

000 PUZZLEWORLD: A BENCHMARK FOR MULTIMODAL, 001 002 OPEN-ENDED REASONING IN PUZZLEHUNTS 003 004

005 **Anonymous authors**

006 Paper under double-blind review

007 008 ABSTRACT 009

010 Puzzlehunts are a genre of complex, multi-step puzzles lacking well-defined prob-
011 lem definitions. In contrast to conventional reasoning benchmarks consisting
012 of tasks with clear instructions and constrained environments, puzzlehunts re-
013 quires discovering the underlying problem structure from multimodal evidence
014 and iterative reasoning, mirroring real-world domains such as scientific discov-
015 ery, exploratory data analysis, or investigative problem-solving. Despite progress
016 in foundation models, their performance on open-ended settings remains largely
017 untested. We introduce PUZZLEWORLD, a comprehensive benchmark of 667
018 puzzlehunt-style problems designed to assess step-by-step, open-ended, and cre-
019 ative multimodal reasoning. Each puzzle is annotated with the final solution, de-
020 tailed reasoning traces, and cognitive skill labels, enabling holistic benchmarking
021 and fine-grained diagnostic analysis. Most state-of-the-art models achieve only 1-
022 4% final answer accuracy. On PUZZLEWORLD, the best model solves only 14%
023 of puzzles and reaches 40% stepwise accuracy, matching human puzzle novices
024 but falling significantly behind puzzle enthusiasts. To demonstrate the value of
025 our reasoning annotations, we show that fine-tuning a small model on reasoning
026 traces boosts stepwise accuracy from 4% to 11%, which translates to improve-
027 ments in downstream visual reasoning tasks. Our detailed error analysis reveals
028 that current models exhibit myopic reasoning, are bottlenecked by the limitations
029 of language-based inference, and lack sketching capabilities crucial for visual and
030 spatial reasoning. We will publicly release PUZZLEWORLD to support future
031 work on building more general, open-ended, and creative reasoning systems.

032 1 INTRODUCTION

033 Recent advances in language and multimodal reasoning (Liang et al., 2024b) have enabled sig-
034 nificant progress in step-by-step problem-solving (Wei et al., 2022; Yao et al., 2023), transparent
035 reasoning (Creswell & Shanahan, 2022; Luo et al., 2023), and enhanced human-AI collaboration
036 (Wu et al., 2022; Chen et al., 2025b). Such progress has been fuelled by and evaluated on com-
037 prehensive benchmarks, particularly in domains like mathematics (Lu et al., 2024) and code (Jiang
038 et al., 2024). However, these benchmarks are largely confined to narrow, well-defined environments.
039 In coding, tasks are meticulously specified and validated within executable environments (Jimenez
040 et al., 2024). In geometry, models often rely on domain-specific languages to structure their rea-
041 soning (Chervonyi et al., 2025). While valuable, these benchmarks primarily test a model’s ability
042 within a pre-defined problem space, rather than its ability to *discover the problem itself*.

043 In contrast, human reasoning excels in open-ended environments, where the rules are unstated and
044 the objectives are ambiguous. We dynamically form hypotheses, adapt to implicit structures, and
045 reason creatively across modalities to solve problems ranging from deciphering an escape room to
046 novel scientific discovery. To build more generalist AI, we argue that the next frontier for eval-
047 uation lies beyond the current constrained settings. It demands benchmarks that challenge models
048 to operate in less structured, discovery-driven environments that require more flexible and holistic
049 reasoning (Mondorf & Plank, 2024).

050 Puzzles are designed precisely to test these abilities. While some are rigidly formatted like Sudokus,
051 others, like *puzzlehunts*, are intentionally open-ended. In a puzzlehunt, solvers are not given a clear
052 task; they must first infer the nature of the problem from ambiguous clues embedded in text, images,
053 or cultural references before devising and executing a solution.

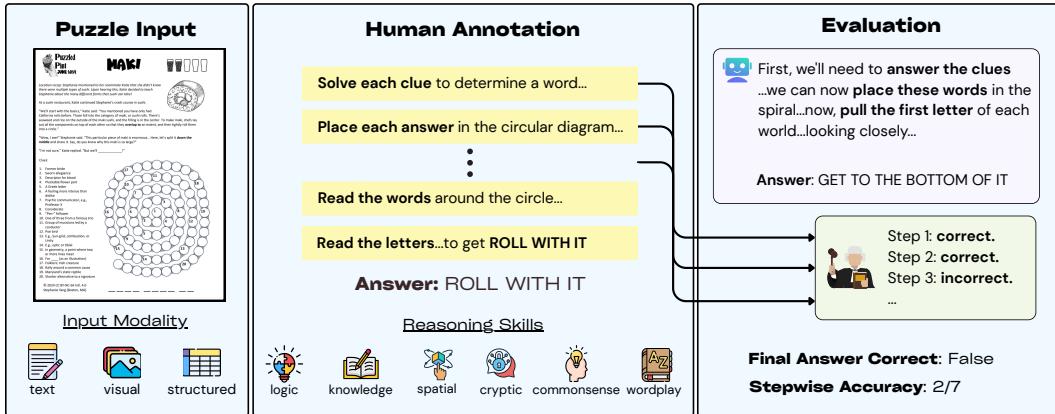


Figure 1: **Overview of PUZZLEWORLD:** PUZZLEWORLD is a dataset of complex puzzles that lack explicit instructions, requiring solvers to deduce the final answer from nuanced, multimodal cues from the puzzle content as well as external domain-specific knowledge. The raw puzzles and solutions are sourced from Puzzled Pint, and the solutions, which are PNG images, are transcribed into a sequence of reasoning steps by human annotators. These annotations enable us to measure the accuracy of the final answer and the step-by-step progress made towards the solution. Best-viewed zoomed in and in color, high-resolution puzzles are in Appendix E.3.

Beyond their entertainment value, puzzlehunts model the essential challenges of real-world discovery and analysis. They demand compositional thinking, lateral reasoning, and the resilience to pursue leads, backtrack from dead ends, and manage uncertainty. Unlike current AI benchmarks that present well-specified tasks, puzzlehunts compel solvers to discover both *what* the problem is and *how* to solve it. This dual challenge makes them uniquely suited for evaluating general-purpose reasoning systems under conditions that more closely resemble open-ended scenarios like scientific investigation, intelligence analysis, or exploratory design.

To bridge this gap, we introduce PUZZLEWORLD, a benchmark of 667 real-world puzzlehunt problems curated from Puzzled Pint (Puzzled Pint, 2025), a monthly puzzlehunt event with content released under a Creative Commons license. These puzzles offer an open-ended, compositional challenge beyond prior benchmarks focused on instruction-following or task completion, and will grow with new puzzle releases. For each puzzle, we provide fine-grained annotation of its solution, input modalities, cognitive reasoning skills it exercises, and a manually curated step-by-step solution trace. These rich annotations support diagnostic analysis, model training, and detailed evaluation of models’ reasoning capabilities. An overview of PUZZLEWORLD is provided in Figure 1.

PUZZLEWORLD enables us to systematically study the multimodal and multi-step reasoning capabilities of today’s best foundation models. Most state-of-the-art models achieve only 1-4% final answer accuracy, with the best model solving only 14% of puzzles and reaching 40% stepwise accuracy. We additionally find that detailed annotations are important, as fine-tuning a model on annotated reasoning traces significantly improves a small model’s performance, both within PUZZLEWORLD and on other visual reasoning datasets. We also conduct detailed error analysis on models’ performance on PUZZLEWORLD, yielding tangible directions for future work in improving multimodal open-ended reasoning in AI. Together, these elements position PUZZLEWORLD as a rigorous resource for evaluating and improving general-purpose multimodal reasoning in AI systems. In the long run, we believe PUZZLEWORLD can catalyze more general and adaptable AI for mathematical and logical reasoning, open-ended scientific discovery, and assistive agents.

2 RELATED WORK

Large Language Model (LLM) Reasoning. LLMs have demonstrated remarkable emergent capabilities, often matching or even surpassing human performance across a wide range of tasks (Street et al., 2024). Notably, models such as GPT-4 (Achiam et al., 2023) and Claude (Anthropic, 2025) have achieved strong results not only on traditional NLP benchmarks—like question answering, summarization, and translation (Widyassari et al., 2022; Soares & Parreira, 2020; Singh et al., 2017), but also in more complex domains such as mathematical reasoning, programming, and log-

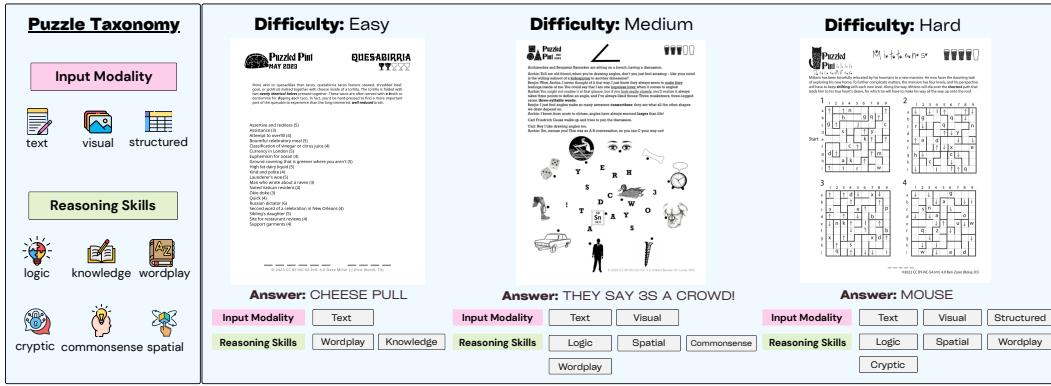


Figure 2: **Overview of samples from PUZZLEWORLD.** **Left:** To gain a deeper understanding of model performance on PUZZLEWORLD, each puzzle is annotated with the input modalities of the puzzle content, the reasoning skills required to solve the puzzle, and step-by-step reasoning steps. **Right:** Example modality and reasoning skill annotations on three puzzles. High-resolution puzzle images are in Appendix E.3.

ical deduction (Ahn et al., 2024; Jiang et al., 2024; Lam et al., 2024). These abilities suggest that LLMs are beginning to exhibit general-purpose reasoning skills, making them increasingly relevant to both academic research and practical applications. However, despite these impressive capabilities, understanding the full extent and limitations of LLM reasoning remains a crucial open question, underscoring the need for benchmarks that rigorously assess their capability for flexible, holistic reasoning (Mondorf & Plank, 2024; Chang et al., 2024).

Reasoning Benchmarks. Numerous reasoning benchmarks have been proposed to evaluate various cognitive skills, including visual mathematical reasoning (Lu et al., 2024), spatial understanding (Wang et al., 2024a), analogical reasoning (Yiu et al., 2024), and social reasoning (Li et al., 2025; Mathur et al., 2025). However, few have addressed abstract, open-ended problems that demand holistic reasoning. HEMM (Liang et al., 2024a), SciBench (Wang et al., 2024b), MMMU (Yue et al., 2024a), MMMU-Pro (Yue et al., 2024b), MMT-Bench (Ying et al., 2024), and Olympiad-Bench (He et al., 2024) test multimodal reasoning across various disciplines in academic and real-world contexts. While these tasks are broad and challenging, they typically involve well-defined questions that closely resemble the training distributions of large models. As such, they primarily assess in-distribution reasoning rather than creativity or adaptability. ARC-AGI (Chollet, 2019) tests the ability to reason and adapt to new situations through abstract visual pattern recognition tasks that require minimal prior knowledge, yet it lacks the open-ended, exploratory nature of real-world problem solving. In contrast, PUZZLEWORLD targets open-ended reasoning through puzzlehunts that lack explicit instructions. Solving these tasks requires creatively piecing together subtle hints, often across many modalities, into coherent multi-step reasoning chains.

Puzzle Benchmarks. A growing line of work has explored the use of puzzles to test the reasoning capabilities of AI systems. PuzzleVQA (Chia et al., 2024) consists of 2,000 puzzles that require abstracting patterns from visual puzzles to answer multiple-choice questions. AlgoVQA (Ghosal et al., 2024) is a visual puzzle benchmark requiring algorithmic reasoning. PUZZLES (Estermann et al., 2024) tests the ability of RL agents to perform algorithmic reasoning on a set of 40 puzzles. While valuable for evaluating specific skills, these benchmarks focus on narrow domains with constrained task formats, and modern models generally perform well on these benchmarks (Chia et al., 2024; Moskvichev et al., 2023; Yue et al., 2024a). On the other hand, the unstructured nature of the puzzlehunt problems in PUZZLEWORLD requires models to interpret ambiguous cues, explore creative strategies, and integrate information across diverse modalities and knowledge areas. In contrast to previous benchmarks that isolate strictly structured vertical reasoning (Chen et al., 2025a) or narrative-based lateral thinking (Huang et al., 2024), puzzlehunt puzzles require an integrated blend of lateral thinking, symbolic abstraction, and visual–spatial reasoning. The closest to our benchmark is EnigmaEval (Wang et al., 2025), which also evaluates AI’s reasoning capabilities on puzzlehunts. However, EnigmaEval is a closed-source evaluation-only dataset and does not include manually annotated step-by-step solutions. The open-sourced puzzles and rich annotations in PUZZLEWORLD support fine-grained analysis of intermediate reasoning and failure modes, facilitating the development and evaluation of more robust, general-purpose reasoning models.

