasks, they often fail to fully capture the breadth, diversity, and dynamic nature of real-world human reasoning. Further, the creation of high-quality, complex multimodal reasoning benchmarks typically requires significant manual effort and expert annotation, which is costly and time-consuming. To address these limitations, we introduce Big Escape Bench, a novel multimodal reasoning benchmark derived from popular reality shows and television programs. Big Escape Bench leverages unique characteristics of TV content, providing a rich source of challenging and realistic multimodal reasoning problems. Key advantages include: questions guaranteed to be human-solvable and of moderate difficulty; problems reflecting diverse, real-world scenarios and knowledge domains; high inherent quality due to content generated by professional program teams. Notably, we develop an automated pipeline to construct the data from these programs into a standardized benchmark format, significantly reducing the manual effort compared to traditional dataset construction. We have conducted extensive experiments to evaluate state-of-the-art (SOTA) LLMs and Multimodal Large Language Models (MLLMs) on Big Escape Bench. Our results reveal a surprising performance gap: while the questions are easily solved by human viewers (about 60% in accuracy), the performance of even the most advanced models (best 40.50% in accuracy) is significantly lower than human-level accuracy. Big Escape Bench serves as a valuable tool for identifying current limitations of MLLMs and fostering future research towards more human-like multimodal reasoning.

1 Introduction

Recent years have witnessed unprecedented progress in the reasoning capabilities of LLMs (Guo et al., 2025; Jaech et al., 2024) and MLLMs (Team, 2024; Anthropic, 2025; Huang et al., 2025; Xu et al., 2024), with state-of-the-art (SOTA) systems achieving human-competitive performance on specialized tasks such as mathematical problem solving (Cobbe et al., 2021; Hendrycks et al., 2021; Liu et al., 2024b; Gao et al., 2025; Lin et al., 2025; Pei et al., 2025), code generation (Austin et al., 2021; Chen et al., 2021; Jain et al., 2025; Zhuo et al., 2025), and constrained visual question answering (Yue et al., 2024; He et al., 2024; Chen et al., 2025b). However, these successes often rely on benchmarks that prioritize narrow, domain-specific expertise (e.g., MATH (Liu et al., 2024b) for math, HumanEval (Chen et al., 2021) for coding) or static, artificially constructed multimodal tasks (e.g., image captioning or VQA datasets). However, such benchmarks are not sufficient to capture the breadth, diversity, and dynamic nature of real-world reasoning, where humans seamlessly integrate multimodal information, adapt to novel contexts, and apply commonsense knowledge to solve open-ended problems.

A critical gap persists in evaluating models on reasoning tasks that mirror the complexity of human challenges. Existing benchmarks face several key limitations: (a) The scope of many existing benchmarks is limited, disproportionately emphasizing performance in specific technical domains, such as math and code, while overlooking the assessment of more general, contextually embedded reasoning abilities critical for real-world understanding. (b) Benchmarks constructed

[†] Equal Contribution.

Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM² 2025), pages 488–503

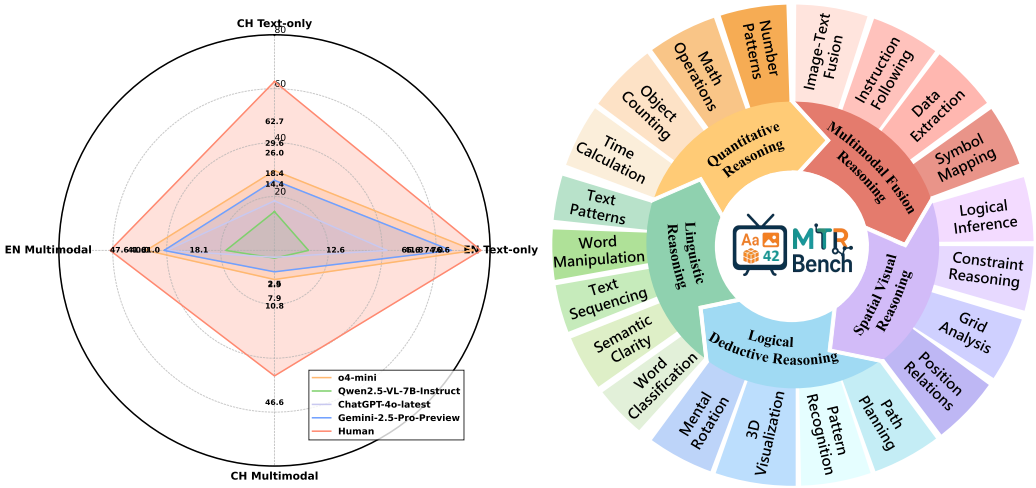

Figure 1: ***Big Escape Benchmark* comprises 252 reasoning tasks that assess 5 reasoning categories across 21 problem types.** It provides bilingual (Chinese / English) evaluation of both textual and visual reasoning categories.

through static, manual processes often result in homogeneous question sets, thereby failing to capture the innovation and rich variability inherent in dynamic, real-world scenarios. (c) The development of complex and high-fidelity multimodal reasoning datasets typically incurs substantial human costs, stemming from the requirement for labor-intensive annotation and expert validation processes. For instance, benchmarks like MMMU (Yue et al., 2024) or GPQA (Rein et al., 2024), while comprehensive, focus on academic subjects and rely on curated, domain-specific content. This leaves open the question of whether current models can generalize to more diverse, complex, and real-world reasoning demands.

To address these challenges, we introduce *Big Escape Benchmark*, a novel multimodal reasoning benchmark derived from popular reality shows and television programs (e.g., *The Great Escape* and *The 1% Club*). TV content has unique characteristics that offer untapped resources for benchmarking: questions are designed by professional production teams to challenge human contestants, ensuring they are inherently solvable, contextually grounded, and dynamically varied. By leveraging these resources, *Big Escape Benchmark* offers significant benefits, including (1) **Human-aligned difficulty**: All problems are vetted for solvability by human participants, ensuring a balanced evaluation of model capabilities without artificial extremes (e.g., trivial or impossibly niche questions); (2) **Diverse and real-world knowledge**: Questions span broad domains (e.g., logic, commonsense, cultural references) and tasks, reflecting the integrative demands of real-life reasoning; (3) **Sustainable innovation**: Since the TV shows update continuously through live broadcasts, the benchmark resists data contamination and encourages models to handle novel and unseen challenges.

Beyond the conceptual strengths of *Big Escape Benchmark*, the benchmark collection pipeline also introduces methodological innovation. Specifically, we develop an automated pipeline to extract, preprocess, and standardize TV content into a scalable benchmark, minimizing manual annotation while preserving the richness of the original material. We leverage an automated pipeline that begins with accurate transcript generation using tools like Videolingo, followed by GPT-4o-mini (Hurst et al., 2024) for refinement. Subsequently, a sophisticated LLM, Claude-3.7-sonnet (Anthropic, 2025), is employed to analyze dialogue and extract problem instances along with relevant clues from the video content. Importantly, this approach not only reduces costs but also enables future expansion to new programs or regions.

