PUZZLEWORLD: A BENCHMARK FOR MULTIMODAL, OPEN-ENDED REASONING IN PUZZLEHUNTS

Anonymous authors
Paper under double-blind review

ABSTRACT

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

1 INTRODUCTION

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

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

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

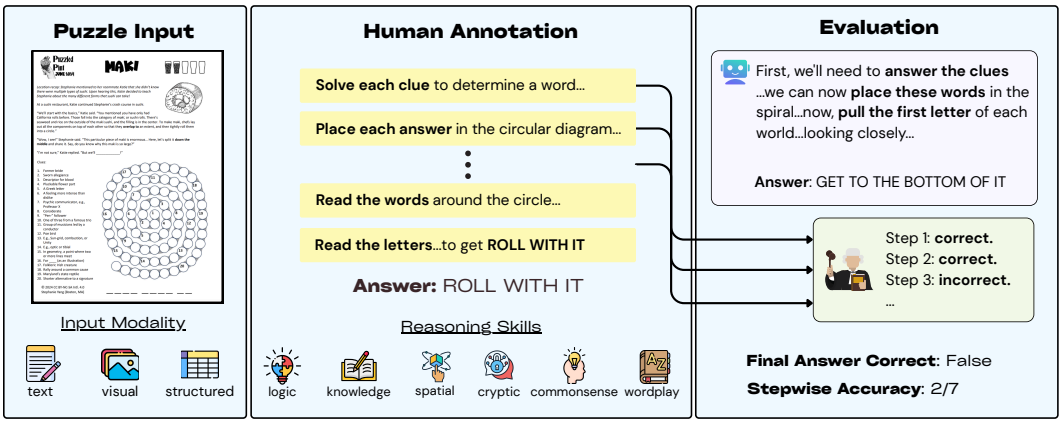


Figure 1: **Overview of PUZZLEWORLD**: PUZZLEWORLD is a dataset of complex puzzles that lack explicit instructions, requiring solvers to deduce the final answer from nuanced, multimodal cues from the puzzle content as well as external domain-specific knowledge. The raw puzzles and solutions are sourced 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 C.2.

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 & Parreiras, 2020; Singh et al., 2017), but also in more complex domains such as mathematical reasoning, programming, and log-

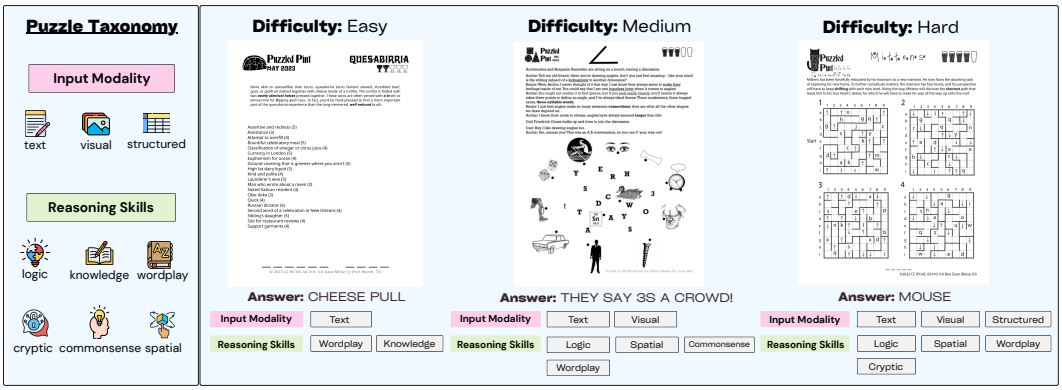
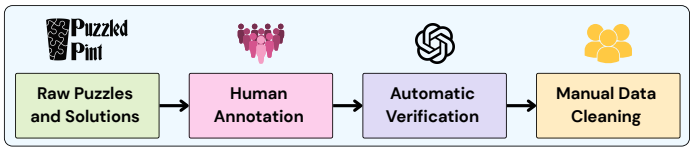


Figure 2: **Overview of samples from PUZZLEWORLD. Left:** To gain a deeper understanding of model performance on PUZZLEWORLD, each puzzle is annotated with the input modalities of the puzzle content, the reasoning skills required to solve the puzzle, and step-by-step reasoning steps. **Right:** Example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

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 OlympiadBench (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. These tasks often demand 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.