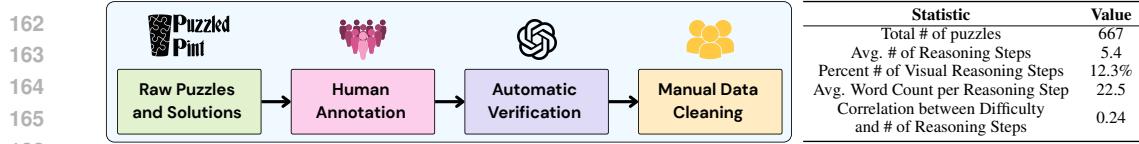


Figure 3: **Dataset construction procedure and statistics:** **Left:** First, we source raw puzzles and solutions from Puzzled Pint. As the Puzzled Pint solutions are often not correctly parsed by OCR, each puzzle’s metadata and reasoning steps are human-annotated. We use GPT-4o to automatically flag ambiguous and inconsistent annotations. Finally, two human verifiers perform a manual data cleaning on the flagged puzzles to ensure a consistent annotation format. **Right:** We summarize the statistics of our dataset. The average number of reasoning steps is high, and the steps are relatively complex, as shown by the high average word count.

3 TAXONOMIZING MULTIMODAL REASONING IN PUZZLEHUNTS

To understand how solving puzzlehunts engages reasoning capabilities evaluated separately in benchmarks like MMMU (Yue et al., 2024b), we analyze puzzle solutions and classify them along two dimensions: input modality and reasoning mechanism. This taxonomy provides a comprehensive evaluation framework that captures both the form in which information is presented and the cognitive strategies required for reasoning.

3.1 PUZZLE INPUT MODALITIES

We consider three puzzle input modalities: **Text**, encompassing textual information such as instructions, narratives, or word puzzles, testing the model’s ability to extract relevant linguistic information; **Visual**, which includes unstructured visuals like images, icons, and typography, challenging the models to interpret visual semantics and patterns; and **Structured**, which refers to systematically organized visual information, such as tables, graphs, grids, matrices, and charts. Table 1 shows the distribution of puzzles across modality and difficulty.

3.2 PUZZLE REASONING MECHANISMS

We identify six core cognitive abilities essential for effective puzzle-solving in PUZZLE-WORLD. These include **logic**, which covers inferential reasoning such as deduction and causal inference; **wordplay**, involving flexible linguistic interpretation through puns, anagrams, and homophones; **spatial reasoning**, which tests an AI’s ability to mentally manipulate objects and navigate structures; and **cryptic decoding**, which requires recognizing and applying transformations like ciphers and hidden encodings. In addition, **knowledge-based reasoning** leverages domain-specific facts from areas such as science or history, while **commonsense reasoning** draws on implicit real-world expectations. This taxonomic approach enables targeted evaluation and analysis of AI reasoning capabilities across different cognitive dimensions. By mapping specific puzzles and reasoning tasks to combinations of modalities and mechanisms, we can identify areas of strength and weakness in AI systems, track progress over time, and guide future development efforts toward more balanced reasoning capabilities.

4 CREATING PUZZLEWORLD

4.1 DATA COLLECTION AND PRE-PROCESSING

We collected our puzzle corpus from Puzzled Pint (2025), an organization that publishes puzzles under Creative Commons (CC BY-NC-SA Intl. 4.0). Their repository contains monthly puzzles designed for collaborative solving, covering a diverse range of puzzle types and difficulties. This allowed us to obtain more than 700 raw puzzles spanning from 2010 to 2025.

Each puzzle in our dataset consists of its original PDF containing the puzzle content, a single-phrase answer, and a solution document. Unlike Wang et al. (2025), we deliberately preserved the original puzzle format rather than transcribing content into separate text and images. This decision was motivated by the importance of spatial relationships in puzzle layouts to the solving process. Fur-

thermore, Wang et al. (2025) showed that the best foundation models are not primarily constrained by OCR capabilities. Instead, we devote our manual effort to construct fine-grained annotations of puzzle reasoning steps, ensuring that the annotations accurately capture the intended solution pathways while maintaining the integrity of the original puzzle presentation.

4.2 DATA ANNOTATION

To facilitate AI’s reasoning capabilities, we designed a comprehensive annotation structure for PUZZLEWORLD. Each puzzle is represented by a standardized metadata and visual assets. To prevent ambiguity, we discard puzzles that have incomplete solutions, multiple ground truth answers, or require physical activity to solve the puzzles. This leaves us with 667 annotated puzzles.

4.2.1 METADATA SCHEMA

Each puzzle is annotated using a JSON schema comprising several fields: a descriptive **title**; **flavor text** providing narrative context; a **difficulty** label (easy, medium, or hard); **solution** representing the canonical answer; a **reasoning** field of an ordered sequence of steps leading to the solution; a **modality** tag specifying the input types involved; a list of **skills** capturing the cognitive abilities required for solving; and a **source** field attributing the data origin. Figure 4 illustrates an example annotation.

4.2.2 REASONING ANNOTATION

A key contribution of our annotations is the decomposition of puzzle-solving into reasoning steps. Each step is formalized as a tuple $\langle e, f \rangle$ where e represents the textual explanation and f denotes an optional figure illustrating the reasoning. To ensure annotation consistency, we loosely require each step to begin with an atomic operation, such as pattern discovery or sketching, followed by the intermediate outcome of that operation. This structured annotation enables fine-grained analysis of an AI’s reasoning trajectory.

4.3 VERIFICATION OF ANNOTATIONS AND DATA CONTAMINATION

To ensure annotation quality and integrity, we implemented a two-stage verification protocol. First, we used GPT-4o to flag each puzzle annotation for correctness and reasoning coherence. This automated screening identified reasoning steps exhibiting ambiguity or logical discontinuities that might impede systematic analysis, which has flagged 12.11% of the dataset. Subsequently, two human verifiers independently reviewed all flagged annotations, applying corrections where necessary. This verification process resulted in modifications to 10.93% of the initially annotated puzzles. As an additional quality assurance measure, we conducted manual verification of a random subset comprising 5% of the dataset. In this evaluation, 96.5% of the verified annotations are marked as correct by the verifiers, demonstrating the high reliability of our annotation methodology. Finally, we verify whether frontier models has memorized any of the puzzles in PUZZLEWORLD. We describe our procedure in E.1, where we find no evidence of data contamination.

4.4 DATASET STATISTICS

We summarize key statistics in Figure 3 (right). The average number of reasoning steps is above 5, and the average word count per reasoning step is above 20, demonstrating the complexity of the reasoning traces. Additionally, 12.3% of the steps have a visual intermediate output, highlighting the importance of sketching and spatial reasoning to solve puzzles. The correlation between puzzle difficulty and # of reasoning steps is 0.24. While we expect difficulty and # of reasoning steps to be positively correlated, the magnitude of the correlation is relatively low, as the difficulty of the puzzles also stems from their open-ended nature. Figure 5 shows the distribution of puzzles by modalities, reasoning skills, number of reasoning steps, and difficulty.

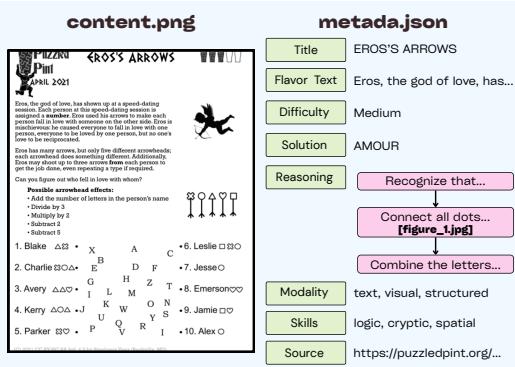


Figure 4: Illustration of metadata schema: All puzzles are annotated with accompanying metadata, which includes the title, flavor text, difficulty, final answer, reasoning steps, input modalities, reasoning skills, and the link to the puzzle.

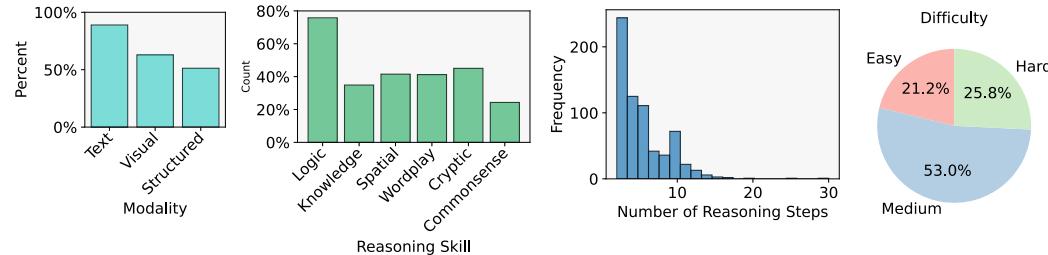


Figure 5: **PUZZLEWORLD dataset statistics.** Distributions of modalities and reasoning skills are balanced. While the majority of puzzles are of medium difficulty, there is significant number of easy and hard puzzles. The number of reasoning steps follows a long-tail distribution, with many solutions requiring more than 5 steps and some hard puzzles requiring up to 30 steps of reasoning.

5 EXPERIMENTS

In this section, we evaluate frontier closed and open-source multimodal LLMs on the PUZZLEWORLD dataset. We detail the evaluation setup, present quantitative results, and conduct qualitative error analysis to understand model behavior in open-ended, multimodal puzzle reasoning.

5.1 EXPERIMENTAL SETUP

We evaluate frontier closed-source reasoning models on PUZZLEWORLD, including GPT-o3 (OpenAI, 2025), GPT-4o (Achiam et al., 2023), Claude Opus 4 (Anthropic, 2025), Gemini-2.5-Pro (Comanici et al., 2025), [Gemini-3-Pro \(Google, 2025\)](#), and Grok 4 (xAI, 2025). We also evaluate open-source models Qwen QVQ (Qwen, 2024), InternVL3 (Zhu et al., 2025), and Kimi VL A3B (Team et al., 2025). We prompt each model with a comprehensive prompt as in Wang et al. (2025), followed by the puzzle images and transcribed flavor text. See Appendix F for the evaluation prompt.

We also provide a human baseline on PUZZLEWORLD, considering three tiers of puzzlehunter expertise: **Novice**, with no prior puzzlehunt experience; **Enthusiasts**, who showed interest or have occasionally participated (1-2 sessions) in puzzlehunts, and **Experts** among the top teams at monthly Puzzled Pint meetings. We gathered 9 Novices and 9 Enthusiasts across high school and college ages. We sampled 5% puzzles from PUZZLEWORLD and assigned each participant to solve four puzzles. Participants were given an hour to solve each puzzle, matching the usual expected time at a live session, and were asked to provide paragraph explanations for their solution. For Experts, we use statistics from Puzzled Pint sessions in Syracuse, New York, and Bangalore, India, dating from January 2023 to June 2025. The statistics suggest that expert puzzlehunters consistently solve all five puzzles within one to two hours, which is on average less than the time prescribed to our human participants. We thus assume that human Experts achieve perfect accuracy on PUZZLEWORLD.

5.1.1 AUTOMATIC EVALUATION METRICS

Beyond final answer accuracy, we additionally evaluate the models' *stepwise accuracy* by comparing their solution with the annotated ground truth reasoning steps. Since puzzles can have multiple solution pathways, we define the stepwise accuracy score of a candidate solution to be the *last* annotated reasoning step it successfully executed out of all the reasoning steps. We implement an LLM judge (Zheng et al., 2023) with GPT-4o to determine the stepwise score of each candidate solution. For each reasoning step in the reference solution, the LLM judge determines if the step is met by the candidate response. To evaluate LLM judge's reliability, we compared its stepwise evaluations on 20 random puzzles against human evaluations. The LLM judge achieved a Pearson correlation of $r = 0.829$ ($p = 6.3 \times 10^{-6}$) and a mean absolute error (MAE) of 0.083 with respect to human scores, indicating strong alignment with human judgment.

5.2 RESULTS

5.2.1 OVERALL PERFORMANCE OF FRONTIER MODELS

We report the models' performance in Table 2. All models exhibit extremely low final answer accuracy on PUZZLEWORLD, with most achieving close to 1-4%. GPT-o3 attains the highest overall accuracy at 14.22%, matching human Novice performance, while the best-performing open-source model, QVQ-72B-Preview, reaches just 1.36%. All models perform significantly worse than human Enthusiasts and Experts. Although the uniformly low accuracy underscore our benchmark's difficulty, it offers limited insight into the models' reasoning capabilities.