We have conducted extensive experiments evaluating multiple advanced LLMs (e.g., DeepSeek V3 (DeepSeek-AI et al., 2025), Grok 3 beta (X.ai, 2025)) and MLLMs (e.g., Qwen2.5-VL-Instruct (Bai et al., 2025), GPT-4o-latest (Hurst et al., 2024), Gemini-2.5 (Google, 2025), o4-mini (OpenAI, 2025)) on our *Big Escape Benchmark*. While human viewers can easily solve these problems with high accuracy (about 60%), the performance of even the most advanced models (e.g., leading proprietary models like Claude-3.7-Sonnet (Anthropic, 2025) and Gemini-2.5-Pro (Google, 2025)) test falls considerably short, trailing human performance by over

30%. Our analysis reveals a significant performance gap between open-source and proprietary models. We also find that while model scaling and the integration of sophisticated reasoning mechanisms can yield high performance, these approaches often encounter diminishing returns or introduce efficiency trade-offs. Furthermore, we observe that wrong reasoning ideas, rather than incorrect information extraction, are a primary driver of model failures; indeed, models with strong reasoning capabilities can exhibit a tendency to overthink textual information. This stark contrast underscores that despite rapid advancements, LLMs and MLLMs still face substantial challenges in robustly performing the diverse, dynamic, and context-dependent reasoning at which humans excel.

2 Related works

LLM reasoning. Enhancing reasoning capabilities is one of the core objectives for LLMs (Qu et al., 2025; Ke et al., 2025). Early approaches introduced explicit prompting techniques like Chain-of-Thought (CoT) (Wei et al., 2022). Subsequently, large reasoning models (LRMs) such as o1 (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025) leveraged reinforcement learning (RL) algorithms (Schulman et al., 2017; Rafailov et al., 2023; Shao et al., 2024) and test-time scaling to significantly improve model reasoning performance (Team et al., 2025; Huang and Chang, 2023; Snell et al., 2025; Zeng et al., 2025; Team, 2025). These models primarily focus on tasks with high reasoning requirements in domains such as mathematics and code. Recently, the deep thinking paradigm has been extended to the domain of multimodal model reasoning (Team, 2024; Anthropic, 2025; Huang et al., 2025; Xu et al., 2024), thereby promoting advancements in multimodal reasoning capabilities.

Reasoning benchmarks. Evaluating the reasoning capabilities of LLMs has spurred the development of a diverse array of benchmarks. These initially covered established domains such as mathematical reasoning (Cobbe et al., 2021; Hendrycks et al., 2021; Liu et al., 2024b; Gao et al., 2025; Pan et al., 2025), coding (Austin et al., 2021; Chen et al., 2021; Jain et al., 2025; Zhuo et al., 2025), and other disciplines (Clark et al., 2018; Rein et al., 2024). To probe broader and more general cognitive abilities, many benchmarks now focus on puzzles collated from various online websites and other repositories (Wang et al., 2025; Toh et al., 2025; Estermann et al., 2024; Gui et al., 2024; Chia et al., 2024). Notable examples include comprehensive puzzle collections like Big-bench (Srivastava et al., 2022), BBH (Suzgun et al., 2022), and BBEH (Kazemi et al., 2025). Other benchmarks concentrate on specific puzzle formats, such as FINEREASON (Chen et al., 2025a) with tasks like Sudoku, Graph Coloring, and the Game of 24, and CrossWordBench (Leng et al., 2025) which employs crossword puzzles. The scope of reasoning evaluation has also expanded to incorporate visual information, leading to multimodal benchmarks (Yue et al., 2024; He et al., 2024; Chen et al., 2025b). An emerging trend in this landscape is the diversification of problem sources: beyond traditional website collection, recent efforts utilize logical reasoning puzzles from real-world examinations (Song et al., 2025; Bi et al., 2025; Cai et al., 2025) and even based on physical objects like LEGO bricks (Tang et al., 2025).

3 *Big Escape Benchmark*

3.1 Data source

To overcome existing benchmarks' limitations in capturing the complexity of real-world human reasoning, *Big Escape Benchmark* utilizes data sourced from popular television programs. This approach generates problems distinct from those found in narrowly-focused or synthetic datasets, fostering a more authentic and comprehensive evaluation. For its initial construction, *Big Escape Benchmark* curates content from internationally recognized shows such as China's *The Great Escape*, America's *Escape! with Janet Varney*, and Britain's *The 1% Club*. These programs, rich in puzzles, escape room scenarios, and intricate questions, serve as a valuable resource for assessing nuanced reasoning abilities. The international diversity of these sources also infuses varied cultural and contextual elements, thereby expanding the benchmark's coverage and challenging models towards more effective generalization.

3.2 Data collection pipeline

We developed a multi-stage data collection and curation pipeline to convert rich television content into standardized, high-quality problems for *Big Escape Benchmark*, and to address the inef-

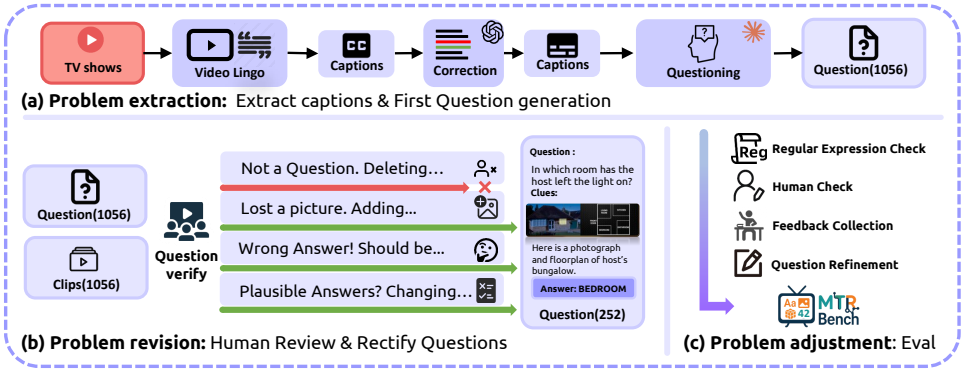

Figure 2: **Pipeline of *Big Escape Benchmark*.** **(a)** We illustrate that by utilizing the VideoLingo framework and LLMs, we can extract and enhance puzzle data from video transcripts. **(b)** This process extends the meticulous validation performed by human reviewers for the extracted puzzles, ensuring logical coherence and filtering for solvability. **(c)** We confirm the effectiveness of our method after the benchmark undergoes iterative refinement through automated validation and feedback from culturally knowledgeable respondents, optimizing both clarity and difficulty.

ficiencies and scalability limitations of traditional manual dataset creation. This pipeline, comprising problem extraction, revision, and adjustment stages, ensures the reliability and rigor of the resulting problems.

Problem extraction. The initial phase of our data pipeline focuses on accurately extracting problem instances from video content. This process commences with the generation of high-fidelity textual transcripts. For this, we employ VideoLingo, an advanced framework for robust subtitle extraction and correction. VideoLingo transcribes timestamped dialogue from raw video footage and performs real-time correction of speech recognition errors. These initial transcripts are then meticulously refined using the GPT-4o-mini model (Hurst et al., 2024) to yield corrected and accurately timestamped textual data.