Statistic	Value
Total # of puzzles	667
Avg. # of Reasoning Steps	5.4
Percent # of Visual Reasoning Steps	12.3%
Avg. Word Count per Reasoning Step	22.5
Correlation between Difficulty and # of Reasoning Steps	0.24

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

Table 1: **Count of puzzles across modalities and difficulties.** Across all modalities, the distribution of difficulties is similar.

	Easy	Medium	Hard
Text	131	322	151
Visual	90	226	111
Structured	59	181	108

We identify six core cognitive abilities essential for effective puzzle-solving in PuzzleWorld. 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.

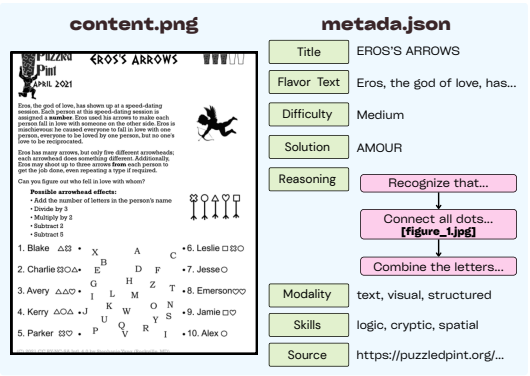


Figure 4: **Illustration of metadata schema:** All puzzles are annotated with accompanying metadata, which includes the title, flavor text, difficulty, final answer, reasoning steps, input modalities, reasoning skills, and the link to the puzzle.

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 C.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.

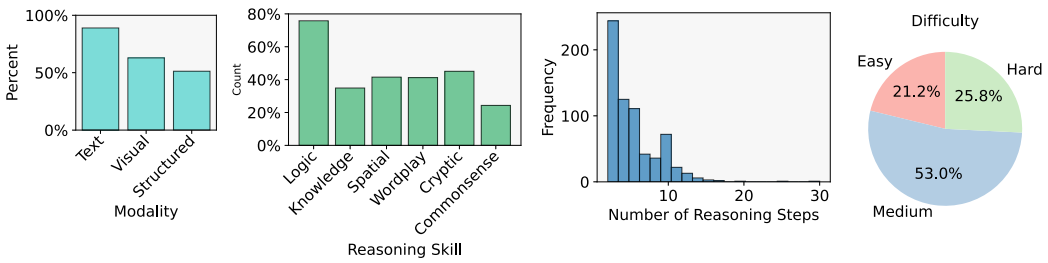


Figure 5: **PUZZLEWORLD dataset statistics.** Distributions of modalities and reasoning skills are balanced. While the majority of puzzles are of medium difficulty, there is significant number of easy and hard puzzles. The number of reasoning steps follows a long-tail distribution, with many solutions 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 today's best closed-source reasoning models on PUZZLEWORLD, including GPT-o3 (OpenAI, 2025), GPT-4o (Achiam et al., 2023), Claude Opus 4 with Extended Thinking (Anthropic, 2025), Gemini-2.5-Pro (Comanici et al., 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 D 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 we gathered 9 Novices and 9 Enthusiasts across high school and college ages. We sampled 5% puzzles from PUZZLEWORLDand 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.

		Overall		Text		Visual		Structured	
	Model	**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
	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	*not applicable for human baseline*					
	Human Enthusiast	44.44	51.70						
	Human Expert	**100.0**	**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 reasoningdespite 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 D.2.

Table 3: **Fine-tuned model accuracy on PUZZLEWORLD.** Stepwise accuracy improves when fine-tuning Intern-VL3 on reasoning traces, while final answer accuracy remains unchanged.