Table 2: **Model performance.** Accuracy (Acc) and stepwise accuracy (Step) are reported overall and per modality. Models struggle significantly on PUZZLEWORLD, most achieve only 1-4% answer accuracy. The best model, GPT-o3, solves only 14% of puzzles and reaches 40% stepwise accuracy, matching human Novice performance but falling behind Enthusiasts.

Model	Overall		Text		Visual		Structured		
	Acc	Step	Acc	Step	Acc	Step	Acc	Step	
Open	QVQ-72B-Preview	1.36	30.23	1.33	29.25	0.63	27.96	1.18	32.40
	InternVL3-78B	0.89	15.49	0.83	14.80	0.47	14.48	1.15	17.97
	Kimi VL A3B	1.33	19.10	1.16	17.91	0.94	18.84	1.72	21.41
Closed	GPT-o3	14.22	39.81	15.16	39.92	8.96	33.38	13.53	41.28
	GPT-4o	1.83	22.09	1.92	20.00	0.73	20.20	2.77	28.09
	Claude Opus 4	4.50	24.56	4.20	23.77	4.04	22.60	4.37	26.93
	Gemini 2.5 Pro	7.65	31.61	8.07	31.09	4.99	29.06	6.71	32.34
	Gemini 3 Pro	18.0	39.99	20.30	39.34	14.71	38.81	20.25	39.99
	Grok 4	3.33	13.79	3.85	13.64	3.70	14.19	1.56	11.22
Human	Human Novice	13.89	23.10	16.98	25.32	11.00	22.70	16.67	24.92
	Human Enthusiast	44.44	51.70	44.14	52.58	44.00	52.20	54.17	57.81
	Human Expert	100.0							

To address this, our stepwise evaluation metrics provide a more nuanced view of models’ reasoning performance. These metrics reveal that models with poor final answer accuracy, such as InternVL3, still demonstrate good intermediate reasoning, achieving up to 15.49% stepwise accuracy. Similarly, while QVQ-72B-Preview lags behind closed-source models in final answer accuracy, it outperforms many of them in stepwise accuracy (30.2%), reflecting more coherent reasoning despite not reaching the correct final output. These two metrics enable PUZZLEWORLD to remain highly challenging while offering detailed diagnostics for model evaluation and development.

In terms of input modalities, models generally perform best on text-based puzzles, with significantly lower accuracy on puzzles involving unstructured visual inputs. Interestingly, most models achieve better performance on structured puzzles, such as crosswords where the spatial format constrained, over unstructured visual puzzles. In contrast, puzzles involving free-form visuals remain difficult, with models often achieving less than half their text puzzle accuracy on these inputs. These trends highlight current models’ persistent weaknesses in visual grounding and spatial reasoning.

5.3 IMPROVING REASONING ON DOWNSTREAM TASKS WITH PUZZLEWORLD

To explore whether PUZZLEWORLD can support model improvement, we fine-tuned an 8B Intern-VL3 model with supervised fine-tuning on annotated reasoning traces from 80% of the dataset, and evaluated performance on the 20% held-out test set. As a control, we fine-tuned the same model using only the final answers, without access to reasoning traces. Full details are provided in Appendix F.2.

Our results in Table 3 highlight the value of PUZZLEWORLD’s annotations. Fine-tuning on reasoning traces doubles the model’s stepwise accuracy—from 4.78% (base model) to 11.00%. In contrast, fine-tuning on final answers alone impairs performance, reducing stepwise accuracy to 2.96% and driving answer accuracy to zero. Despite the improved stepwise accuracy, the fine-tuned model’s answer accuracy remained at 0.76%. This suggests that while fine-tuning enhanced model’s intermediate reasoning, it was insufficient to solve additional puzzles completely. This result underscores both the difficulty of PUZZLEWORLD and the limitations of naive fine-tuning approaches in addressing such complex reasoning challenges.

We then explore PUZZLEWORLD’s detailed stepwise annotation can improve models on downstream reasoning tasks. We finetuned a model on 80% of PUZZLEWORLD and evaluated it on two benchmarks: a Rebus puzzles dataset (Lee et al., 2025) involving visual metaphors without explicit instructions, and the MathVista dataset (Lu et al., 2024), ranging from general visual question answering to domain-specific geometry questions. Our results are shown in Table 4.

378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431

Table 4: Fine-tuned model performance on downstream reasoning tasks. Finetuning on PUZZLEWORLD leads to performance gains, on visually-oriented tasks (Rebus puzzles, geometry, visual question answering) but slightly reducing it on problems less dependent on pure visual knowledge.

Dataset	Task	Base Model	Fine-tuned on PUZZLEWORLD
Rebus Puzzles	Puzzle reasoning	3.2%	5.1%
MathVista	Geometry problem solving	65.87%	66.35%
	Textbook question answering	63.92%	60.13%
	Math word problem	62.37%	59.14%
	Visual question answering	32.40%	39.11%

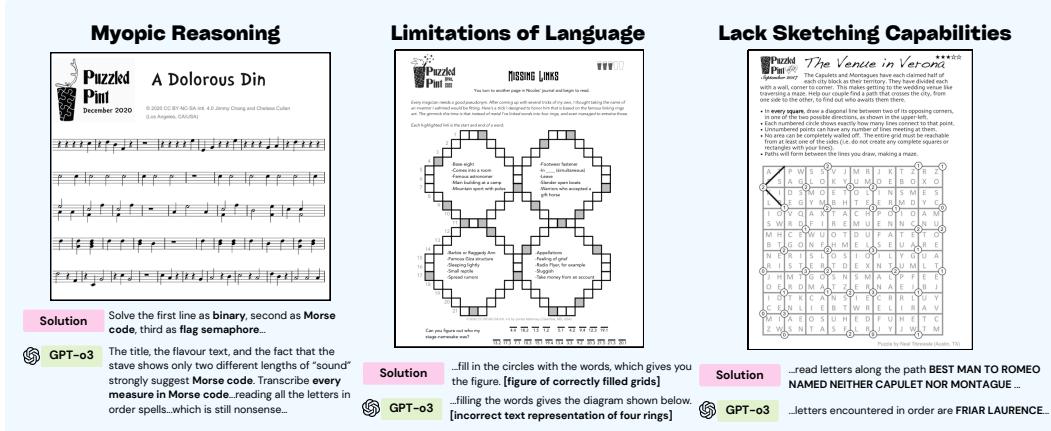


Figure 6: Example puzzle errors. **Left:** (myopic reasoning) The model is unable to backtrack when it hits a dead end. **Middle:** (language bottleneck) The model misrepresents the visual contents due to inherent limitations of texts. **Right:** (sketching errors) The model fails to execute the visual sketching steps to obtain correct intermediate outputs. High-resolution images are in Appendix E.3.

Finetuning on PUZZLEWORLD yields notable performance gains on Rebus puzzles, where the model’s accuracy increased from 3.2% to 5.1%. On MathVista, the model shows significant improvement in geometry problem solving and visual question answering, but its performance slightly decreased on tasks outside of PUZZLEWORLD’s reasoning skills, such as textbook question answering and math word problems. This performance improvement suggests that the skills learned from PUZZLEWORLD are not merely task-specific. They represent transferable, general-purpose reasoning capabilities, making our dataset a valuable tool for enhancing models’ capabilities.

5.4 DETAILED ERROR ANALYSIS

We highlight the main sources of errors by the best reasoning multimodal LLMs on PUZZLEWORLD, focusing on GPT-o3. See Figure 6 for example errors from each category.

Myopic reasoning. Despite strong performance on conventional benchmarks, frontier models often exhibit *myopic commitment* in their reasoning. Rather than exploring alternatives or revisiting prior steps, models tend to fixate on early, surface-level hypotheses, resulting in reasoning trajectories that are locally coherent but globally misaligned with the puzzle. For example, in Figure 6, solving the puzzle requires interpreting musical notes using a mix of binary, Morse code, and flag semaphores. Instead, GPT-o3 identifies a Morse code reference early on and rigidly adheres to it—even as contradictions arise—demonstrating a lack of backtracking and verification.

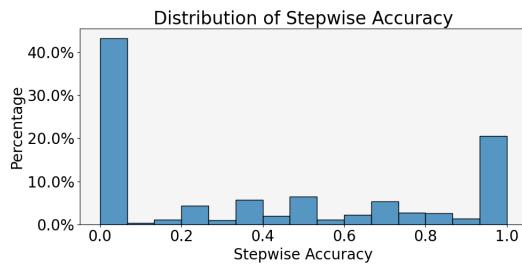


Figure 7: Stepwise accuracy distribution of GPT-o3. GPT-o3 receives stepwise accuracy of 0 for most puzzles, highlighting the model’s myopic reasoning tendencies and its inability to backtrack after committing to an incorrect first step.

432 To further examine this behavior, we analyze the stepwise accuracy distribution of GPT-o3 (Figure 7). We find that, on most puzzles, the model receive a score of 0, meaning the model often
 433 fails to correctly identify even the first step of the solution. Once committed to an incorrect path,
 434 the model rarely recovers, highlighting its brittle reasoning and a lack of dynamic self-correction,
 435 especially when it cannot rely on external environments for verification.
 436

437 **Limitations of language.** Modern multi-
 438 modal models rely heavily on language-based
 439 reasoning strategies, such as chain-of-thought
 440 and code generation. However, this dependence
 441 becomes a bottleneck in puzzles with complex
 442 visual structure. In Figure 6, the puzzle is com-
 443 posed of four interlocking loops arranged in a
 444 clover-like pattern. This layout is visually intui-
 445 tive, but difficult to represent in text.

446 While GPT-o3 correctly solves the word clues,
 447 it fails to capture the layout when converting the
 448 puzzle into text, as shown in Figure 8. This ul-
 449 timately leads the model to derive an incorrect
 450 answer. This example highlights a broader lim-
 451 itation: when faced with highly complex struc-
 452 tured inputs, models that default to textual reasoning often lose critical spatial information. This
 453 inherent mismatch between visual intuition and language-centric inference poses a fundamental
 454 challenge to models, especially those that depend on textual or code-based reasoning chains.

455 **Multimodal reasoning needs sketching.** While fron-
 456 tier models have made notable progress in logical deduc-
 457 tion and arithmetic reasoning, they consistently underper-
 458 form on spatial tasks that require sketching, drawing, and
 459 manipulating visual structure, such as decoding based on
 460 spatial arrangements or tracing paths through grids and
 461 mazes. In Figure 6, the model correctly solves the indi-
 462 vidual clues in a grid-based puzzle but fails to trace the
 463 intended path, resulting in an incorrect final answer. Hu-
 464 mans naturally rely on sketching or mental imagery to
 465 reason through such spatial challenges, using external or
 466 internal visualizations to keep track of evolving structure.
 467 The absence of such capabilities in current models reveals
 468 a critical gap: without the ability to sketch and update a
 469 persistent visual representation, models are prone to fail-
 ure in tasks that depend on spatial coherence.

470 To understand the impact of sketching to model performance, we manually analyzed 30 puzzles
 471 where GPT-o3 produced incorrect answers. For each failure, we annotated the reasoning step re-
 472 sponsible for the error with its corresponding reasoning skill. As shown in Figure 9, we found that
 473 53.33% of these bottleneck steps involved spatial reasoning or sketching-related capabilities. This
 474 highlights a gap in models’ ability to manipulate visual structure during inference. Incorporating
 475 sketch-like visual memory and reasoning (Wu et al., 2024; Hu et al., 2024; Chen et al., 2025b) may
 476 offer a promising direction toward more robust and spatially grounded reasoning AI.

477 6 CONCLUSION

478 This paper presents PUZZLEWORLD, a large-scale benchmark of 667 puzzlehunt-style problems de-
 479 signed to assess multi-step, open-ended multimodal reasoning. The diversity of puzzles and richly
 480 annotated reasoning traces enable holistic benchmarking and fine-grained diagnostics. PUZZLE-
 481 WORLD presents a unique challenge to modern multimodal reasoning, with the best model solv-
 482 ing only 14% of puzzles. Our error analysis reveals that current models exhibit myopic reason-
 483 ing, are bottlenecked by the limitations of language, and lack sketching capabilities. This makes
 484 PUZZLEWORLD uniquely well-suited for evaluating general-purpose reasoning systems under con-
 485 ditions that more closely resemble real-world open-ended scenarios, such as scientific discovery,
 exploratory data analysis, or investigative problem-solving.

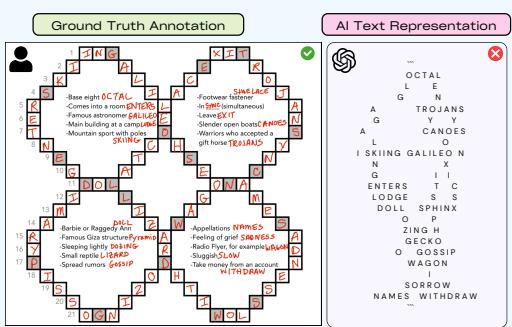


Figure 8: **Limitations of text.** An example failure case where GPT-o3 fails to represent a complex structured puzzle into text.