With these high-quality subtitles established, the subsequent crucial step is the automated extraction of problem-specific information. This involves analyzing participant dialogue to pinpoint a puzzle's introduction and resolution, and to extract pertinent clues embedded within the conversational context. Critically, this stage requires the model to logically infer and differentiate between various solution attempts and the definitive answer, thereby ensuring the accurate isolation of key information for each puzzle. Given these demanding requirements for accuracy and nuanced understanding, we evaluated several leading language models, including Gemini, DeepSeek, ChatGPT, and Claude. Claude-3.7-sonnet-thinking (Anthropic, 2025) demonstrated superior performance in fulfilling these requirements and was thus selected to implement this automated extraction. Specific prompt engineering strategies and comprehensive templates are detailed in Appendix C.3.

Problem revision.

This protocol comprises two key stages: (1) **Screening**: This phase validates each problem's inherent solvability (i.e., it was demonstrably solved in the source program) and its alignment with *Big Escape Benchmark*'s core principles. Problems are excluded if unsuitable for a Q&A format (e.g., those requiring physical interaction by the solver) or if they lack a clear solution derivable from the available clues, thereby maintaining task integrity and ecological validity. (2) **Refinement**: This phase optimizes selected problems. Reviewers craft clear Q&A phrasing and supplement critical missing information, especially visual clues, to preserve the original puzzle's multimodal nature. To establish a single, verifiable correct answer for each Q&A problem grounded in the source material, reviewers add disambiguating context or constraining elements if the initial phrasing could permit unintended plausible solutions. This process ensures a unique logical reasoning path to the intended answer, even if other interpretations were considered and ruled out during the review.

The outcome is a curated set of problems, each featuring an unambiguous question, a verified solution, and all necessary textual and visual clues, thereby upholding *Big Escape Benchmark*'s high standards for accuracy, logical coherence, and appropriate difficulty.

https://github.com/Huanshere/VideoLingo

Problem adjustment. Following the revision stage, problems undergo a final adjustment phase designed to maximize dataset integrity and human alignment. This phase begins with an internal answer verification step, where regular expression tools, guided by predefined criteria, standardize annotated answers. This process ensures consistent formatting (e.g., case, spacing), resulting in unambiguous, programmatically evaluable solutions.

Subsequently, an external human evaluation is conducted using participants entirely naive to both the problem development process and the original program content. Crucially, these evaluators are distinct from any expert human group whose performance might be reported as a human baseline for *Big Escape Benchmark* (see Section 4.1). Participants are selected for relevant cultural knowledge, allowing them to attempt solutions under objective conditions, mimicking real-world problem-solving. Their responses, success rates, and common answer patterns provide crucial empirical data for assessing problem difficulty, identifying potential ambiguities, and guiding final adjustments to problem wording or structure. This iterative feedback loop enhances overall problem coherence and fairness.

The overarching goal of this adjustment stage is to ensure that *Big Escape Benchmark* not only effectively challenges multimodal language models but also remains well-calibrated against general human reasoning capabilities.

3.3 Dataset statistics and splits

Table 1: **The Statistics of *Big Escape Benchmark*.** *Big Escape Benchmark* encompasses a comprehensive, equilibrated corpus of interrogatives in both Chinese and English languages, incorporating both textual and multimodal question formats.

Category	Statistics
Total Questions	252
Chinese	113
- CH Textonly	50
- CH Multimodal	63
English	139
- EN Textonly	57
- EN Multimodal	82
Chinese / English	46.4% / 53.6%
Text-only / Multimodal	42.4% / 57.6%

The comprehensive data collection pipeline described previously yields *Big Escape Benchmark*, a dataset comprising 252 carefully curated multimodal reasoning questions. Sourced from diverse television programs, these questions are presented in their original languages, encompassing both Chinese and English content, and require either text-based reasoning or the interpretation of visual clues. Accordingly, *Big Escape Benchmark* is organized into four distinct subsets based on language (Chinese or English) and clue modality. Comprehensive dataset statistics are provided in Table 1 and Table 4.

To facilitate a more nuanced analysis of the reasoning skills tested, problems within *Big Escape Benchmark* are further mapped to 21 fine-grained types and 5 overarching reasoning categories, as outlined in Figure 1. Detailed descriptions of this categorization process and its criteria can be found in Appendix B.1 and Appendix B.2.

Furthermore, as the source television programs are continually updated, *Big Escape Benchmark* will be regularly expanded in future releases. This will ensure its continued relevance and the introduction of novel reasoning challenges.

3.4 Comparison with other benchmarks

Current multimodal reasoning benchmarks often suffer from limited diversity, typically being confined to a narrow range of question types and similar prompts. Our novel benchmark for text-visual reasoning directly addresses this deficiency by leveraging rich content from real-world television programs. It introduces 21 distinct question types, each accompanied by unique prompts, a significant expansion compared to existing benchmarks, which usually feature fewer than ten. Critically, all tasks are presented in a question-answering (QA) format. This strategic choice minimizes the likelihood of correct answers obtained through guessing, a prevalent issue in multiple-choice settings, thereby emphasizing genuine inferential abilities. The data originates from human-intensive reasoning tasks within detective television series; each question is manually verified for authenticity and complexity, contrasting with datasets that are programmatically generated or directly adopt publicly available web data. This comprehensive methodology facilitates a more rigorous evaluation of a model's capacity for diverse reasoning and effective generalization.

Table 2: **Comparison of *Big Escape Benchmark* with existing benchmarks.** *Big Escape Benchmark* uniquely offers the most diverse reasoning types, exclusively Q&A format, and sources data from real-world TV shows rather than web content or code generation. MCQ means Muti-Choices Questions.

Benchmark	Question Types	Answer Type	Source	Content Type	Language
MC (Todd et al., 2024)	2	MCQ	Internet	Text	English
DOTP (Webb et al., 2020)	2	MCQ	Code Generation	Images	English
VAP (Hill et al., 2019)	3	MCQ	Human	Images	English
G-set (Mańdziuk and Żychowski, 2019)	4	MCQ	Code Generation	Images	English
ARC (Chollet, 2019)	4	MCQ	Code Generation	Images	English
RAVEN (Zhang et al., 2019)	5	MCQ	Code Generation	Images	English
VisualPuzzles (Song et al., 2025)	5	MCQ	Internet, Textbook	Images	English
MARVEL(Jiang et al., 2024)	5	MCQ	Internet	Images	English
KOR Bench (Ma et al., 2024)	5	Q&A	Internet	Text	English
VisuLogic (Xu et al., 2025)	6	MCQ	Internet	Images	English
MMIQ (Cai et al., 2025)	8	MCQ	Internet	Images	English
CipherBank (Li et al., 2025)	9	Q&A	Synthetic	Text	English
PuzzleVQA (Chia et al., 2024)	10	MCQ	Internet	Images	English
VERIFY (Bi et al., 2025)	10	MCQ	Internet	Images	English
LEGO-Puzzles (Tang et al., 2025)	11	MCQ	Internet	Images	English
Big Escape Benchmark	**21**	**Q&A**	**TV Shows**	**Text & Images**	**English & Chinese**

Table 3: **Full evaluation results of 32 models on *Big Escape Benchmark*.** Gray indicates the best performance for each task among all models and light gray indicates the best result among open-source models. Futhermore, reasoning models are highlighted by light yellow.