Model	**Acc.**	**Step.**
Base	**0.76%**	4.78%
Fine-tuned (Answer-only)	0.00%	2.96%
Fine-tuned (Reasoning)	**0.76%**	**11.00%**

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.

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%**

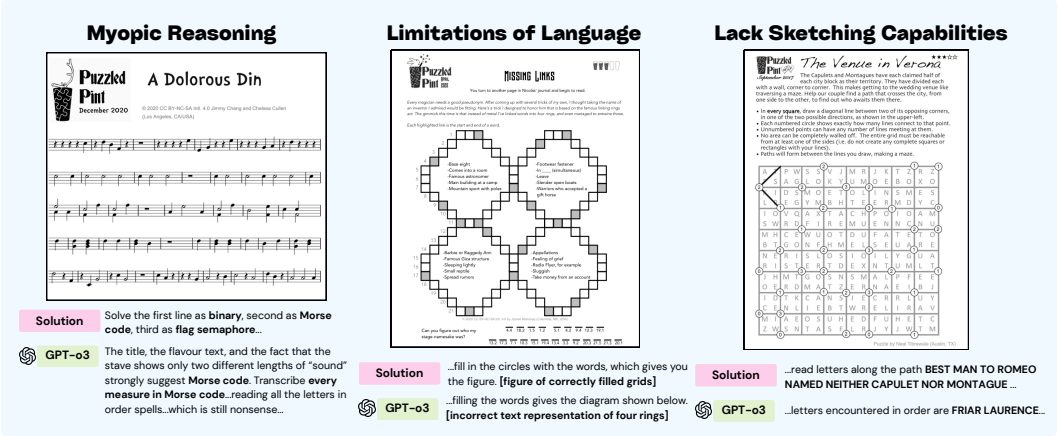


Figure 6: **Example puzzle errors. Left:** (myopic reasoning) The model is unable to backtrack when it hits a dead end. **Middle:** (language bottleneck) The model misrepresents the visual contents due to inherent limitations of texts. **Right:** (sketching errors) The model fails to execute the visual sketching steps to obtain correct intermediate outputs. High-resolution images are in Appendix C.2.

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.

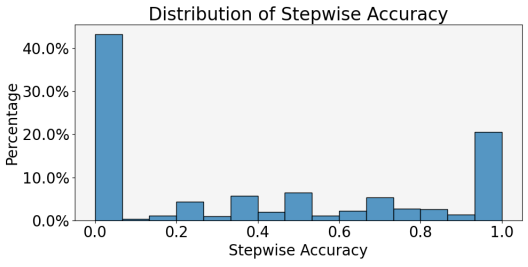


Figure 7: **Stepwise accuracy distribution of GPT-o3.** GPT-o3 receives stepwise accuracy of 0 for most puzzles, highlighting the model's myopic reasoning tendencies and its inability to backtrack after committing to an incorrect first step.

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 fails to correctly identify even the first step of the solution. Once committed to an incorrect path, the model rarely recovers, highlighting its brittle reasoning and a lack of dynamic self-correction, especially when it cannot rely on external environments for verification.

Limitations of language. Modern multimodal models rely heavily on language-based reasoning strategies, such as chain-of-thought and code generation. However, this dependence becomes a bottleneck in puzzles with complex visual structure. In Figure 6, the puzzle is composed of four interlocking loops arranged in a clover-like pattern. This layout is visually intuitive, but difficult to represent in text.

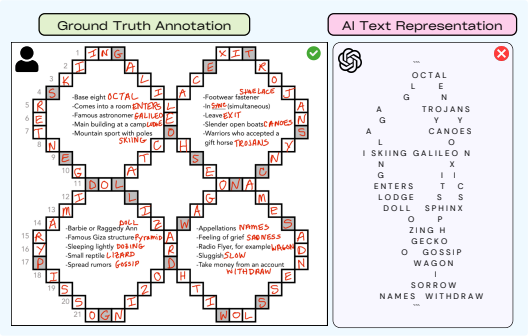


Figure 8: **Limitations of text**. An example failure case where GPT-o3 fails to represent a complex structured puzzle into text.