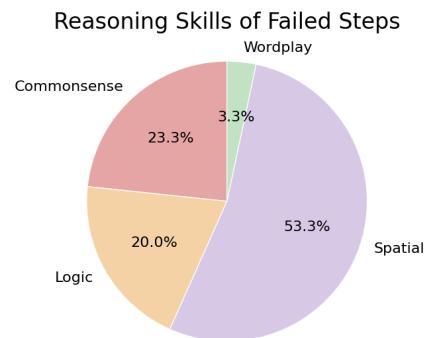


Figure 9: **Reasoning skills of failed steps.** We annotated the bottleneck steps with their reasoning skills.

486 ETHICS AND REPRODUCIBILITY STATEMENT
487

488 This research focuses on developing a benchmark to support the creation of models with robust open-
489 ended, multistep, multimodal reasoning. All data sources are cited and employed within the scope of
490 their intended use and applicable copyright licenses. To promote transparency and reproducibility,
491 we provide detailed data collection and annotation process in Section 4, evaluation setup in Section
492 5.1, and compute details in Section F of the appendix. We will publicly release the PUZZLEWORLD
493 benchmark and code to facilitate reproducibility and further research.

494 REFERENCES
495

496 Mit mystery hunt. <https://web.mit.edu/puzzle/www/>, 1981–.

497 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo
498 Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint*
499 *arXiv:2303.08774*, 2023.

500 Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. Large language models for
501 mathematical reasoning: Progresses and challenges, 2024. URL <https://arxiv.org/abs/2402.00157>.

502 Anthropic. Introducing claude 4. 2025. URL <https://www.anthropic.com/news/claude-4>.

503 Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang
504 Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent*
505 *systems and technology*, 15(3):1–45, 2024.

506 Guizhen Chen, Weiwen Xu, Hao Zhang, Hou Pong Chan, Chaoqun Liu, Lidong Bing, Deli Zhao, Anh Tuan
507 Luu, and Yu Rong. FineReason: Evaluating and improving LLMs’ deliberate reasoning through reflective
508 puzzle solving. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar
509 (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume*
510 *1: Long Papers)*, pp. 6685–6715, Vienna, Austria, July 2025a. Association for Computational Linguistics.
511 ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.333. URL [https://aclanthology.org/2025.acl-long.333/](https://aclanthology.org/2025.acl-long.333).

512 Steven-Shine Chen, Jimin Lee, and Paul Pu Liang. Interactive sketchpad: A multimodal tutoring system for
513 collaborative, visual problem-solving. *arXiv preprint arXiv:2503.16434*, 2025b.

514 Yuri Chervonyi, Trieu H Trinh, Miroslav Olšák, Xiaomeng Yang, Hoang Nguyen, Marcelo Menegali, June-
515 hyuk Jung, Vikas Verma, Quoc V Le, and Thang Luong. Gold-medalist performance in solving olympiad
516 geometry with alphageometry2. *arXiv preprint arXiv:2502.03544*, 2025.

517 Nathan A Chi, Teodor Malchev, Riley Kong, Ryan A Chi, Lucas Huang, Ethan A Chi, R Thomas McCoy, and
518 Dragomir Radev. Modeling: A novel dataset for testing linguistic reasoning in language models. *arXiv*
519 *preprint arXiv:2406.17038*, 2024.

520 Yew Ken Chia, Vernon Toh Yan Han, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. PuzzleVQA:
521 Diagnosing Multimodal Reasoning Challenges of Language Models with Abstract Visual Patterns, 2024.
522 URL <http://arxiv.org/abs/2403.13315>.

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

524 Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel
525 Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning,
526 multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*,
527 2025.

528 Antonia Creswell and Murray Shanahan. Faithful reasoning using large language models. *arXiv preprint*
529 *arXiv:2208.14271*, 2022.

530 Benjamin Estermann, Luca A. Lanzendorfer, Yannick Niedermayr, and Roger Wattenhofer. PUZZLES: A
531 Benchmark for Neural Algorithmic Reasoning. *Advances in Neural Information Processing Systems*, 37:
532 127059–127098, December 2024.

533 Deepanway Ghosal, Vernon Toh Yan Han, Chia Yew Ken, and Soujanya Poria. Are Language Models Puzzle
534 Prodigies? Algorithmic Puzzles Unveil Serious Challenges in Multimodal Reasoning, 2024. URL <http://arxiv.org/abs/2403.03864>.

540 Google. Gemini 3 Developer Guide. <https://ai.google.dev/gemini-api/docs/gemini-3>,
 541 2025.

542

543 Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie
 544 Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging
 545 benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems, 2024.

546

547 Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay
 548 Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *arXiv*
 549 preprint *arXiv:2406.09403*, 2024.

550

551 Shulin Huang, Shirong Ma, Yinghui Li, Mengzuo Huang, Wuhe Zou, Weidong Zhang, and Haitao Zheng. Late-
 552 val: An interactive llms evaluation benchmark with incomplete information from lateral thinking puzzles. In
 553 *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources*
 554 and *Evaluation (LREC-COLING 2024)*, pp. 10186–10197, 2024.

555

556 Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. A survey on large language models for
 557 code generation, 2024. URL <https://arxiv.org/abs/2406.00515>.

558

559 Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan.
 560 Swe-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference*
 561 on *Learning Representations*, 2024.

562

563 Long Hei Matthew Lam, Ramya Keerthy Thatikonda, and Ehsan Shareghi. A closer look at logical reasoning
 564 with llms: The choice of tool matters, 2024. URL <https://arxiv.org/abs/2406.00284>.

565

566 Heekyung Lee, Jiaxin Ge, Tsung-Han Wu, Minwoo Kang, Trevor Darrell, and David M Chan. Puzzled by
 567 puzzles: When vision-language models can't take a hint. *arXiv preprint arXiv:2505.23759*, 2025.

568

569 Hengzhi Li, Megan Tjandrasuwita, Yi R Fung, Armando Solar-Lezama, and Paul Pu Liang. Mimeqa: Towards
 570 socially-intelligent nonverbal foundation models. *arXiv preprint arXiv:2502.16671*, 2025.

571

572 Paul Pu Liang, Akshay Goindani, Talha Chafekar, Leena Mathur, Haofei Yu, Ruslan Salakhutdinov, and Louis-
 573 Philippe Morency. Hemm: Holistic evaluation of multimodal foundation models. In *The Thirty-eight Con-
 574 ference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024a.

575

576 Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learn-
 577 ing: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10):1–42, 2024b.

578

579 Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei
 580 Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation
 581 models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.

582

583 Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. Reasoning on graphs: Faithful and inter-
 584 pretable large language model reasoning. *arXiv preprint arXiv:2310.01061*, 2023.

585

586 Leena Mathur, Marian Qian, Paul Pu Liang, and Louis-Philippe Morency. Social genome: Grounded social
 587 reasoning abilities of multimodal models. *arXiv preprint arXiv:2502.15109*, 2025.

588

589 Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Fara-
 590 jtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models.
 591 In *The Thirteenth International Conference on Learning Representations*, 2025.

592

593 Philipp Mondorf and Barbara Plank. Beyond accuracy: Evaluating the reasoning behavior of large language
 594 models—a survey. *arXiv preprint arXiv:2404.01869*, 2024.

595

596 Arseny Moskvichev, Victor Vikram Odouard, and Melanie Mitchell. The ConceptARC Benchmark: Evaluating
 597 Understanding and Generalization in the ARC Domain. 2023. doi: 10.48550/ARXIV.2305.07141.

598

599 OpenAI. Openai o3 and o4-mini system card. <https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf>,
 600 2025. Accessed: 2025-05-16.

601

602 Puzzled Pint. Puzzled pint. <https://puzzledpint.org/>, 2025. CC BY-NC-SA Intl. 4.0.

603

604 Qwen. Qvq: To see the world with wisdom, December 2024. URL <https://qwenlm.github.io/blog/qvq-72b-preview/>.

594 Shashi Pal Singh, Ajai Kumar, Hemant Darbari, Lenali Singh, Anshika Rastogi, and Shikha Jain. Machine
 595 translation using deep learning: An overview. In *2017 international conference on computer, communications
 596 and electronics (comptelix)*, pp. 162–167. IEEE, 2017.

597 Marco Antonio Calijorne Soares and Fernando Silva Parreiras. A literature review on question answering
 598 techniques, paradigms and systems. *Journal of King Saud University-Computer and Information Sciences*,
 599 32(6):635–646, 2020.

600 Winnie Street, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael McKibben, Tatenda
 601 Kanyere, Alison Lentz, Blaise Aguera y Arcas, and Robin I. M. Dunbar. Llms achieve adult human perfor-
 602 mance on higher-order theory of mind tasks, 2024. URL <https://arxiv.org/abs/2405.18870>.

603 Hai-Long Sun, Zhun Sun, Houwen Peng, and Han-Jia Ye. Mitigating visual forgetting via take-along visual
 604 conditioning for multi-modal long CoT reasoning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova,
 605 and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Com-
 606 putational Linguistics (Volume 1: Long Papers)*, pp. 5158–5171, Vienna, Austria, July 2025. Association
 607 for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.257. URL
 608 <https://aclanthology.org/2025.acl-long.257/>.

609 Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. A benchmark for learning
 610 to translate a new language from one grammar book. *arXiv preprint arXiv:2309.16575*, 2023.

611 Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang,
 612 Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

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

616 Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth
 617 a thousand words? delving into spatial reasoning for vision language models. In *The Thirty-Eighth Annual
 618 Conference on Neural Information Processing Systems*, 2024a.

619 Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba,
 620 Shichang Zhang, Yizhou Sun, and Wei Wang. SciBench: Evaluating College-Level Scientific Problem-
 621 Solving Abilities of Large Language Models. In *Proceedings of the Forty-First International Conference on
 622 Machine Learning*, 2024b.

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

626 Adhika Pramita Widjassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy
 627 Affandy, and De Rosal Ignatius Moses Setiadi. Review of automatic text summarization techniques &
 628 methods. *Journal of King Saud University-Computer and Information Sciences*, 34(4):1029–1046, 2022.

629 Tongshuang Wu, Michael Terry, and Carrie Jun Cai. Ai chains: Transparent and controllable human-ai inter-
 630 action by chaining large language model prompts. In *Proceedings of the 2022 CHI conference on human
 631 factors in computing systems*, pp. 1–22, 2022.

632 Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Mind’s eye of llms:
 633 Visualization-of-thought elicits spatial reasoning in large language models. In *The Thirty-eighth Annual
 634 Conference on Neural Information Processing Systems*, 2024.

635 xAI. Grok-4. <https://x.ai/news/grok-4>, 2025. Accessed: 2025-09-23.

636 Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of
 637 thoughts: Deliberate problem solving with large language models. *Advances in neural information process-
 638 ing systems*, 36:11809–11822, 2023.

639 Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin,
 640 Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali
 641 Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. Mmt-bench: A comprehensive multimodal
 642 benchmark for evaluating large vision-language models towards multitask agi, 2024.

643 Eunice Yiu, Maan Qraitem, Anisa Noor Majhi, Charlie Wong, Yutong Bai, Shiry Ginosar, Alison Gopnik,
 644 and Kate Saenko. Kiva: Kid-inspired visual analogies for testing large multimodal models. *arXiv preprint
 645 arXiv:2407.17773*, 2024.

648 Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang,
 649 Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng,
 650 Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. Mmmu: A massive multi-
 651 discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*,
 652 2024a.

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

657 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuo-
 658 han Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in*
 659 *Neural Information Processing Systems*, 36:46595–46623, 2023.

660 Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyuan Luo, Zhangchi Feng, and Yongqiang Ma.
 661 Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual*
 662 *Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok,
 663 Thailand, 2024. Association for Computational Linguistics. URL <http://arxiv.org/abs/2403.13372>.

665 Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie
 666 Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal
 667 models. *arXiv preprint arXiv:2504.10479*, 2025.

668
 669
 670
 671
 672
 673
 674
 675
 676
 677
 678
 679
 680
 681
 682
 683
 684
 685
 686
 687
 688
 689
 690
 691
 692
 693
 694
 695
 696
 697
 698
 699
 700
 701

702 **A LIMITATIONS AND BROADER IMPACT**
703

704 To ensure consistency and standardization across the dataset, we excluded puzzles involving under-
 705 explored or difficult-to-represent modalities such as audio, video, or interactive file-based inputs. As
 706 a result, PUZZLEWORLD may not fully capture the breadth of sensory and interaction-based reasoning
 707 found in some real-world, more challenging puzzlehunts. **We discuss the potential incorporation**
 708 **of more challenging puzzles in Appendix C.** Additionally, unlike Wang et al. (2025) that uses human
 709 annotators to transcribe textual and visual components separately, we preserve the puzzle content in
 710 its original image format and focus annotation efforts on intermediate reasoning traces. While this
 711 allows PUZZLEWORLD to provide richer annotation of the solution reasoning process, it may also
 712 introduce variability in model performance depending on the quality of their OCR capabilities. **We**
 713 **discuss whether OCR is a limiting capability for frontier models in Appendix D.2.**

714 Finally, our evaluation pipeline relies on LLM-based judges to automatically assess generated reasoning
 715 traces. To address this, we adopted careful prompting and cross-checking. For example, we
 716 enforce that each individual annotated step is evaluated separately by the LLM judge to determine
 717 whether it matches the model generated solution. We tested the alternative approach, where the
 718 LLM judge is prompted with the full candidate solution and outputs the latest ground truth step that
 719 the candidate response achieved. However this approach was more prone to hallucinations, as LLM
 720 judge sometimes outputs a stepwise accuracy greater than 1. As such, our approach of running the
 721 LLM judge to output a boolean on each ground truth step helps mitigate potential hallucination.
 722 Nevertheless, we acknowledge that the use of LLM-based evaluations may be subject to instability
 723 or bias, and the metrics should be taken with caution.