Models	CH Text-only		EN Text-only		CH Multimodal		EN Multimodal		Overall	
	pass@1	pass@5	pass@1	pass@5	pass@1	pass@5	pass@1	pass@5	pass@1	pass@5
Proprietary LLM										
Grok-3-Beta	20.80	38.00	54.04	59.65	-	-	-	-	37.42	48.83
Doubao-1.5-Pro-32k (250115)	24.80	34.00	21.05	36.84	-	-	-	-	22.93	35.42
Doubao-1.5-Thinking-Pro (250415)	34.80	44.00	55.79	61.40	-	-	-	-	45.30	52.70
Open-source LLM										
DeepSeek-V3-0324	26.40	40.00	48.77	64.91	-	-	-	-	37.59	52.46
DeepSeek-R1	28.80	44.00	54.39	71.93	-	-	-	-	41.60	57.97
Llama-3.3-70B-Instruct	6.80	12.00	8.42	24.56	-	-	-	-	7.61	18.28
Llama-4-Scout-17B-16E-Instruct	12.00	16.00	10.88	22.81	-	-	-	-	11.44	19.41
Llama-4-Maverick-17B-128E-Instruct	12.80	38.60	23.86	38.60	-	-	-	-	18.33	38.60
Qwen2.5-7B-Instruct	2.80	12.00	4.91	10.53	-	-	-	-	3.86	11.27
Qwen2.5-14B-Instruct	9.20	16.00	7.37	15.79	-	-	-	-	8.29	15.90
Qwen2.5-32B-Instruct	13.20	22.00	8.42	14.04	-	-	-	-	10.81	18.02
Qwen2.5-72B-Instruct	12.40	22.81	11.23	26.32	-	-	-	-	11.82	24.57
QwQ-32B	14.00	24.00	42.11	49.12	-	-	-	-	28.06	36.56
Proprietary MLLM										
Gemini-2.5-Flash-Preview (250417)	18.00	30.00	28.77	56.14	7.30	12.70	30.24	59.76	21.08	40.84
Gemini-2.5-Pro-Preview (250506)	26.00	36.00	65.61	84.21	7.94	17.46	40.98	62.2	35.13	49.97
ChatGPT-4o-latest (250326)	18.40	30.00	41.75	59.65	2.54	14.29	40.00	63.41	25.67	41.84
GPT-4.1 (250414)	22.00	32.00	40.00	68.42	8.57	14.29	35.12	54.88	26.42	42.40
GPT-4.1-mini (250414)	18.40	26.00	37.89	54.39	6.03	9.52	28.54	47.56	22.71	34.37
o4-mini (250416)	29.60	42.00	74.04	87.72	10.79	14.29	47.56	75.61	40.50	55.70
Claude-3.7-Sonnet (250219)	19.20	32.00	40.00	59.65	3.17	7.94	28.54	53.66	22.73	38.31
Claude-3.7-Sonnet (thinking-32k-250219)	26.80	42.00	68.07	82.46	7.94	17.46	36.10	58.54	34.73	49.32
Doubao-1.5-Vision-Pro (250328)	22.00	32.00	16.49	29.82	1.59	6.35	24.88	37.80	16.24	26.49
Doubao-1.5-Thinking-Pro-m (250415)	29.60	32.00	44.21	61.40	6.35	14.29	28.54	54.88	27.18	40.64
Open-source MLLM										
Qwen2.5-VL-7B-Instruct	2.40	6.00	3.86	8.77	1.59	3.17	6.34	28.05	3.55	11.50
Qwen2.5-VL-32B-Instruct	13.20	18.00	9.82	22.81	2.54	6.35	16.59	41.46	10.54	22.16
Qwen2.5-VL-72B-Instruct	14.40	24.00	12.63	21.05	2.86	7.94	18.05	47.56	11.99	25.14
Llama-3.2-11B-Vision-Instruct	2.00	4.00	3.86	10.53	1.59	4.76	9.76	24.39	4.30	10.92
Llama-3.2-90B-Vision-Instruct	10.00	16.00	8.42	24.56	3.17	7.94	10.24	30.49	7.96	19.78
InternVL3-8B-Instruct	4.00	8.00	5.26	5.26	0.00	0.00	15.61	34.15	6.22	11.85
InternVL3-14B-Instruct	4.00	8.00	7.02	10.53	1.59	1.59	23.17	32.93	8.95	13.26
InternVL3-38B-Instruct	8.00	12.00	10.53	10.53	1.59	0.00	21.95	34.15	10.52	14.17
InternVL3-78B-Instruct	6.00	10.00	8.77	10.53	0.00	1.59	17.07	36.59	7.96	14.68
Human										
Human Expert Avg.	62.67	71.67	76.61	88.89	46.56	65.61	60.98	78.86	61.70	76.26

4 Experiments

4.1 Experiment Setup

To comprehensively evaluate model capabilities, our experimental setup encompasses a diverse range of models, standardized evaluation frameworks, and rigorous human performance baselines.

Evaluation models. Our evaluation includes a total of 32 models, comprising 13 LLMs and 19 MLLMs. The LLMs feature open-source models such as DeepSeek-V3-0324 (Liu et al., 2024a), DeepSeek-R1 (Guo et al., 2025), Llama-3.3-70B-Instruct (Grattafiori et al., 2024), QwQ-32B (Team, 2025), the Qwen2.5-Instruct series (7B, 32B, 72B) (Yang et al., 2024), and the Llama4 series (Scout-17B-16E-Instruct, Maverick-17B-128E-Instruct). Proprietary LLMs include Grok-3-Beta and Doubao-1.5-Pro (Thinking). For MLLMs, we assess open-source models including the Qwen2.5-VL-Instruct series (7B, 32B, 72B) (Yang et al., 2024), QVQ-72B-Preview, and the Llama-3.2-Vision-Instruct series (11B, 90B) (Grattafiori et al., 2024). Evaluated proprietary MLLMs include Gemini-2.5-Flash-Preview, Gemini-2.5-Pro-Preview, ChatGPT-4o-latest (Hurst et al., 2024), GPT-4.1 (mini), o4-mini, Claude-3.7-Sonnet (thinking), Doubao-1.5-Vision-Pro, and Doubao-1.5-Thinking-Pro-m.

Evaluation Protocol. We use OpenCompass (Contributors, 2023) for text-based tasks and VLMEvalKit (Duan et al., 2024) for multimodal benchmarks. Following common practice, we report both `Pass@1` and `Pass@5` (Li et al., 2024), which measure whether at least one correct answer appears among the top-1 or top-5 generated outputs, we define `Pass@N` as follows:

$$\texttt{Pass@N} = \mathop{\mathbb{E}}_{\text{Problems}}[\min(c, 1)]. \quad (1)$$

All models are prompted with chain-of-thought instructions by appending "*Let's think step by step*" to the inputs (detailed prompts are provided in Figure 4). For `Pass@1`, we use greedy decoding; for `Pass@5`, we apply sampling with temperature set to 0.6. The maximum output length is set to 4,096 tokens, extended to 32,768 for models with long-context capabilities. For API-based models, we average results over multiple runs to account for potential non-determinism.