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

Multimodal reasoning needs sketching. While frontier models have made notable progress in logical deduction and arithmetic reasoning, they consistently underperform on spatial tasks that require sketching, drawing, and manipulating visual structure, such as decoding based on spatial arrangements or tracing paths through grids and mazes. In Figure 6, the model correctly solves the individual clues in a grid-based puzzle but fails to trace the intended path, resulting in an incorrect final answer. Humans naturally rely on sketching or mental imagery to reason through such spatial challenges, using external or internal visualizations to keep track of evolving structure. The absence of such capabilities in current models reveals a critical gap: without the ability to sketch and update a persistent visual representation, models are prone to failure in tasks that depend on spatial coherence.

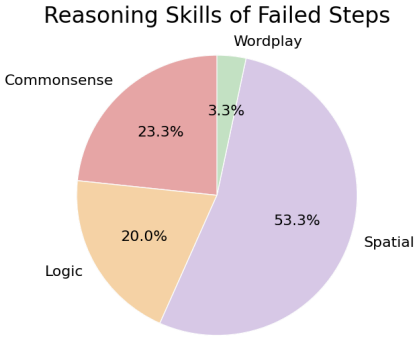


Figure 9: **Reasoning skills of failed steps.** We annotated the bottleneck steps with their reasoning skills.

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

6 Conclusion

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

Ethics and Reproducibility Statement

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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, 2025. Accessed: 2025-05-16.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A Limitations and Broader Impact

To ensure consistency and standardization across the dataset, we excluded puzzles involving under-explored or difficult-to-represent modalities such as audio, video, or interactive file-based inputs. As a result, PuzzleWorld may not fully capture the breadth of sensory and interaction-based reasoning found in some real-world, more challenging puzzlehunts. Additionally, unlike Wang et al. (2025) that uses human annotators to transcribe textual and visual components separately, we preserve the puzzle content in its original image format and focus annotation efforts on intermediate reasoning traces. While this allows PuzzleWorld to provide richer annotation of the solution reasoning process, it may also introduce variability in model performance depending on the quality of their OCR capabilities.

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

One possible concern with our grading scheme is that a model might "hallucinate" the correct final answer without engaging in proper reasoning. However, such a case is extremely rare in PuzzleWorld. The high-quality, human-designed puzzles are deliberately constructed to discourage superficial guessing, and even experienced human solvers cannot easily infer the answer without following the intended logic. In a thorough manual inspection of models' puzzle responses, we did not find any case where the model arrived at the final answer without demonstrating the necessary reasoning.