724 One possible concern with our grading scheme is that a model might "hallucinate" the correct final
 725 answer without engaging in proper reasoning. However, such a case is extremely rare in PUZ-
 726 ZLEWORLD. The high-quality, human-designed puzzles are deliberately constructed to discourage
 727 superficial guessing, and even experienced human solvers cannot easily infer the answer without
 728 following the intended logic. In a thorough manual inspection of models' puzzle responses, we did
 729 not find any case where the model arrived at the final answer without demonstrating the necessary
 730 reasoning. **In Appendix D.3, we discuss whether our stepwise scoring mechanisms successfully**
 731 **credit alternative solution paths.**

732 Our goal in releasing PUZZLEWORLD is to advance research in general-purpose, multimodal rea-
 733 soning systems. However, we recognize that increasingly capable AI models, especially those
 734 skilled at complex reasoning, carry risks of misuse. These include the potential for externalizing
 735 or replacing human reasoning in settings where authenticity or creativity is essential, such as ed-
 736 ucation, scientific authorship, or collaborative problem-solving. While our dataset does not pose
 737 direct risks on its own, we support future work that includes safeguards to mitigate misuse and
 738 encourages the responsible deployment of reasoning-capable AI systems in alignment with human
 739 values.

740 **B USE OF LARGE LANGUAGE MODELS**
741

742 This project used Large Language Models (LLMs) to assist dataset verification and model evalua-
 743 tions. Details are described in the Sections 4.3 and 5.1. We also acknowledge the use of LLMs
 744 to assist with correcting grammatical errors and improving clarity of the writing. This assistance
 745 was limited to language refinement and did not affect the core methodology, scientific rigour, or
 746 originality of the research. We confirm that no AI-generated content has been presented as our own
 747 intellectual contribution.

748 **C FUTURE DIRECTIONS**
749

750 We note that some prior benchmarks (Chen et al., 2025a) adopt a *structured reflective reasoning*
 751 paradigm, centered on structured, text-based puzzles with explicit rules, enabling precise state verifi-
 752 cation and surgical analysis of multi-step logical reasoning. In contrast, PUZZLEWORLDemphasizes
 753 multimodal, discovery-driven reasoning in which the puzzle rules are implicit and must be inferred
 754 by the solver. This setting offers a more holistic evaluation of generalist reasoning capabilities that
 755 requires a broader range of cognitive skills, though at the cost of less precise, LLM-approximated
 intermediate state verification due to the absence of explicit rule structures.

756 An exciting direction for future work is to extend PUZZLEWORLD with structured, verifiable
 757 reflective-reasoning frameworks akin to those in Chen et al. (2025a). However, we caution that
 758 PUZZLEWORLD’s open-ended and multimodal puzzles often involve diverse, creative reasoning
 759 steps that are difficult to formalize as explicit rules, making it challenging to achieve the same level
 760 of rule-based granularity as in more structured puzzle domains in Chen et al. (2025a). That said,
 761 integrating reflective reasoning paradigms with PUZZLEWORLD could yield richer diagnostic tools
 762 and may ultimately benefit the training and evaluation of frontier reasoning models.

763 We acknowledge that, beyond PuzzledPint, extensive puzzlehunt-style puzzle sources could be inte-
 764 grated to expand PUZZLEWORLD, such as MIT Mystery Hunt (MIT, 1981–). While MIT Mystery
 765 Hunt is an attractive source of challenging, high-quality puzzles, we note several considerations that
 766 make it challenging to integrate it to PUZZLEWORLD’s initial release.

- 768 • Unlike PuzzledPint (released under Creative Commons), MIT Mystery Hunt puzzles do not
 769 have a centralized copyright policy. Permissions would need to be obtained from individual
 770 authors, which adds substantial overhead.
- 771 • Mystery Hunt puzzles frequently involve audio, video, interactive web tools, custom code,
 772 and even physical components. These are exciting modalities for future generalist AI sys-
 773 tems, but they are difficult to standardize under PUZZLEWORLD’s stepwise annotation
 774 format and are not well supported by current frontier models.
- 775 • Mystery Hunt puzzles are intentionally designed to challenge top teams and often require
 776 multiple people to solve each puzzle. Given that frontier models already struggle on the
 777 entry-level PuzzledPint puzzles, including much harder Mystery Hunt puzzles would likely
 778 provide limited practical diagnostic value at this stage.

780 Therefore, for the present, we prioritize the more accessible PuzzledPint puzzles, as they are both
 781 suitably challenging and sufficiently structured to support reliable stepwise annotation and informa-
 782 tive intermediate diagnostics. That said, as models and annotation tools improve, future work that
 783 extends PUZZLEWORLD to incorporate more advanced modalities and higher-difficulty sources –
 784 potentially including Mystery Hunt puzzles – would offer a richer evaluation suite.

785 D ADDITIONAL DISCUSSION

786 D.1 ON THE ABSTRACT NATURE OF PUZZLEHUNTS

787 We note that PUZZLEWORLD, like prior puzzle datasets (Chia et al., 2024; Ghosal et al., 2024;
 788 Estermann et al., 2024), is abstract in nature rather than drawn from real-world tasks such as math-
 789 ematics or physics. While this means that PUZZLEWORLD does not directly evaluate a model’s
 790 ability on any specific application, this abstraction is intentionally designed to assess generalist,
 791 open-ended reasoning while reducing pattern memorization. For example, prior work has shown
 792 that symbolic variants of math datasets can reveal brittle reasoning behavior in models that oth-
 793 erwise appear strong (Mirzadeh et al., 2025). PUZZLEWORLD’s abstraction thus helps isolate and
 794 measure a model’s generalist and genuine reasoning ability beyond domain-specific patterns.

795 Existing reasoning benchmarks – ranging from math and coding tasks – typically operate in closed-
 796 ended, well-defined environments. These datasets assess correctness under fixed rules, but they do
 797 not require models to navigate underspecified or open-ended problem spaces. In contrast, many
 798 real-world tasks (e.g., exploratory data analysis, investigative research, science discovery) involve
 799 ambiguous signals, multiple possible solutions, and iterative hypothesis testing.

800 PUZZLEWORLD fills this gap in two ways that previous abstract reasoning benchmarks do not. First,
 801 unlike traditional puzzle datasets (e.g. Sudoku) that evaluate within a limited rule set, our puzzle-
 802 hunt puzzles require leveraging diverse reasoning competencies to dynamically form and evaluate
 803 hypotheses from nuanced multimodal signals. Second, our manually annotated reasoning traces
 804 enable systematic analysis of model behavior in these open-ended settings, revealing characteristic
 805 failure modes of frontier models such as “myopic commitment” (Section 5.4). Our transfer experi-
 806 ments in Table 4 has also shown that capabilities exercised in PUZZLEWORLD transfer to real-world
 807 benchmarks, suggesting that puzzlehunt-style reasoning is a useful proxy for generalist reasoning.
 808 Nonetheless, we acknowledge that no single benchmark can represent all forms of reasoning.

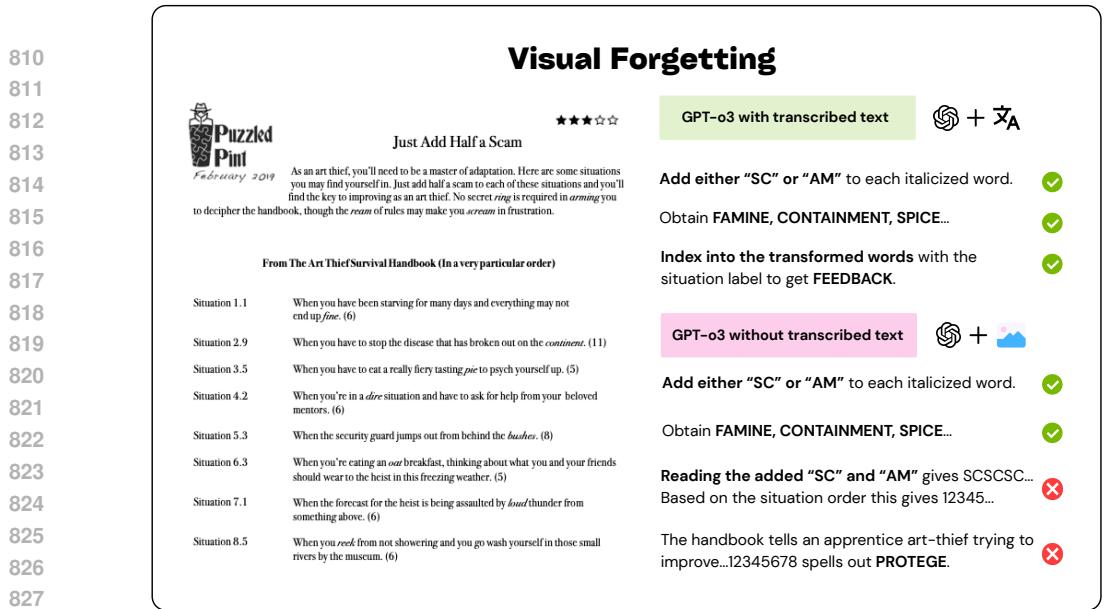


Figure 10: **Example error of visual forgetting.** GPT-o3 failed to correctly reference the text clues in the puzzle image, but it solved the puzzle once the clues were provided in transcribed text form.

D.2 ARE REASONING MODELS BOTTLENECKED BY TEXT RECOGNITION?

We acknowledge that multimodal large language models may be bottlenecked by text recognition in images. In the construction of PUZZLEWORLD, since Wang et al. (2025) has shown that today’s best multimodal reasoning models are not primarily constrained by optical character recognition (OCR) capabilities, we devoted our manual effort to construct fine-grained annotations. Although each puzzle includes a human-transcribed “flavor text”, this may not capture all textual content.

We thus provide an analysis on whether OCR is limiting the performance of frontier models on PUZZLEWORLD. We selected a 5% subset of strictly text-only puzzles, manually transcribed all textual content, and compared GPT-o3’s performance with and without the full transcription. A paired t-test showed no significant difference ($t = 0.52, p > 0.1$), suggesting that text recognition limitations are not a major bottleneck for the frontier multimodal models.

That said, we did observe isolated cases where GPT-o3 improved from partial to full solutions once transcription was provided. Upon examination, these instances appear to stem from model hallucination in later stages of reasoning when relying only on images, whereas the transcription helps anchor its reasoning more reliably. For example, in Figure 10, GPT-o3 solves the puzzle correctly when given the transcribed text. When relying solely on the image, GPT-o3 can solve the intermediate text clues but fails during extraction: it misreferences the visual text, as shown by its inability to recover the added letters against the italicized words in the image, and hallucinates a final answer only tangentially related to the flavor text. This aligns with patterns of visual forgetting observed in prior works (Sun et al., 2025).

D.3 CAN PUZZLEWORLD’S STEPWISE SCORING CREDIT CREATIVE SOLUTIONS?

Due to the PUZZLEWORLD’s open-ended nature, a legitimate concern is whether PUZZLEWORLD’s stepwise scoring mechanism can fairly credit creative, alternative reasoning paths, rather than being limited to the annotated reference chain. In puzzlehunts, the solution steps are intentionally designed to be interlocking and sequentially dependent. It is thus hard to reach the correct answer while following a completely different solution path, and correct partial chains must partially converge with the annotated ground-truth, even if the solver temporarily deviates or makes logic leaps.

That said, human puzzlehunters do commonly make educated guesses that skip over intermediate steps. For example, after identifying three of the four final answer letters (“I _ L R”), a solver might correctly infer “ICLR” without fully completing the preceding clue. PUZZLEWORLD’s step-level scoring explicitly accounts for this behavior: we identify the latest ground-truth step that appears in the candidate solution. In this scenario, because the solver correctly arrives at the final answer, our scoring grants full credit, even if several intermediate steps were omitted or approximated.

864 Table 5: **GPT-o3’s performance on PUZZLEWORLD by difficulty.** GPT-o3’s accuracy monoton-
 865 ically decreases as difficulty increases.

	Difficulty	Accuracy	Stepwise
866	Easy	18.46%	40.58%
867	Medium	14.76%	39.38%
868	Hard	9.64%	39.71%
869			
870			
871			

872 Table 6: **GPT-o3’s accuracy on PUZZLEWORLD by difficulty and modality.** Within each modal-
 873 ity, GPT-o3’s accuracy consistency decreases with increased difficulty, with a relatively well per-
 874 formance on easy structured puzzles.