Human evaluation. To establish a reference baseline, we recruit three science and engineering undergraduate students to solve the benchmark puzzles under consistent constraints: no external tools can be used and a 5-minute time limit per problem. Each participant provides one primary answer and, when applicable, up to four additional guesses. We compute `Pass@1` and `Pass@5` in the same way as for models.

4.2 Overall results

Human performance remains substantially higher than all models. As shown in Table 3, human experts outperform all models across every setting, achieving an overall `pass@1` of 61.70% and `pass@5` of 76.26%. In comparison, the best-performing model, o4-mini, reaches only 40.50% `pass@1` and 55.70% `pass@5`, indicating a gap of over 20 percentage points. Even with the relaxed `pass@5` setting, the gap persists, highlighting that current models—despite their progress—still fall significantly short in solving complex reasoning tasks with human-like consistency.

Proprietary models outperform open-source counterparts by a wide margin. We observe a consistent and substantial performance gap between proprietary and open-source models, particularly in the multimodal setting. For example, o4-mini achieves 10.79% and 47.56% on Chinese and English multimodal tasks (under `pass@1`), whereas the strongest open-source MLLM, Qwen2.5-VL-72B, reaches only 3.17% and 18.05%. In the text-only setting, the gap narrows: DeepSeek-R1 performs competitively with proprietary models, achieving 41.60% overall `pass@1`, surpassing Claude-3.7-Sonnet(22.73%) and approaching o4-mini (40.50%). This suggests that open-source LLMs are catching up in text-based reasoning, but still lag in multimodal understanding.

Reasoning-specialized models improve performance but incur higher cost. Several reasoning-enhanced models (e.g., DeepSeek-R1, Doubao-Thinking-Pro, Claude-3.7-Thinking) outperform their non-reasoning counterparts in `pass@1` accuracy, attributed to their ability to produce explicit chain-of-thought (CoT) rationales. For instance, Doubao-Thinking-Pro achieves 45.3% `pass@1`, compared to 22.93% for the non-reasoning variant. However, this performance gain comes at the cost of significantly longer

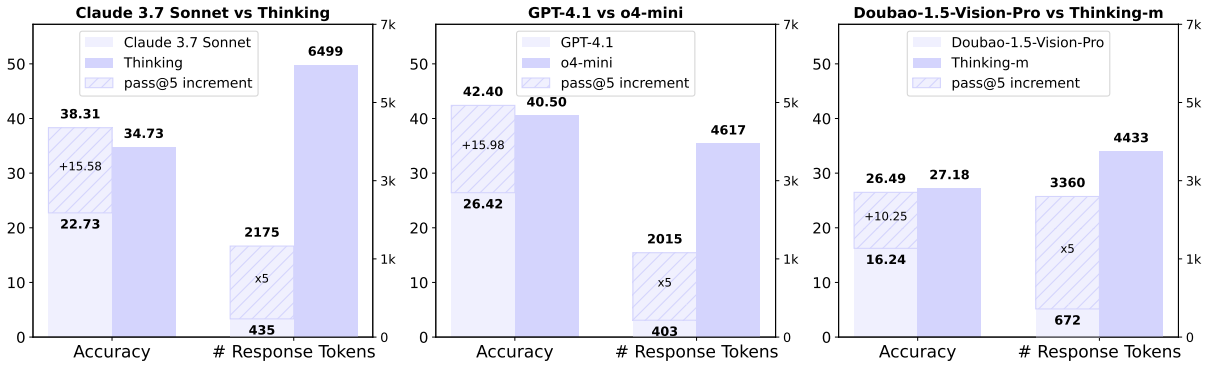

Figure 3: **Comparison of accuracy and average number of total completion tokens of reasoning models and their general counterparts.** It highlighting that calculating `Pass@N` using 5 samples from general models can achieve performance comparable or superior to reasoning models, with reduced token expenditure.

outputs and increased token usage. Moreover, baseline models using sampling (`pass@5`) often reach similar or better performance with far less decoding overhead. These results suggest that while reasoning traces help, they trade off efficiency and are not always necessary.

Scaling model size improves performance, but with diminishing returns. Larger models generally yield better results, yet the improvements taper off at higher scales. For example, in the Qwen2.5-VL-Instruct series, `pass@1` increases from 3.55% (7B) to 10.54% (32B), but only marginally further to 11.99% (72B). A similar pattern is observed in InternVL3 and LLaMA-Vision series. This diminishing return highlights that parameter count alone is not sufficient to overcome the reasoning difficulty posed by our benchmark, and future gains will likely depend on architectural advances or training strategies beyond simple scaling.

***Big Escape Benchmark* presents a challenging benchmark across both text and multimodal domains.** Across all tasks and model types, scores on *Big Escape Benchmark* remain low relative to standard benchmarks. Even the strongest models achieve only 40–45% `pass@1` on average, with particularly low scores in the Chinese multimodal setting (e.g., <11% `pass@1` for top models). The consistently large gap between model and human performance, the underperformance of large open-source MLLMs, and the limited benefits of scale all point to the intrinsic difficulty of the benchmark. This confirms *Big Escape Benchmark* as a reliable stress test for evaluating fine-grained reasoning in both unimodal and multimodal contexts.

5 Conclusion

We introduce *Big Escape Benchmark*, a novel multimodal reasoning benchmark derived from reality TV shows, addressing the diversity, dynamism, and creation-cost limitations of current benchmarks. *Big Escape Benchmark* features human-solvable, diverse, high-quality problems via an automated pipeline. Experiments revealed a significant performance gap: humans achieve approximately 60% accuracy, while top models reach only about 40.50%. This highlights that even advanced MLLMs struggle with human-like, context-dependent reasoning. Our analysis indicates that flawed reasoning approaches are the primary error source. *Big Escape Benchmark* offers a valuable tool to identify MLLM limitations and guide future research towards more robust multimodal reasoning.

References

Anthropic. 2025. Claude 3.7 sonnet and claude code.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Jing Bi, Junjia Guo, Susan Liang, Guangyu Sun, Luchuan Song, Yunlong Tang, Jinxi He, Jiarui Wu, Ali Vosoughi, Chen Chen, et al. 2025. Verify: A benchmark of visual explanation and reasoning for investigating multimodal reasoning fidelity. *arXiv preprint arXiv:2503.11557*.

Huanqia Cai, Yijun Yang, and Winston Hu. 2025. Mm-iq: Benchmarking human-like abstraction and reasoning in multimodal models. *arXiv preprint arXiv:2502.00698*.

Guizhen Chen, Weiwen Xu, Hao Zhang, Hou Pong Chan, Chaoqun Liu, Lidong Bing, Deli Zhao, Anh Tuan Luu, and Yu Rong. 2025a. Finereason: Evaluating and improving llms' deliberate reasoning through reflective puzzle solving. *arXiv preprint arXiv:2502.20238*.