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

B Use of Large Language Models

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

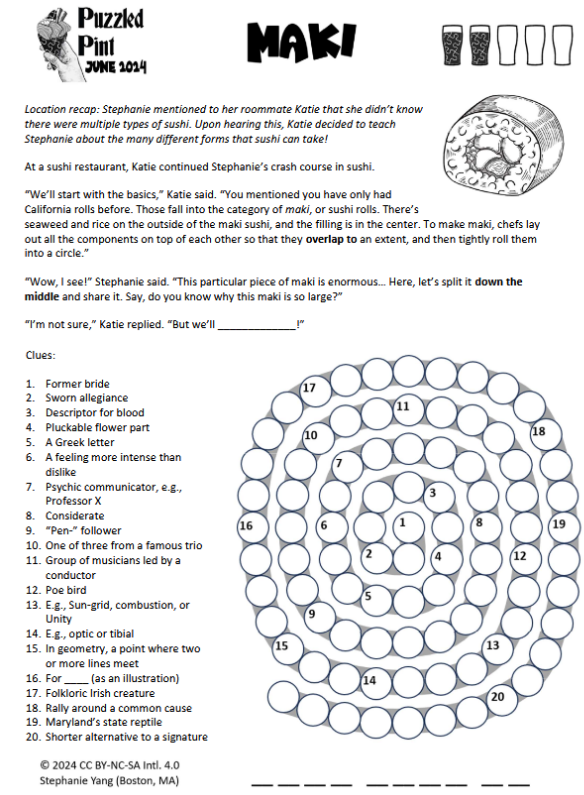
C PuzzleWorld details

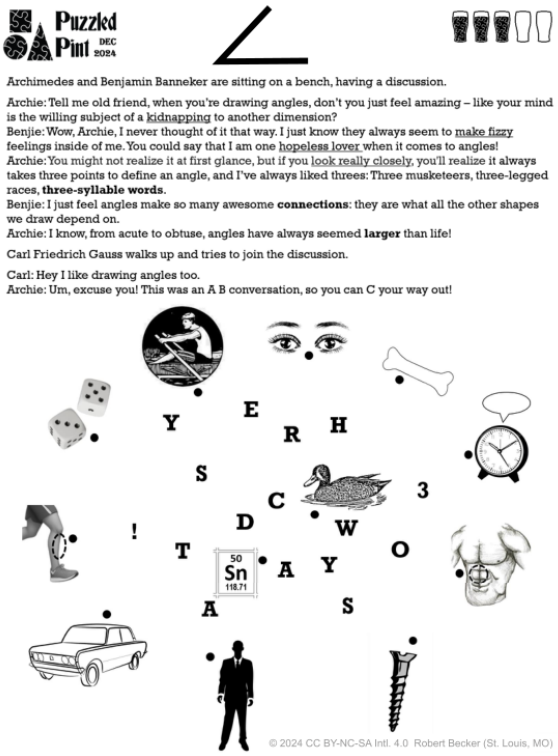
C.1 Checking for dataset contamination

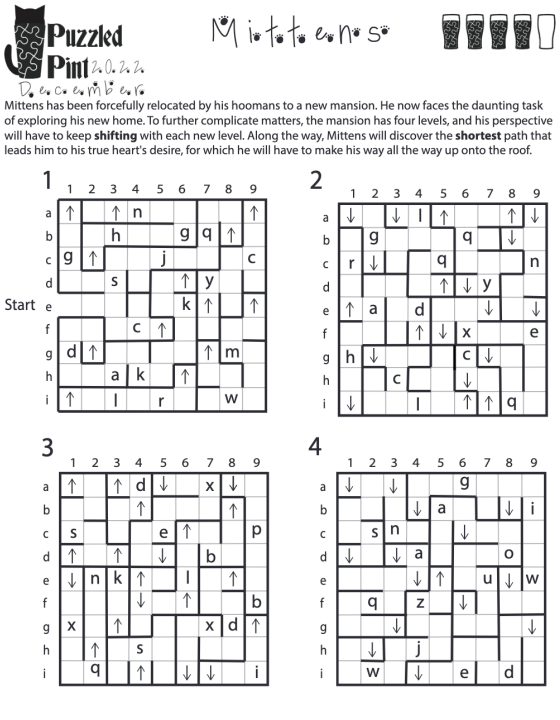
To assess the possibility of data contamination, we test whether GPT-o3 (OpenAI, 2025) has memorized any of the puzzles in our dataset. Specifically, inspired by prior work Tanzer et al. (2023); Chi et al. (2024), we prompt the model to reconstruct the flavor text for 40 randomly sampled puzzles out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