	Difficulty	Text	Visual	Structured
875	Easy	19.5%	11.9%	20.4%
876	Medium	15.3%	8.9%	13.5%
877	Hard	10.6%	7.6%	9.6%
878				
879				
880				

881 As a proxy for evaluating how robust our step-level crediting is to alternative partial chains, we
 882 randomly sampled 10% of PUZZLEWORLD and automatically paraphrased GPT-o3’s solutions us-
 883 ing GPT-5-mini. We then compared the LLM-judge’s step-level scores on the original and para-
 884 phrased solutions. We observe strong positive correlation between the two sets of scores ($r = 0.885$,
 885 $p < 0.001$) with a mean absolute error of 0.066, indicating a high degree of score stability across
 886 semantically equivalent but structurally varied solutions. Although this does not exhaustively cover
 887 all possible reasoning paths, it indicates that the LLM-judge can reliably credit alternative solutions.

E PUZZLEWORLD DETAILS

E.1 CHECKING FOR DATASET CONTAMINATION

893 To assess the possibility of data contamination, we test whether GPT-o3 (OpenAI, 2025) has memo-
 894 rized any of the puzzles in our dataset. Specifically, inspired by prior work Tanzer et al. (2023); Chi
 895 et al. (2024), we prompt the model to reconstruct the flavor text for 40 randomly sampled puzzles out
 896 of the 84 that were answered correctly. We then use GPT-4o (Achiam et al., 2023) to automatically
 897 evaluate the similarity between the reconstructed and original flavor texts. We find a reconstruction
 898 accuracy of 0%, suggesting little to no evidence of data leakage. Furthermore, since Puzzled Pint
 899 (Puzzled Pint, 2025) publishes new puzzles on a monthly basis, our dataset can be continuously
 900 updated to mitigate the risk of model overfitting on released content.

E.2 DIFFICULTY LABEL ANALYSIS

902 We acknowledge that the difficulty labels in PUZZLEWORLD are obtained from the original Puz-
 903 zledPint sources, without manual calibration. According to the PuzzledPint website (Puzzled Pint,
 904 2025), all submitted puzzles undergo internal playtesting by an editorial team, where the difficulty
 905 tags are revised and finalized by expert editors and reflect a community-standard difficulty assess-
 906 ment. We thus treat them as expert annotations rather than ad-hoc metadata.

907 To understand whether the human-labelled difficulty labels meaningfully reflect task difficulty, we
 908 inspect GPT-o3’s performance broken down by difficulty in Table 5. As expected, we observe that
 909 GPT-o3’s accuracy decreases as difficulty level increases. Another valid concern is that human-
 910 perceived difficulty and AI-perceived difficulty might not align. For example, certain diagram-
 911 heavy puzzles may be straightforward for humans but require nontrivial spatial reasoning for AIs.
 912 To investigate this, we further break down GPT-o3’s performance by modality and difficulty.

913 As observed in Table 6, within each modality, accuracy consistently decreases with difficulty. No-
 914 tably, GPT-o3 performs relatively well on easy structured puzzles – likely due to their clean, regular
 915 layouts – but its performance drops sharply on harder structured puzzles where the diagrams become
 916 more irregular and visually complex. Overall, both aggregated and modality-level results show that
 917 PuzzledPint’s difficulty labels provide a meaningful and consistent difficulty metric for AI models.

918
919

E.3 PUZZLEWORLD IMAGE SAMPLES

920

We provide high-resolution images of puzzle samples used in this paper.

921

922

923

924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

946

947

948

949

950

951

952

953

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971



MAKI



Location recap: Stephanie mentioned to her roommate Katie that she didn't know there were multiple types of sushi. Upon hearing this, Katie decided to teach Stephanie about the many different forms that sushi can take!

At a sushi restaurant, Katie continued Stephanie's crash course in sushi.

"We'll start with the basics," Katie said. "You mentioned you have only had California rolls before. Those fall into the category of *maki*, or sushi rolls. There's seaweed and rice on the outside of the maki sushi, and the filling is in the center. To make maki, chefs lay out all the components on top of each other so that they **overlap** to an extent, and then tightly roll them into a circle."

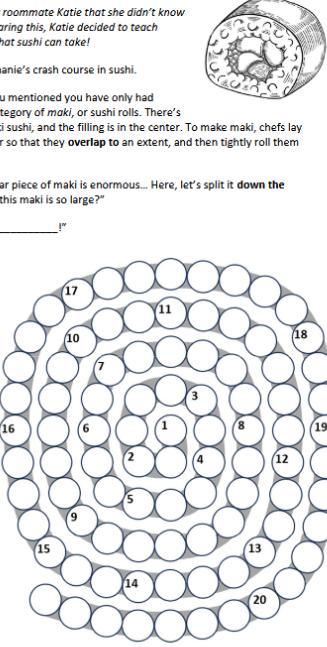
"Wow, I see!" Stephanie said. "This particular piece of maki is enormous... Here, let's split it **down the middle** and share it. Say, do you know why this maki is so large?"

"I'm not sure," Katie replied. "But we'll _____!"

Clues:

1. Former bride
2. Sworn allegiance
3. Descriptor for blood
4. Pluckable flower part
5. A Greek letter
6. A feeling more intense than distress
7. Psychic communicator, e.g., Professor X
8. Considerate
9. "Pen" follower
10. One of three from a famous trio
11. Group of musicians led by a conductor
12. Poe bird
13. E.g., sun-grid, combustion, or Unity
14. E.g., optic or tibial
15. In geometry, a point where two or more lines meet
16. For _____ (as an illustration)
17. Folkloric Irish creature
18. Rally around a common cause
19. Maryland's state reptile
20. Shorter alternative to a signature

© 2024 CC BY-NC-SA Intl. 4.0
Stephanie Yang (Boston, MA)



QUESABIRRIA



More akin to quesadillas than tacos, quesabirria tacos feature stewed, shredded beef, goat, or jackfruit melted together with cheese inside of a tortilla. The tortilla is folded with two **nearly identical halves** pressed together. These tacos are often served with **a** broth or consomme for dipping each taco. In fact, you'd be hard-pressed to find a more important part of the quesabirria experience than the long-simmered, **well-reduced** broth.

Assertive and reckless (5)

Assistance (3)

Attempt to overfill (4)

Bountiful celebratory meal (5)

Classification of vinegar or citrus juice (4)

Currency in London (5)

Euphemism for ocean (4)

Ground covering that is greener where you aren't (5)

High fat dairy liquid (5)

Kind and polite (4)

Launderer's woe (5)

Man who wrote about a raven (3)

Noted Vatican resident (4)

Okie doke (3)

Quick (4)

Russian dictator (6)

Second word of a celebration in New Orleans (4)

Sibling's daughter (5)

Site for restaurant reviews (4)

Support garments (4)

972

973

974

975

976

977

978

979

980

981

982

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

1001

1002

1003

1004

1005

1006

1007

1008

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

1020

1021

1022

1023

1024

1025



Archimedes and Benjamin Banneker are sitting on a bench, having a discussion.
 Archie: Tell me old friend, when you're drawing angles, don't you just feel amazing – like your mind is the willing subject of a kidnapping to another dimension?
 Benjie: Wow, Archie, I never thought of it that way. I just know they always seem to make fizzy feelings inside of me. You could say that I am one hopeless lover when it comes to angles!
 Archie: You might not realize it at first glance, but if you look really closely, you'll realize it always takes three points to define an angle, and I've always liked threes: Three musketeers, three-legged races, three-syllable words.
 Benjie: I just feel angles make so many awesome connections: they are what all the other shapes we draw depend on.
 Archie: I know, from acute to obtuse, angles have always seemed larger than life!
 Carl Friedrich Gauss walks up and tries to join the discussion.
 Carl: Hey I like drawing angles too.
 Archie: Um, excuse you! This was an A conversation, so you can C your way out!



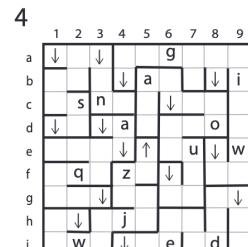
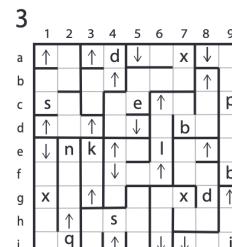
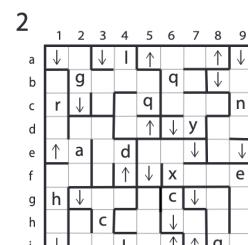
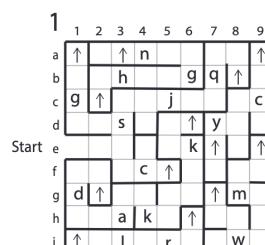
© 2024 CC BY-NC-SA Intl. 4.0 Robert Becker (St. Louis, MO)



Mittens



Mittens has been forcefully relocated by his hoomans to a new mansion. He now faces the daunting task of exploring his new home. To further complicate matters, the mansion has four levels, and his perspective will have to keep shifting with each new level. Along the way, Mittens will discover the shortest path that leads him to his true heart's desire, for which he will have to make his way all the way up onto the roof.

---
©2022 CC BY-NC-SA Intl. 4.0 Ben Zoon (Boise, ID)

1026

1027

1028

1029

1030

1031

1032

1033

1034

1035

1036

1037

1038

1039

1040

1041

1042

1043

1044

1045

1046

1047

1048

1049

1050

1051

1052

1053

1054

1055

1056

1057

1058

1059

1060

1061

1062

1063

1064

1065

1066

1067

1068

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079



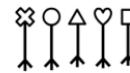
Eros, the god of love, has shown up at a speed-dating session. Each person at this speed-dating session is assigned a **number**. Eros used his arrows to make each person fall in love with someone on the other side. Eros is mischievous: he caused everyone to fall in love with one person, everyone, to be loved by one person, but no one's love to be reciprocated.

Eros has many arrows, but only five different arrowheads; each arrowhead does something different. Additionally, Eros may shoot up to three arrows **from** each person to get the job done, even repeating a type if required.

Can you figure out who fell in love with whom?

Possible arrowhead effects:

- Add the number of letters in the person's name
- Divide by 3
- Multiply by 2
- Subtract 2
- Subtract 5



1. Blake	$\Delta \bowtie$	•	X	A	C	• 6. Leslie	$\square \bowtie \circ$		
2. Charlie	$\bowtie \circ \Delta$	•	E	B	D	F	• 7. Jesse	\circ	
3. Avery	$\Delta \Delta \heartsuit$	•	G	H	Z	T	• 8. Emerson	$\heartsuit \heartsuit$	
4. Kerry	$\Delta \circ \Delta$	•	J	K	W	O	N	• 9. Jamie	$\square \heartsuit$
5. Parker	$\bowtie \heartsuit$	•	P	U	Q	Y	S	• 10. Alex	\circ
			V	R	I				

(C) 2021 CC BY-NC-SA Int. 4.0 by Stephanie Yang (Rockville, MD)



© 2020 CC BY-NC-SA Int. 4.0 Jimmy Chang and Chelsea Cullen

(Los Angeles, CA USA)



Add another critie to the Christmas croak's wrongs:
the hoot's gone and roared all of Whoo-ville's songs!
These crities are off-beat, off-beat and off-key.
It's a critie and off-beat, off-beat and off-key.
Mark through the tones to hear the sounds underneath.
And burn the lines of note taken by the thief!

Legend:
 $\frac{1}{2}$ = 1 count rest
 $\frac{1}{4}$ = whole measure rest
 $\frac{1}{8}$ = 1 count note
 $\frac{1}{16}$ = 2 count note

1080

1081

1082

1083

1084

1085

1086

1087

1088

1089

1090

1091

1092

1093

1094

1095

1096

1097

1098

1099

1100

1101

1102

1103

1104

1105

1106

1107

1108

1109

1110

1111

1112

1113

1114

1115

1116

1117

1118

1119

1120

1121

1122

1123

1124

1125

1126

1127

1128

1129

1130

1131

1132

1133



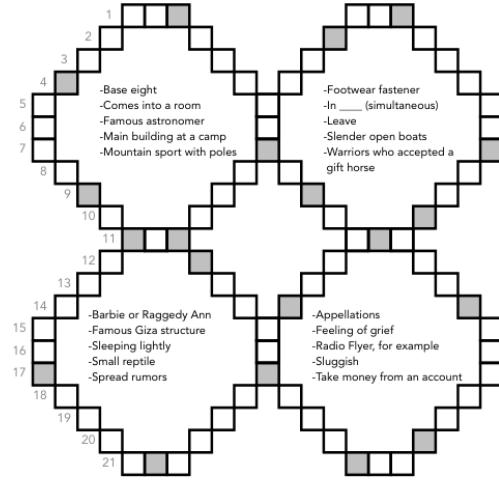
MISSING LINKS



You turn to another page in Nicolas' journal and begin to read.