Jiacheng Chen, Tianhao Liang, Sherman Siu, Zhengqing Wang, Kai Wang, Yubo Wang, Yuansheng Ni, Ziyan Jiang, Wang Zhu, Bohan Lyu, Dongfu Jiang, Xuan He, Yuan Liu, Hexiang Hu, Xiang Yue, and Wenhu Chen. 2025b. MEGA-bench: Scaling multimodal evaluation to over 500 real-world tasks. In *The Thirteenth International Conference on Learning Representations*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Yew Ken Chia, Vernon Toh Yan Han, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. 2024. Puzzlevqa: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. *arXiv preprint arXiv:2403.13315*.

François Chollet. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, et al. 2025. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201.

Benjamin Estermann, Luca Lanzendörfer, Yannick Niedermayr, and Roger Wattenhofer. 2024. Puzzles: A benchmark for neural algorithmic reasoning. *Advances in Neural Information Processing Systems*, 37:127059–127098.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, et al. 2025. Omni-MATH: A universal olympiad level mathematic benchmark for large language models. In *The Thirteenth International Conference on Learning Representations*.

Google. 2025. Gemini 2.5.

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

Jiayi Gui, Yiming Liu, Jiale Cheng, Xiaotao Gu, Xiao Liu, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2024. Logicgame: Benchmarking rule-based reasoning abilities of large language models. *arXiv preprint arXiv:2408.15778*.

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

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Felix Hill, Adam Santoro, David GT Barrett, Ari S Morcos, and Timothy Lillicrap. 2019. Learning to make analogies by contrasting abstract relational structure. *arXiv preprint arXiv:1902.00120*.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *Preprint*, arXiv:2503.06749.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2025. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*.

Yifan Jiang, Kexuan Sun, Zhivar Sourati, Kian Ahrabian, Kaixin Ma, Filip Ilievski, Jay Pujara, et al. 2024. Marvel: Multidimensional abstraction and reasoning through visual evaluation and learning. *Advances in Neural Information Processing Systems*, 37:46567–46592.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K Jain, Virginia Aglietti, Disha Jindal, Peter Chen, et al. 2025. Big-bench extra hard. *arXiv preprint arXiv:2502.19187*.

Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, Peifeng Wang, Silvio Savarese, et al. 2025. A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems. *arXiv preprint arXiv:2504.09037*.

Jixuan Leng, Chengsong Huang, Langlin Huang, Bill Yuchen Lin, William W Cohen, Haohan Wang, and Jiaxin Huang. 2025. Crosswordbench: Evaluating the reasoning capabilities of llms and lvlms with controllable puzzle generation. *arXiv preprint arXiv:2504.00043*.

Chen Li, Weiqi Wang, Jingcheng Hu, Yixuan Wei, Nanning Zheng, Han Hu, Zheng Zhang, and Houwen Peng. 2024. Common 7b language models already possess strong math capabilities. *Preprint*, arXiv:2403.04706.

Yu Li, Qizhi Pei, Mengyuan Sun, Honglin Lin, Chenlin Ming, Xin Gao, Jiang Wu, Conghui He, and Lijun Wu. 2025. Cipherbank: Exploring the boundary of llm reasoning capabilities through cryptography challenges. *Preprint*, arXiv:2504.19093.

Honglin Lin, Zhuoshi Pan, Yu Li, Qizhi Pei, Xin Gao, Mengzhang Cai, Conghui He, and Lijun Wu. 2025. Metaladder: Ascending mathematical solution quality via analogical-problem reasoning transfer. *Preprint*, arXiv:2503.14891.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Hongwei Liu, Zilong Zheng, Yuxuan Qiao, Haodong Duan, Zhiwei Fei, Fengzhe Zhou, Wenwei Zhang, Songyang Zhang, Dahua Lin, and Kai Chen. 2024b. Mathbench: Evaluating the theory and application proficiency of llms with a hierarchical mathematics benchmark. *arXiv preprint arXiv:2405.12209*.

Kaijing Ma, Xinrun Du, Yunran Wang, Haoran Zhang, Zhoufutu Wen, Xingwei Qu, Jian Yang, Jiaheng Liu, Minghao Liu, Xiang Yue, et al. 2024. Kor-bench: Benchmarking language models on knowledge-orthogonal reasoning tasks. *arXiv preprint arXiv:2410.06526*.

Jacek Mańdziuk and Adam Żychowski. 2019. Deepiq: A human-inspired ai system for solving iq test problems. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

OpenAI. 2025. Openai o4-mini.

Zhuoshi Pan, Yu Li, Honglin Lin, Qizhi Pei, Zinan Tang, Wei Wu, Chenlin Ming, H. Vicky Zhao, Conghui He, and Lijun Wu. 2025. Lemma: Learning from errors for mathematical advancement in llms. *Preprint*, arXiv:2503.17439.

Qizhi Pei, Lijun Wu, Zhuoshi Pan, Yu Li, Honglin Lin, Chenlin Ming, Xin Gao, Conghui He, and Rui Yan. 2025. Mathfusion: Enhancing mathematic problem-solving of llm through instruction fusion. *Preprint*, arXiv:2503.16212.

Xiaoye Qu, Yafu Li, Zhaochen Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong Liu, Shuxian Liang, Junxian He, Peng Li, Wei Wei, Jing Shao, Chaochao Lu, Yue Zhang, Xian-Sheng Hua, Bowen Zhou, and Yu Cheng. 2025. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. *Preprint*, arXiv:2503.21614.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*.

Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. 2025. Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge. *arXiv preprint arXiv:2504.10342*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Kexian Tang, Junyao Gao, Yanhong Zeng, Haodong Duan, Yanan Sun, Zhening Xing, Wenran Liu, Kaifeng Lyu, and Kai Chen. 2025. Lego-puzzles: How good are mllms at multi-step spatial reasoning? *arXiv preprint arXiv:2503.19990*.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, et al. 2025. Kimi k1.5: Scaling reinforcement learning with llms. *Preprint*, arXiv:2501.12599.

Qwen Team. 2024. Qvq: To see the world with wisdom.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Graham Todd, Tim Merino, Sam Earle, and Julian Togelius. 2024. Missed connections: Lateral thinking puzzles for large language models. In *2024 IEEE Conference on Games (CoG)*, pages 1–8. IEEE.

Vernon YH Toh, Yew Ken Chia, Deepanway Ghosal, and Soujanya Poria. 2025. The jumping reasoning curve? tracking the evolution of reasoning performance in gpt-[n] and o-[n] models on multimodal puzzles. *arXiv preprint arXiv:2502.01081*.

Clinton J Wang, Dean Lee, Cristina Menghini, Johannes Mols, Jack Doughty, Adam Khoja, Jayson Lynch, Sean Hendryx, Summer Yue, and Dan Hendrycks. 2025. Enigmaeval: A benchmark of long multimodal reasoning challenges. *arXiv preprint arXiv:2502.08859*.

Taylor Webb, Zachary Dulberg, Steven Frankland, Alexander Petrov, Randall O'Reilly, and Jonathan Cohen. 2020. Learning representations that support extrapolation. In *International conference on machine learning*, pages 10136–10146. PMLR.

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

X.ai. 2025. Grok 3 beta.

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, et al. 2025. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Yunhua Zhou, and Xipeng Qiu. 2025. Revisiting the test-time scaling of o1-like models: Do they truly possess test-time scaling capabilities? *Preprint*, arXiv:2502.12215.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. 2019. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5317–5327.