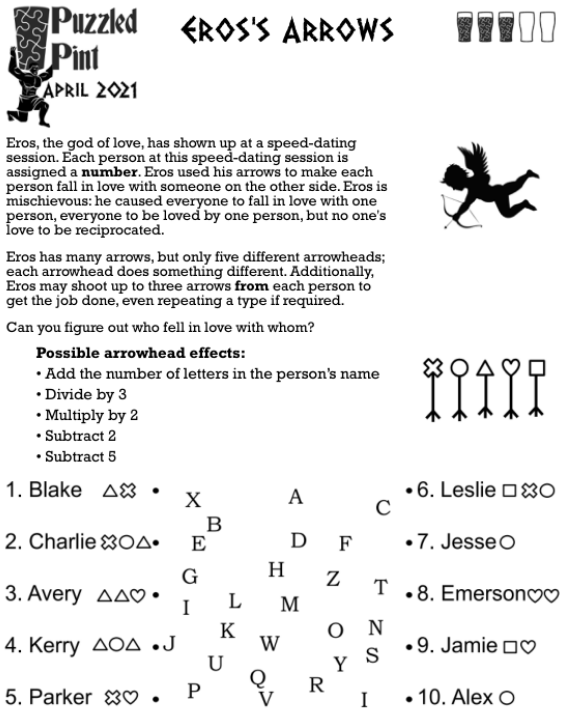
C.2 PUZZLEWORLD IMAGE SAMPLES

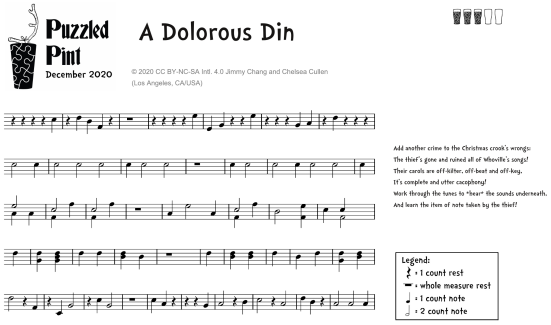
We provide high-resolution images of puzzle samples used in this paper.




Puzzled Pint
DEC 2024
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 B conversation, so you can C your way out!
Y E R H S C 3 D W ! T O A Y A S
50 Sn 118.71
© 2024 CC BY-NC-SA Intl. 4.0 Robert Becker (St. Louis, MO)


Puzzled Pint
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.
1
2
3
4
Start
©2022 CC BY-NC-SA Intl. 4.0 Ben Zoon (Boise, ID)


Puzzled Pint
APRIL 2021
EROS'S ARROWS
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
2. Charlie
3. Avery
4. Kerry
5. Parker
6. Leslie
7. Jesse
8. Emerson
9. Jamie
10. Alex
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
(C) 2021 CC BY-NC-SA Intl. 4.0 by Stephanie Yang (Rockville, MD)


Puzzled Pint
December 2020
A Dolorous Din
© 2020 CC BY-NC-SA Intl. 4.0 Jimmy Chang and Chelsea Cullen (Los Angeles, CA/USA)
Add another crime to the Christmas crook's wrongs:
The thief's gone and ruined all of Whoville's songs!
Their carols are off-kilter, off-beat and off-key,
It's complete and utter cacophony!
Work through the tunes to "hear" the sounds underneath,
And learn the item of note taken by the thief!
Legend:
𝄽 : 1 count rest
𝄻 : whole measure rest
♩ : 1 count note
𝅗𝅥 : 2 count note

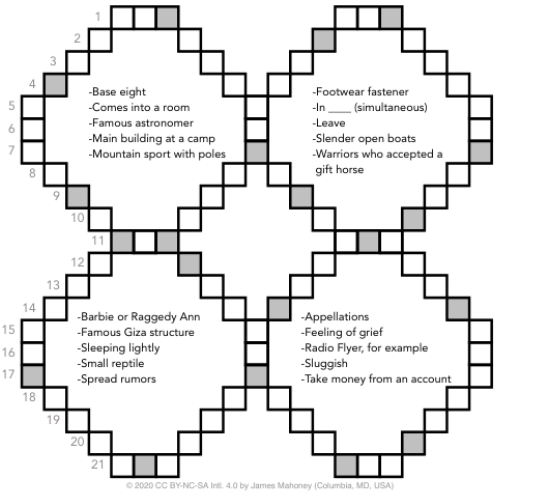
Puzzled Pint April 2020

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.