Every magician needs a good pseudonym. After coming up with several tricks of my own, I thought taking the name of an inventor I admired would be fitting. Here's a trick I designed to honor him that is based on the famous linking rings act. The gimmick this time is that instead of metal I've linked words into four rings, and even managed to entwine those.

Each highlighted link is the start and end of a word.



© 2020 CC BY-NC-SA Int'l. 4.0 by James Mahoney (Columbia, MO, USA)

Can you figure out who my stage-names was?

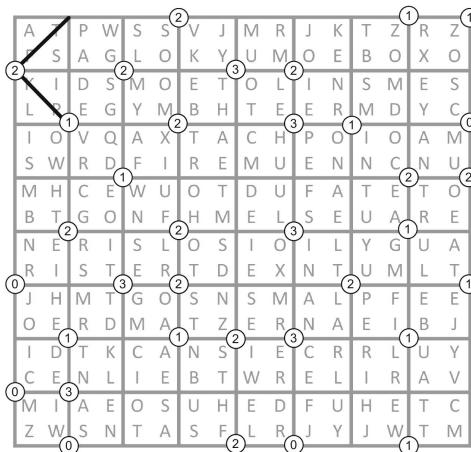
4.4 18.2 1.5 1.2 5.1 4.2 9.4 12.3 19.1
13.2 17.3 7.1 10.3 15.1 19.4 13.4 3.3 9.2 20.3 21.5 21.3 20.1



The Venue in Verona ★★★★★

The Capulets and Montagues have each claimed half of each city block as their territory. They have divided each with a wall, corner to corner. This makes getting to the wedding venue like traversing a maze. Help our couple find a path that crosses the city, from one side to the other, to find out who awaits them there.

- In **every square**, draw a diagonal line between two of its opposing corners, in one of the two possible directions, as shown in the upper-left.
- Each numbered circle shows exactly how many lines connect to that point.
- Unnumbered points can have any number of lines meeting at them.
- No area can be completely walled off. The entire grid must be reachable from at least one of the sides (i.e. do not create any complete squares or rectangles with your lines).
- Paths will form between the lines you draw, making a maze.



Puzzle by Neal Tibrewala (Austin, TX)

1134

F BENCHMARKING DETAILS

1135

1136

F.1 COMPUTE RESOURCES

1137

1138

1139

All evaluations and experiments in this paper were conducted on a remote cluster equipped with two NVIDIA H200 GPUs (each with 141 GB HBM3 memory). Runtime for each model evaluation varied between 10–48 hours depending on the model size and architecture.

1140

1141

F.2 SUPERVISED FINETUNING

1142

1143

1144

1145

1146

1147

All fine-tuning experiments were conducted using the LLaMA Factory framework (Zheng et al., 2024). For fine-tuning on the 8B InternVL3 model, we used LoRA fine-tuning with a rank of 8, a learning rate of 1×10^{-6} , and trained for 3 epochs on PUZZLEWORLD. For the transfer experiments, we fine tune a Qwen2.5-VL-7B-Instruct model, using full parameter finetuning with a learning rate of 1×10^{-5} for 5 epochs on PUZZLEWORLD. We did not perform extensive hyperparameter tuning for any of these experiments.

1148

F.3 HUMAN BASELINE DETAILS

1149

1150

1151

1152

1153

We provide additional context for the human baseline in Table 2. While the human baseline only covers a subset of PUZZLEWORLD, the resulting baselines reveal clear separations between human tiers and current models—for example, a >30% gap between human enthusiasts and GPT-o3. This gap is substantially larger than the uncertainty introduced by the sample size, suggesting that the current human results provide a reliable directional reference point.

1154

1155

1156

1157

1158

1159

Given our resource constraints, we prioritized depth and quality in the human evaluation rather than exhaustiveness. To reduce individual variability in a task as open-ended as PuzzleWorld, each sampled puzzle was solved by at least two independent participants, each allotted up to one hour and asked to produce detailed reasoning traces. Since we assign multiple participants to each puzzle to ensure consistency, each puzzle modality has more than 40 human attempts, which we consider sufficient for a modality-level human baseline.

1160

1161

1162

1163

Finally, we note that the same LLM-judge protocol is applied to both human and model solutions, ensuring that the comparison is fair and step-symmetric. We observe that, unlike AI models, visual and structured modalities do not pose additional difficulty for human solvers, further highlighting the multimodal limitations of current foundation models.

1164

1165

F.4 ROBUSTNESS OF LLM-JUDGE TO IMPERFECT ANNOTATIONS

1166

1167

1168

1169

1170

1171

1172

1173

PUZZLEWORLD contains extensive manual annotation to provide fine-grained analysis on models' performance. In this section, we include an ablation on the impact of potential human annotation noise on our step-level grading accuracy. To simulate annotation noise, we select 10% of PUZZLEWORLD and use GPT-5-mini to paraphrase each step of the ground truth annotation. Then, we evaluate GPT-o3's output on both the original and paraphrased ground truth. We observe strong positive correlation between the two sets of scores ($r = 0.898, p < 0.001$) with a mean absolute error of 0.056. This indicates that PUZZLEWORLD's stepwise evaluation pipeline is robust to moderate noise in the stepwise annotations.

1174

F.5 PROMPT FOR BENCHMARKING

1175

1176

Below is the system prompt template for benchmarking models on PUZZLEWORLD, which is adapted from Wang et al. (2025).

1177

1178

1179

1180

1181

1182

1183

1184

1185

1186

1187

You will be presented with a puzzle to solve. The puzzle may not have specific instructions, but you know that the answer to the puzzle is a word or short phrase (or rarely, a number).

Do not ask any questions about how to proceed, just do your best to solve the puzzle. Here are some tips for solving puzzles of this type:

General Tips:

- Puzzles will often have multiple steps to get to the answer word. You can usually tell you are on the right track if the intermediate answers agree with the title, flavor, or theme of the puzzle.
- You can usually find hints in the introductory text. For example references to "in the dark" or "sight" are often hints something is encoded with braille.
- Puzzles often incorporate acrostics: a clue where the first letter, syllable, or word of each line, paragraph, or other recurring feature spells out a word or message.
- If you end up with a garbled "alphabet soup", then look for a clue on how to order them.
- Indexing is one of the most common puzzle mechanisms. Try indexing when you have a list of words or phrases and a corresponding list of numbers. Count into the word or phrase by the given number and record the letter in that position. For example: "2 Cake, 6 Pudding, 5 Shortening" gives you "ant".

1188 - Alpha-numeric codes are also very common. If you end up with a list of numbers try replacing
 1189 the numbers with the corresponding letters like this: 1 = A, 2 = B, 3 = C... 26 = Z.
 1190 Occasionally, these types of codes will "wrap around", so don't despair if you see a
 1191 number greater than 26. Just subtract 26 and try again. In this scenario 27 (27-26 = 1) =
 1192 A, 28 (28-26 = 2) = B etc. If you try this and it doesn't work, try other numeric codes
 1193 such as ASCII.
 1194 - Often a puzzle repeats a strategy multiple times.
 1195 You will likely need to backtrack frequently, so make sure to write out your steps as you go.
 1196 If you get stuck, try to think of a new way to approach the puzzle. Try:
 1197 - Rereading the title and the flavor text. These are the most important hints about what type
 1198 of strategies, themes or cultural references might be used to solve the puzzle.
 1199 - Checking for pop culture references
 1200 - Checking for references to a song/poem/book/movie/TV show
 1201 For strings, examples of strategies you might try include:
 1202 - Alphabetizing
 1203 - Using leftover letters to spell something
 1204 - Rearranging the letters (aka anagrams or "transposing")
 1205 - Seeing if there are any acronyms
 1206 - Diagonalizing (taking the first letter of the first answer, the second letter of the second
 1207 answer, etc.)
 1208 - Looking for unusual letter frequencies
 1209 - Puns and homophones
 1210 - Shifting from letters to numbers
 1211 For numbers, try:
 1212 - Shifting from numbers to letters
 1213 - Using it as a phone number
 1214 - Treating numbers as dates
 1215 - Treating numbers as ASCII numbers
 1216 - Seeing if there are any strange sequences
 1217 - Seeing if prime numbers are involved
 1218 For images, try:
 1219 - Looking at it in a mirror
 1220 - Squinting at it from far away
 1221 - Tilting it
 1222 - Looking at it upside down
 1223 - Looking through it
 1224 - Transcribing it neatly

1213 We additionally append the user prompt:

1214 Your task is to solve the following puzzle. The attached images are presented in the order
 1215 they are referenced in the text.
 1216
 1217 The puzzle's title is: {}
 1218 The puzzle's flavor text is: {}
 1219 ---
 1220 Write out a step-by-step solution to the puzzle. At the end of your solution, write your
 1221 answer in the following format:
 1222 Answer: <answer>

1222 Below is the prompt for LLM judge:

1223 Answer Equivalence Instructions:
 1224 Using the puzzle and the reference solution, grade the candidate solution as follows.
 1225 For every reasoning step of the reference solution, output True if the candidate solution both
 1226 includes
 1227 the step and achieves the same intermediate result of the step, otherwise False.
 1228 Explain why the candidate's solution did or did not get the reasoning step correct.
 1229 Do not add more steps than there are in the reference solution and evaluate every step
 1230 in the reference solution.
 1231 There is a exception in scoring for the last reasoning step. Identify the candidate output
 1232 solution.
 1233 If the candidate output solution is the exact same as the reference solution answer of
 1234 \'{puzzle_solution}\',
 1235 then output final step as true.

1233 G ANNOTATOR DETAILS

1234 We employ university undergraduates to assist the human annotation process in PUZZLEWORLD.
 1235 All annotators are compensated at a rate of \$16.00 per hour. Prior to annotation, annotators receive
 1236 detailed guidelines and participate in training sessions to ensure annotation consistency and task
 1237 understanding.