Terry Yue Zhuo, Vu Minh Chien, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widyasari, Imam Nur Bani Yusuf, et al. 2025. Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions. In *The Thirteenth International Conference on Learning Representations*.

A License

Our benchmark, *Big Escape Benchmark*, is constructed using problems derived from publicly broadcast television programs. We do not distribute the original video or audio content from these programs; instead, the benchmark consists of questions, answers, and necessary visual cues (e.g., specific screenshots or descriptions of on-screen information) extracted from limited, essential portions of the source material solely for the purpose of creating a multimodal reasoning evaluation dataset. Similar to other academic benchmarks utilizing copyrighted material (e.g., Hendrycks et al., 2021), we operate under the principle of Fair Use (§107 of the U.S. Copyright Act), which permits the use of copyrighted work for purposes such as criticism, comment, news reporting, teaching, scholarship, or research. In determining whether the use made of a work in any particular case is a fair use, factors to be considered include the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes; the nature of the copyrighted work; the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and the effect of the use upon the potential market for or value of the copyrighted work. Our specific use falls under non-profit research and educational purposes, utilizing only limited, necessary portions relative to the copyrighted work as a whole, and this limited, transformative use for creating a research benchmark is unlikely to substitute for the original work and thus has no significant adverse effect on its market value. We release the *Big Escape Benchmark* benchmark dataset and its associated materials under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0). This license permits users to share and adapt the benchmark for non-commercial purposes, with appropriate attribution, under the same license. Furthermore, the collected problems and data are intended for academic and non-commercial research purposes only, and users are explicitly prohibited from using the *Big Escape Benchmark* benchmark dataset or any part thereof to train models that will be evaluated on this benchmark, or for any commercial purposes. Users are responsible for ensuring their own compliance with applicable copyright laws and the terms of this license.

B Detail Statistics

Table 4: **Other Statistics of *Big Escape Benchmark*.** *Big Escape Benchmark* derived from diverse television programming sources.

Category	Statistics
The Great Escape	113
The 1% Club	125
Escape! with Janet Varney	4
EXIT	5
Catchphrase	5
Avg. Question Len.	133.88 tokens
Different Task Prompts	210

B.1 Reasoning Catagres

Initially, all problems in Big Escape Bench were provided to a LLM tasked with identifying and summarizing the core reasoning abilities required. This analysis yielded five overarching reasoning categories: Multimodal Fusion Reasoning, Spatial Visual Reasoning, Logical Reasoning, Deductive Reasoning and Quantitative Reasoning.

B.2 Problem Types

To systematically categorize the reasoning skills assessed by *Big Escape Benchmark*, a multi-stage classification process was implemented. This process aimed to define problem types with appropriate granularity and ensure alignment with established benchmarks for comparability.

Fine-Grained type generation and standardization. The LLM was employed again (utilizing the prompt detailed in Appendix Figure 6) to perform a finer-grained tagging of problems within the five broad categories. This initial pass resulted in 84 distinct, highly specific problem subtypes.

Standardization and alignment. To ensure the granularity of method's problem types was comparable to existing multimodal benchmarks, we aligned our classifications with the typology used in MMIQ (Cai et al., 2025). MMIQ defines eight primary problem types: Temporal Movement, Spatial Relationship, 2D-Geometry, 3D-Geometry, Logical Operation, Concrete Object, Visual, and Instruction Mathematics. An LLM was tasked with mapping our 84 initial subtypes to these MMIQ categories as a standard.

Final *Big Escape Benchmark* promblem types. This alignment process consolidated the initial 84 subtypes into 21 distinct problem types for *Big Escape Benchmark*. This standardized set ensures that our problem type distribution can be meaningfully compared to other benchmarks while accurately reflecting the diversity of reasoning challenges within Big Escape Bench.

B.3 Error Categories

Shown in Table 5.

C Prompt

C.1 Evaluation prompt

C.2 Problem extraction prompt

C.3 Problem type classification prompt

D Error analysis

To further analyze model performance, we selected three representative models: ChatGPT-4o-latest, as a leading closed-source model; Qwen-2.5-72B-Vision-Instruct, as a prominent open-source MLLM; and o4-mini, noted for its specialized reasoning capabilities. Errors made by these models are categorized into three main types: (1) **Textual Comprehension Errors (TCE)**, subdivided into Omission of Textual Information (OTI), Misinterpretation of Textual Information (MTI), and Exclusive Reliance on Textual Information (TIO). (2) **Visual Comprehension Errors (VCE)**, subdivided into Omission of Visual Clues (OVC), Misinterpretation of Visual Information (MVI), and Exclusive Reliance on Visual Information (VIO). (3) **Reasoning Errors (RE)**, subdivided into Goal Misunderstanding (GM), Wrong Reasoning Idea (WRI), Intermediate Steps Error (ISE), and Conclusion Derivation Error (CDE). Detailed definitions for all error categories and their sub-types are provided in Table 5 of Appendix B.3. Error classification follows a sequential protocol: an error is assigned to a category only if it does not meet the criteria for any higher-priority category in the defined order. A visual breakdown of the error distributions for these selected models across text-only and multimodal tasks is presented in Figure 7.

Reasoning errors dominate and are primarily caused by flawed reasoning strategies. Across both text-only and multimodal tasks, reasoning errors (RE) consistently represent the most frequent failure mode for all evaluated models. In the text-only setting, RE accounts for 91.9% of errors in ChatGPT-4o-latest, 85.7% in o4-mini, and 90.6% in Qwen2.5-VL-72B-Instruct. This trend persists in multimodal scenarios. Within the RE category, the most common root cause is wrong reasoning ideas (WRI). For example, WRI constitutes 61.4% of RE cases in ChatGPT-4o-latest and 76.6% in Qwen2.5-VL-72B-Instruct. These findings suggest that current models frequently fail not due to misunderstanding the question or content, but due to selecting incorrect inferential paths, indicating a fundamental misalignment with human-like reasoning strategies.

Stronger models may over-interpret textual information in multimodal tasks. In multimodal tasks, we observe an emerging trend where models with stronger reasoning ability exhibit a higher proportion of textual comprehension errors (TCE). Notably, o4-mini—despite achieving the fewest total errors—records a TCE rate of 22.5%, substantially higher than ChatGPT-4o-latest (7.4%) and Qwen2.5-VL-72B-Instruct (2.3%). This suggests that more capable models may exhibit a tendency to overanalyze or over-rely on textual information, potentially leading to hallucinations or distraction from relevant visual cues. These results highlight a possible trade-off between general reasoning ability and robustness in multimodal grounding.

Visual interpretation remains a bottleneck for weaker multimodal models. Visual comprehension errors (VCE) are especially prominent among lower-performing models in multimodal tasks, often approaching or exceeding the frequency of reasoning errors. The dominant subcategory is *misinterpretation of visual information* (MVI), where models fail to correctly interpret visual attributes, object states, or spatial relationships. This indicates that while detection of visual elements may be successful, deeper understanding and integration of visual semantics into reasoning remain significant challenges. Improving this capability is essential for advancing performance in complex, vision-grounded reasoning tasks.