Can you figure out who my stage-namesake was?

4.4 18.2 1.5 1.2 5.1 4.2 9.4 12.3 19.1

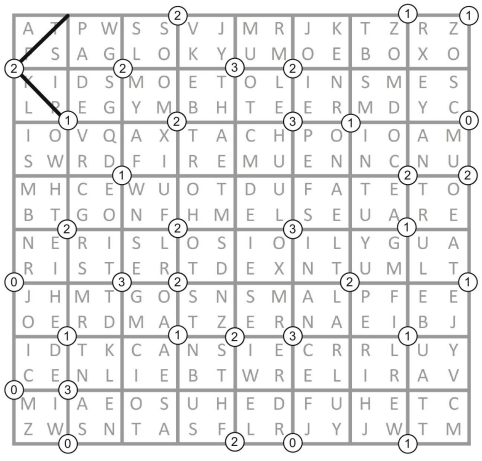
13.2 17.3 7.1 10.3 15.1 19.4 13.4 3.3 9.2 20.3 21.5 21.3 20.1

Puzzled Pint September 2017

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)

D BENCHMARKING DETAILS

D.1 COMPUTE RESOURCES

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.

D.2 SUPERVISED FINETUNING

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.

D.3 PROMPT FOR BENCHMARKING

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

```
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".
- Alpha-numeric codes are also very common. If you end up with a list of numbers try replacing
the numbers with the corresponding letters like this: 1 = A, 2 = B, 3 = C... 26 = Z.
Occasionally, these types of codes will "wrap around", so don't despair if you see a
number greater than 26. Just subtract 26 and try again. In this scenario 27 (27-26 = 1) =
A, 28 (28-26 = 2) = B etc. If you try this and it doesn't work, try other numeric codes
such as ASCII.
- Often a puzzle repeats a strategy multiple times.

You will likely need to backtrack frequently, so make sure to write out your steps as you go.
If you get stuck, try to think of a new way to approach the puzzle. Try:
- Rereading the title and the flavor text. These are the most important hints about what type
of strategies, themes or cultural references might be used to solve the puzzle.
- Checking for pop culture references
- Checking for references to a song/poem/book/movie/TV show

For strings, examples of strategies you might try include:
- Alphabetizing
- Using leftover letters to spell something
- Rearranging the letters (aka anagrams or "transposing")
- Seeing if there are any acronyms
- Diagonalizing (taking the first letter of the first answer, the second letter of the second
answer, etc.)
- Looking for unusual letter frequencies
- Puns and homophones
- Shifting from letters to numbers

For numbers, try:
- Shifting from numbers to letters
- Using it as a phone number
- Treating numbers as dates
- Treating numbers as ASCII numbers
- Seeing if there are any strange sequences
- Seeing if prime numbers are involved

For images, try:
- Looking at it in a mirror
- Squinting at it from far away
- Tilting it
- Looking at it upside down
- Looking through it
- Transcribing it neatly
```

We additionally append the user prompt:

```
Your task is to solve the following puzzle. The attached images are presented in the order
they are referenced in the text.

The puzzle's title is: {}
The puzzle's flavor text is: {}

---
Write out a step-by-step solution to the puzzle. At the end of your solution, write your
answer in the following format:
Answer: <answer>
```

Below is the prompt for LLM judge:

```
Answer Equivalence Instructions:
Using the puzzle and the reference solution, grade the candidate solution as follows.

For every reasoning step of the reference solution, output True if the candidate solution both
includes
the step and achieves the same intermediate result of the step, otherwise False.
Explain why the candidate's solution did or did not get the reasoning step correct.
Do not add more steps than there are in the reference solution and evaluate every step
in the reference solution.
There is a exception in scoring for the last reasoning step. Identify the candidate output
solution.
If the candiate output solution is the exact same as the reference solution answer of
\"{puzzle_solution}\",
then output final step as true.
```