1238 G.1 ANNOTATOR INSTRUCTIONS

1239 We provide the instructions given to annotators below:

1240 # Instructions for Submitting a Puzzle

1242 To submit a puzzle, fork this repository and create a new branch. Then, create a new folder
 1243 '{puzzle_name}' in the 'data/puzzles' folder, and place the following files in it:
 1244 - 'metadata.json': A JSON file containing the metadata of the puzzle
 1245 - 'content.png': The image of the puzzle content
 1246 - 'figure_(N).png': (Optional) Figures illustrating the reasoning steps
 1247 For an example puzzle, see the 'data/puzzles/example' folder. After you are done, create a pull
 1248 request to merge your branch into the main repository.
 1249 Note, please replace any spaces in the puzzle name with '_' when creating the new folder!
 1250 ## Metadata
 1251 The 'metadata.json' file should contain a JSON object with the following fields:
 1252 | Field Name | Type | Description |
 1253 |-----|-----|-----|
 1254 | title | string | The title of the puzzle |
 1255 | flavor text | string | The flavor text of the puzzle, possibly empty |
 1256 | difficulty | string | The difficulty level of the puzzle (easy, medium, hard) |
 1257 | solution | string | The solution to the puzzle |
 1258 | reasoning | Step\[\] | An ordered list of reasoning [steps](#reasoning-step) towards the
 1259 solution |
 1260 | modality | string\[\] | A list of input [modalities](#a-list-of-input-modalities) the puzzle
 1261 contains |
 1262 | skills | string\[\] | A list of [skills](#a-list-of-reasoning-skills) required to solve the
 1263 puzzle |
 1264 | source | url | The link to the puzzle |
 1265 ### Reasoning Step
 1266 The 'reasoning' field should contain a list of 'Step' objects, which are represented as
 1267 dictionaries with the following fields:
 1268 | Field Name | Type | Description |
 1269 |-----|-----|-----|
 1270 | explanation | string | The textual explanation of the step |
 1271 | figure | file path | (Optional) File path to a figure illustrating the step |
 1272 Each of the explanation should begin with one of the following atomic actions:
 1273 - Pattern discovery: discover patterns / insights from current information
 1274 - E.g. discovering that current laser patterns are semaphores
 1275 - Sketching: sketching on or interacting with visual elements
 1276 - E.g. traversing through a maze
 1277 - E.g. connecting the dots
 1278 - Manipulation: manipulating or arranging a sequence of elements
 1279 - E.g. sorting alphabets in order
 1280 - E.g. applying cryptic encoding / decoding
 1281 - Combining / Chaining: combining or chaining multiple pieces of observations
 1282 - E.g. matching patterns in images with text segments
 1283 - Extraction: extracting information from one pattern or observation
 1284 - E.g. extracting letters from semaphore patterns
 1285 (Note: the exact wording of action is not important as long as it resembles one of the above
 1286 categories)
 1287 Each explanation step should consist of one action and the intermediate outcome of the action e.g.
 1288 Identify the pattern that (...), which is (...)
 1289 ## A List of Input Modalities
 1290 | Keyword | Description |
 1291 |-----|-----|
 1292 | 'text' | Textual information |
 1293 | 'visual' | Unstructured visual information e.g. images, icons, fonts, etc. |
 1294 | 'structured' | Structured visual information e.g. tables, graphs, crosswords, etc. |
 1295
 1296 ## A List of Reasoning Skills
 1297 | Keyword | Description |
 1298 |-----|-----|
 1299 | 'logic' | Logic reasoning e.g. rule deduction or inferring conclusion given partial
 1300 information |
 1301 | 'wordplay' | Manipulating words based on linguistic properties e.g. anagrams, homophones, etc. |
 1302 | 'spatial' | Spatial or visual understanding, manipulation and navigation e.g. mazes, connecting
 1303 dots, etc. |
 1304 | 'cryptic' | Encoding and decoding information e.g. ciphers, indexing, etc. |
 1305 | 'knowledge' | Leveraging domain-specific knowledge e.g. history, science, etc. |
 1306 | 'commonsense' | Applying common sense reasoning e.g. physical laws, social norms, etc. |
 1307 | 'tool_use' | Searching through an external database for information unlikely in model's training
 1308 data, such as Google Maps |
 1309
 1310
 1311
 1312
 1313
 1314
 1315
 1316
 1317
 1318
 1319
 1320
 1321
 1322
 1323
 1324
 1325
 1326
 1327
 1328
 1329
 1330
 1331
 1332
 1333
 1334
 1335
 1336
 1337
 1338
 1339
 1340
 1341
 1342
 1343
 1344
 1345
 1346
 1347
 1348
 1349
 1350
 1351
 1352
 1353
 1354
 1355
 1356
 1357
 1358
 1359
 1360
 1361
 1362
 1363
 1364
 1365
 1366
 1367
 1368
 1369
 1370
 1371
 1372
 1373
 1374
 1375
 1376
 1377
 1378
 1379
 1380
 1381
 1382
 1383
 1384
 1385
 1386
 1387
 1388
 1389
 1390
 1391
 1392
 1393
 1394
 1395
 1396
 1397
 1398
 1399
 1400
 1401
 1402
 1403
 1404
 1405
 1406
 1407
 1408
 1409
 1410
 1411
 1412
 1413
 1414
 1415
 1416
 1417
 1418
 1419
 1420
 1421
 1422
 1423
 1424
 1425
 1426
 1427
 1428
 1429
 1430
 1431
 1432
 1433
 1434
 1435
 1436
 1437
 1438
 1439
 1440
 1441
 1442
 1443
 1444
 1445
 1446
 1447
 1448
 1449
 1450
 1451
 1452
 1453
 1454
 1455
 1456
 1457
 1458
 1459
 1460
 1461
 1462
 1463
 1464
 1465
 1466
 1467
 1468
 1469
 1470
 1471
 1472
 1473
 1474
 1475
 1476
 1477
 1478
 1479
 1480
 1481
 1482
 1483
 1484
 1485
 1486
 1487
 1488
 1489
 1490
 1491
 1492
 1493
 1494
 1495
 1496
 1497
 1498
 1499
 1500
 1501
 1502
 1503
 1504
 1505
 1506
 1507
 1508
 1509
 1510
 1511
 1512
 1513
 1514
 1515
 1516
 1517
 1518
 1519
 1520
 1521
 1522
 1523
 1524
 1525
 1526
 1527
 1528
 1529
 1530
 1531
 1532
 1533
 1534
 1535
 1536
 1537
 1538
 1539
 1540
 1541
 1542
 1543
 1544
 1545
 1546
 1547
 1548
 1549
 1550
 1551
 1552
 1553
 1554
 1555
 1556
 1557
 1558
 1559
 1560
 1561
 1562
 1563
 1564
 1565
 1566
 1567
 1568
 1569
 1570
 1571
 1572
 1573
 1574
 1575
 1576
 1577
 1578
 1579
 1580
 1581
 1582
 1583
 1584
 1585
 1586
 1587
 1588
 1589
 1590
 1591
 1592
 1593
 1594
 1595
 1596
 1597
 1598
 1599
 1600
 1601
 1602
 1603
 1604
 1605
 1606
 1607
 1608
 1609
 1610
 1611
 1612
 1613
 1614
 1615
 1616
 1617
 1618
 1619
 1620
 1621
 1622
 1623
 1624
 1625
 1626
 1627
 1628
 1629
 1630
 1631
 1632
 1633
 1634
 1635
 1636
 1637
 1638
 1639
 1640
 1641
 1642
 1643
 1644
 1645
 1646
 1647
 1648
 1649
 1650
 1651
 1652
 1653
 1654
 1655
 1656
 1657
 1658
 1659
 1660
 1661
 1662
 1663
 1664
 1665
 1666
 1667
 1668
 1669
 1670
 1671
 1672
 1673
 1674
 1675
 1676
 1677
 1678
 1679
 1680
 1681
 1682
 1683
 1684
 1685
 1686
 1687
 1688
 1689
 1690
 1691
 1692
 1693
 1694
 1695
 1696
 1697
 1698
 1699
 1700
 1701
 1702
 1703
 1704
 1705
 1706
 1707
 1708
 1709
 1710
 1711
 1712
 1713
 1714
 1715
 1716
 1717
 1718
 1719
 1720
 1721
 1722
 1723
 1724
 1725
 1726
 1727
 1728
 1729
 1730
 1731
 1732
 1733
 1734
 1735
 1736
 1737
 1738
 1739
 1740
 1741
 1742
 1743
 1744
 1745
 1746
 1747
 1748
 1749
 1750
 1751
 1752
 1753
 1754
 1755
 1756
 1757
 1758
 1759
 1760
 1761
 1762
 1763
 1764
 1765
 1766
 1767
 1768
 1769
 1770
 1771
 1772
 1773
 1774
 1775
 1776
 1777
 1778
 1779
 1780
 1781
 1782
 1783
 1784
 1785
 1786
 1787
 1788
 1789
 1790
 1791
 1792
 1793
 1794
 1795
 1796
 1797
 1798
 1799
 1800
 1801
 1802
 1803
 1804
 1805
 1806
 1807
 1808
 1809
 1810
 1811
 1812
 1813
 1814
 1815
 1816
 1817
 1818
 1819
 1820
 1821
 1822
 1823
 1824
 1825
 1826
 1827
 1828
 1829
 1830
 1831
 1832
 1833
 1834
 1835
 1836
 1837
 1838
 1839
 1840
 1841
 1842
 1843
 1844
 1845
 1846
 1847
 1848
 1849
 1850
 1851
 1852
 1853
 1854
 1855
 1856
 1857
 1858
 1859
 1860
 1861
 1862
 1863
 1864
 1865
 1866
 1867
 1868
 1869
 1870
 1871
 1872
 1873
 1874
 1875
 1876
 1877
 1878
 1879
 1880
 1881
 1882
 1883
 1884
 1885
 1886
 1887
 1888
 1889
 1890
 1891
 1892
 1893
 1894
 1895
 1896
 1897
 1898
 1899
 1900
 1901
 1902
 1903
 1904
 1905
 1906
 1907
 1908
 1909
 1910
 1911
 1912
 1913
 1914
 1915
 1916
 1917
 1918
 1919
 1920
 1921
 1922
 1923
 1924
 1925
 1926
 1927
 1928
 1929
 1930
 1931
 1932
 1933
 1934
 1935
 1936
 1937
 1938
 1939
 1940
 1941
 1942
 1943
 1944
 1945
 1946
 1947
 1948
 1949
 1950
 1951
 1952
 1953
 1954
 1955
 1956
 1957
 1958
 1959
 1960
 1961
 1962
 1963
 1964
 1965
 1966
 1967
 1968
 1969
 1970
 1971
 1972
 1973
 1974
 1975
 1976
 1977
 1978
 1979
 1980
 1981
 1982
 1983
 1984
 1985
 1986
 1987
 1988
 1989
 1990
 1991
 1992
 1993
 1994
 1995
 1996
 1997
 1998
 1999
 2000
 2001
 2002
 2003
 2004
 2005
 2006
 2007
 2008
 2009
 2010
 2011
 2012
 2013
 2014
 2015
 2016
 2017
 2018
 2019
 2020
 2021
 2022
 2023
 2024
 2025
 2026
 2027
 2028
 2029
 2030
 2031
 2032
 2033
 2034
 2035
 2036
 2037
 2038
 2039
 2040
 2041
 2042
 2043
 2044
 2045
 2046
 2047
 2048
 2049
 2050
 2051
 2052
 2053
 2054
 2055
 2056
 2057
 2058
 2059
 2060
 2061
 2062
 2063
 2064
 2065
 2066
 2067
 2068
 2069
 2070
 2071
 2072
 2073
 2074
 2075
 2076
 2077
 2078
 2079
 2080
 2081
 2082
 2083
 2084
 2085
 2086
 2087
 2088
 2089
 2090
 2091
 2092
 2093
 2094
 2095
 2096
 2097
 2098
 2099
 2100
 2101
 2102
 2103
 2104
 2105
 2106
 2107
 2108
 2109
 2110
 2111
 2112
 2113
 2114
 2115
 2116
 2117
 2118
 2119
 2120
 2121
 2122
 2123
 2124
 2125
 2126
 2127
 2128
 2129
 2130
 2131
 2132
 2133
 2134
 2135
 2136
 2137
 2138
 2139
 2140
 2141
 2142
 2143
 2144
 2145
 2146
 2147
 2148
 2149
 2150
 2151
 2152
 2153
 2154
 2155
 2156
 2157
 2158
 2159
 2160
 2161
 2162
 2163
 2164
 2165
 2166
 2167
 2168
 2169
 2170
 2171
 2172
 2173
 2174
 2175
 2176
 2177
 2178
 2179
 2180
 2181
 2182
 2183
 2184
 2185
 2186
 2187
 2188
 2189
 2190
 2191
 2192
 2193
 2194
 2195
 2196
 2197
 2198
 2199
 2200
 2201
 2202
 2203
 2204
 2205
 2206
 2207
 2208
 2209
 2210
 2211
 2212
 2213
 2214
 2215
 2216
 2217
 2218
 2219
 2220
 2221
 2222
 2223
 2224
 2225
 2226
 2227
 2228
 2229
 2230
 2231
 2232
 2233
 2234
 2235
 2236
 2237
 2238
 2239
 2240
 2241
 2242
 2243
 2244
 2245
 2246
 2247
 2248
 2249
 2250
 2251
 2252
 2253
 2254
 2255
 2256
 2257
 2258
 2259
 2260
 2261
 2262
 2263
 2264
 2265
 2266
 2267
 2268
 2269
 2270
 2271
 2272
 2273
 2274
 2275
 2276
 2277
 2278
 2279
 2280
 2281
 2282
 2283
 2284
 2285
 2286
 2287
 2288
 2289
 2290
 2291
 2292
 2293
 2294
 2295
 2296
 2297
 2298
 2299
 2300
 2301
 2302
 2303
 2304
 2305
 2306
 2307
 2308
 2309
 2310
 2311
 2312
 2313
 2314
 2315
 2316
 2317
 2318
 2319
 2320
 2321
 2322
 2323
 2324
 2325
 2326
 2327
 2328
 2329
 2330
 2331
 2332
 2333
 2334
 2335
 2336
 2337
 2338
 2339
 2340
 2341
 2342
 2343
 2344
 2345
 2346
 2347
 2348
 2349
 2350
 2351
 2352
 2353
 2354
 2355
 2356
 2357
 2358
 2359
 2360
 2361
 2362
 2363
 2364
 2365
 2366
 2367
 2368
 2369
 2370
 2371
 2372
 2373
 2374
 2375
 2376
 2377
 2378
 2379
 2380
 2381
 2382
 2383
 2384
 2385
 2386
 2387
 2388
 2389
 2390
 2391
 2392
 2393
 2394
 2395
 2396
 2397
 2398
 2399
 2400
 2401
 2402
 2403
 2404
 2405
 2406
 2407
 2408
 2409
 2410
 2411
 2412
 2413
 2414
 2415
 2416
 2417
 2418
 2419
 2420
 2421
 2422
 2423
 2424
 2425
 2426
 2427
 2428
 2429
 2430
 2431
 2432
 2433
 2434
 2435
 2436
 2437
 2438
 2439
 2440
 2441
 2442
 2443
 2444
 2445
 2446
 2447
 2448
 2449
 2450
 2451
 2452
 2453
 2454
 2455
 2456
 2457
 2458
 2459
 2460
 2461
 2462
 2463
 2464
 2465
 2466
 2467
 2468
 2469
 2470
 2471
 2472
 2473
 2474
 2475
 2476
 2477
 2478
 2479
 2480
 2481
 2482
 2483
 2484
 2485
 2486
 2487
 2488
 2489
 2490
 2491
 2492
 2493
 2494
 2495
 2496
 2497
 2498
 2499
 2500
 2501
 2502
 2503
 2504
 2505
 2506
 2507
 2508
 2509
 2510
 2511
 2512
 2513
 2514
 2515
 2516
 2517
 2518
 2519
 2520
 2521
 2522
 2523
 2524
 2525
 2526
 2527
 2528
 2529
 2530
 2531
 2532
 2533
 2534
 2535
 2536
 2537
 2538
 2539
 2540
 2541
 2542
 2543
 2544
 2545
 2546
 2547
 2548
 2549