Table 5: Error case and definition

Error Case	Definition
Textual Comprehension Errors (TCE)	
Omission of Textual Information (OTI)	The model overlooks key textual information provided in the prompt or related context.
Misinterpretation of Textual Information (MTI)	The model incorrectly interprets the provided textual information.
Textual Information Only (TIO)	The model relies solely on textual information, ignoring necessary visual information for problem-solving.
Visual Comprehension Errors (VCE)	
Omission of Visual Information (OVC)	The model overlooks critical visual details or clues essential for understanding or problem-solving.
Misinterpretation of Visual Information (MVI)	The model incorrectly interprets visual information, such as misidentifying objects or their attributes.
Visual Information Only (VIO)	The model relies solely on visual information, ignoring necessary textual information for problem-solving.
Reasoning Errors (RE)	
Goal Misunderstanding (GM)	The model misunderstands the primary objective or the core aspect the question aims to address.
Wrong Reasoning Idea (WRI)	The model understands the goal but employs an incorrect initial reasoning approach.
Intermediate Steps Error (ISE)	The model's overall reasoning approach is sound, but an error occurs in one or more intermediate steps.
Conclusion Derivation Error (CDE)	The model's reasoning approach is correct, but an error is made in deriving the final conclusion.

Prompt 1: ***Prompt for evaluation***

You are playing an escape room puzzle game, and you need to use clues to solve the puzzle in front of you. You must provide a single, definitive answer.

Puzzle:

{task} Clues: {clues} Let's think step by step and put the final answer in \ boxed{{}}. Like this: \ boxed{{THE ANSWER}}.

Figure 4: Prompt for evaluation

Prompt 2: *Prompt for puzzle extraction*

Role: Escape Room Puzzle Extraction and Analysis Expert
Profile
- Language: Chinese
- Description: Accurately extract all puzzles from the subtitles of the show "Escape Room" and conduct systematic logical analysis and organization.
Goal
Comprehensively identify all puzzles and provide complete time ranges, problem statements, requirements, clues, reasoning logic, and correct answers for each puzzle.
Skills
- Accurately identify various types of puzzles and Q&A questions, ensuring nothing is missed.
- Define the complete time range of each puzzle, covering the entire process from appearance to resolution.
- Filter core information, removing irrelevant dialogue and content unrelated to the puzzle.
- Construct a rigorous logical reasoning chain to ensure each puzzle has a unique answer.
Rules
1. **Comprehensive Puzzle Identification**:
- Identify as many puzzles as possible, ensuring none are overlooked.
2. **Precise Time Positioning**:
- Provide the complete time range for each puzzle, including the discovery, thinking, and resolution process. - Time markers must be accurate, formatted as xx:xx:xx,xxx –> xx:xx:xx,xxx.
3. **Information Filtering and Organization**:
- Retain only core information related to the puzzle, removing irrelevant dialogue (such as casual chat or variety show effects).
- Ensure clues and information have internal logical consistency to aid in reasoning and solving.
4. **Logical Reasoning Construction**:
- Build a complete reasoning chain, ensuring logical rigor.
- Ensure each puzzle can be solved to a unique correct answer using the provided clues.
5. **Standardized Output Format**, ensuring clear structure:
#Number#: {Puzzle Number}
#Time#: {xx:xx:xx,xxx –> xx:xx:xx,xxx}
#Task#: {Puzzle Task Description, clearly stating the problem to be solved and the required answer format}

Figure 5: Prompt for problem extraction.

Prompt 3: *Prompt for Problem type classification*

You are now a senior puzzle capability analyzer.
Your task is to conduct a detailed skill point analysis of the **single puzzle** I provide. You need to identify 1-3 of the most core **Fine-grained Skills** that the puzzle tests and classify each skill point into one of the predefined 5 **Macro-Types**.
Definition of Macro-Types (must strictly follow):

1. **Linguistic_Reasoning**: word/letter games, homonym/spelling/idioms, semantic understanding and disambiguation, text structure analysis, etc.
    - *Fine-grained Skills examples*: anagrams, rhyming, word search, sentence completion, synonym/antonym.
2. **Quantitative_Reasoning**: numerical patterns, arithmetic operations, number counting, numeral system conversion, date/time calculation, basic algebra, probability and statistics, etc.
    - *Fine-grained Skills examples*: arithmetic sequence, percentage calculation, unit conversion, basic algebra, counting objects.
3. **Spatial_Visual_Reasoning**: figure rotation/flip, spatial folding, mirror symmetry, geometric figure counting, view transformation (top view/side view), path planning and tracking, map reading, etc.
    - *Fine-grained Skills examples*: mental rotation, pattern folding, 2D to 3D visualization, maze solving, visual pattern recognition.
4. **Logical_Deductive_Reasoning**: rule-based deduction, conditional judgment, permutation and combination, truth deduction, logic grid puzzles, procedural logic, causal relationship analysis, etc.
    - *Fine-grained Skills examples*: deductive inference, conditional logic, truth-table evaluation, constraint satisfaction, sequence deduction.
5. **Multimodal_Fusion_Reasoning**: requires simultaneous integration and reasoning of image and text, audio and text, or multiple sensory information to solve the puzzle.
    - *Fine-grained Skills examples*: image-text matching, audio-based instruction following, visual data interpretation with text query.

Figure 6: Classify problem type prompt.

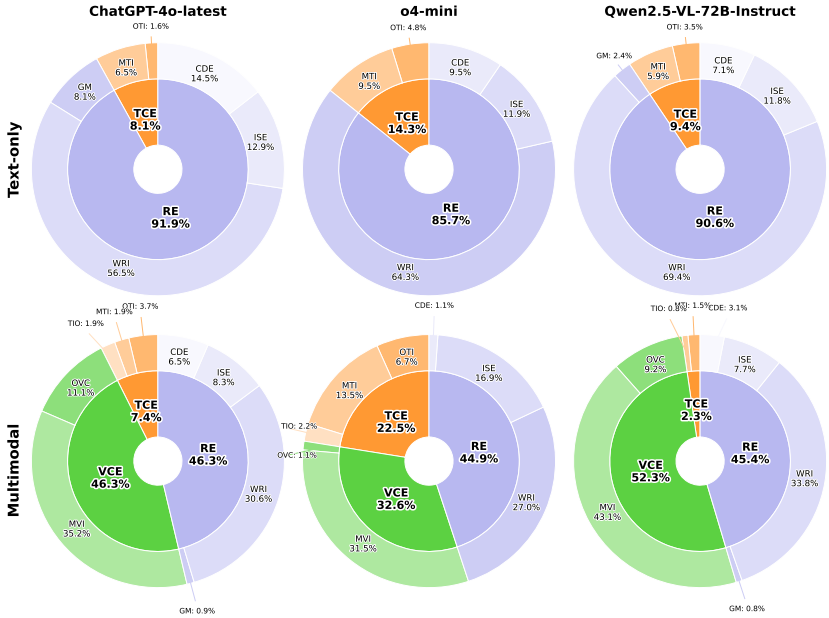

Figure 7: **Error distributions for three selected models across text-only and multimodal tasks.** Each chart illustrates the proportion of main error categories along with their respective sub-categories.