E ANNOTATOR DETAILS

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

E.1 ANNOTATOR INSTRUCTIONS

We provide the instructions given to annotators below:

```
# Instructions for Submitting a Puzzle
To submit a puzzle, fork this repository and create a new branch. Then, create a new folder
`{puzzle_name}` in the `data/puzzles` folder, and place the following files in it:

- `metadata.json`: A JSON file containing the metadata of the puzzle
- `content.png`: The image of the puzzle content
- `figure_{N}.png`: (Optional) Figures illustrating the reasoning steps

For an example puzzle, see the `data/puzzles/example` folder. After you are done, create a pull
request to merge your branch into the main repository.

Note, please replace any spaces in the puzzle name with `_` when creating the new folder!

## Metadata
The `metadata.json` file should contain a JSON object with the following fields:

| Field Name | Type  | Description                                                    |
|--------------|------------|----------------------------------------------------------|
| title    | string  | The title of the puzzle             |
| flavor text | string | The flavor text of the puzzle, possibly empty |
| difficulty | string | The difficulty level of the puzzle (easy, medium, hard) |
| solution | string  | The solution to the puzzle          |
| reasoning | Step\[ \] | An ordered list of reasoning [steps](#reasoning-step) towards the
solution |
| modality | string\[ \] | A list of input [modalities](#a-list-of-input-modalities) the puzzle
contains |
| skills   | string\[ \] | A list of [skills](#a-list-of-reasoning-skills) required to solve the
puzzle |
| source   | url     | Thel link to the puzzle             |

### Reasoning Step
The `reasoning` field should contain a list of `Step` objects, which are represented as
dictionaries with the following fields:
| Field Name | Type  | Description                                  |
|--------------|-----------|-----------------------------------------------------------|
| explanation | string | The textual explanation of the step   |
| figure   | file path | (Optional) File path to a figure illustrating the step |

Each of the explanation should begin with one of the following atomic actions:
- Pattern discovery: discover patterns / insights from current information
  - E.g. discovering that current laser patterns are semaphores
- Sketching: sketching on or interacting with visual elements
  - E.g. traversing through a maze
  - E.g. connecting the dots
- Manipulation: manipulating or arranging a sequence of elements
  - E.g. sorting alphabets in order
  - E.g. applying cryptic encoding / decoding
```

```
- Combining / Chaining: combining or chaining multiple pieces of observations
  - E.g. matching patterns in images with text segments
- Extraction: extracting information from one pattern or observation
  - E.g. extracting letters from semaphore patterns

(Note: the exact wording of action is not important as long as it resembles one of the above
categories)

Each explanation step should consist of one action and the intermediate outcome of the action e.g.
Identify the pattern that (...), which is (...)

### A List of Input Modalities
| Keyword    | Description                                        |
|----------------|-------------------------------------------------------------------|
| `text`     | Textual information                                |
| `visual`   | Unstructued visual information e.g. images, icons, fonts, etc. |
| `structured` | Structured visual information e.g. tables, graphs, crosswords, etc.|


### A List of Reasoning Skills
| Keyword    | Description                                                                                    |
|----------------|------------------------------------------------------------------------------------------------|
| `logic`    | Logic reasoning e.g. rule deduction or inferring conclusion given partial
information |
| `wordplay` | Manipulating words based on linguistic properties e.g. anagrams, homophones, etc. |
| `spatial`  | Spatial or visual understanding, manipulation and navigation e.g. mazes, connecting
dots, etc. |
| `cryptic`  | Encoding and decoding information e.g. ciphers, indexing, etc.  |
| `knowledge` | Leverarging domain-specific knowledge e.g. history, science, etc. |
| `commonsense` | Applying common sense reasoning e.g. physical laws, social norms, etc. |
| `tool_use` | Searching through an external database for information unlikely in model's training
data, such as Google Maps |
```