Lexical Recall or Logical Reasoning: Probing the Limits of Reasoning Abilities in Large Language Models

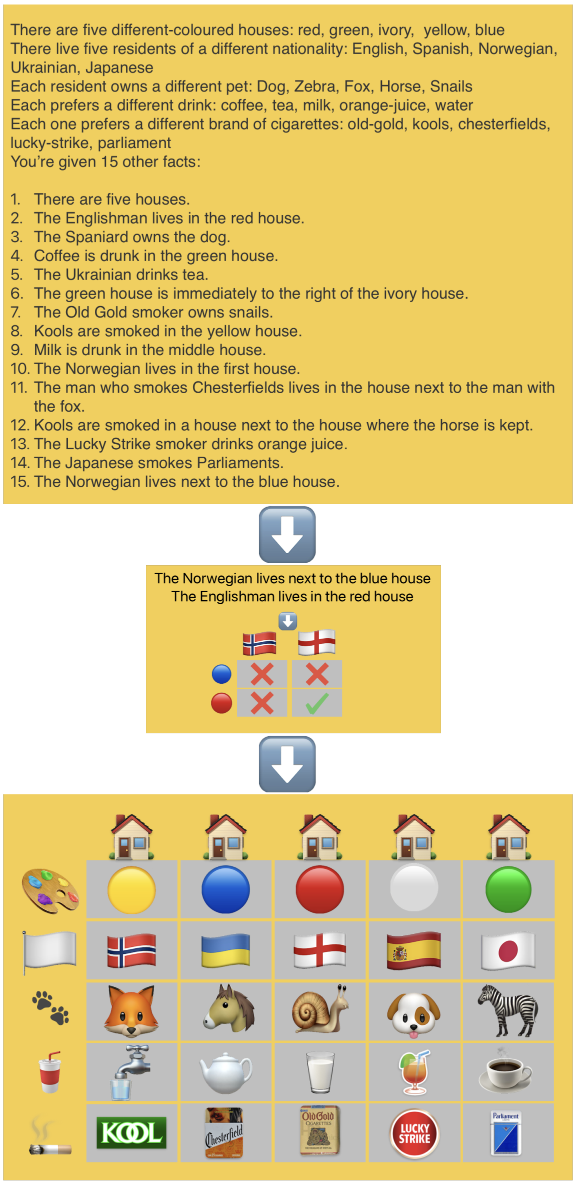
Henrike Beyer and Chris Reed
Centre for Argument Technology
University of Dundee
Dundee DD1 4HN, United Kingdom
2579207@dundee.ac.uk; c.a.reed@dundee.ac.uk

Abstract

Despite the increasing interest in the reasoning abilities of Large Language Models (LLMs), existing work shows limitations in assessing logic abilities independently from lexical memory. We address this gap with Mystery-Zebra. This robust two-part benchmark (4,290 puzzles) challenges the logic abstraction abilities of LLMs in two setups: (1) a lexical obfuscation setup tests the dependence of LLMs on lexical content based on two canonical grid puzzles widely spread on the Internet; (2) a set of new grid puzzles in 42 different sizes and 12 difficulty levels tests how the formal difficulty degree of a puzzle affects LLMs. We test open and closed-weight LLMs on both parts of the benchmark. The results on part two suggest that model sizes up to 70B parameters have only a minor influence when solving newly generated puzzles, while performance mainly relates to the number of items in the puzzle. The results on the first part of the benchmark suggest that the applied obfuscation strategies help to mitigate effects of logic puzzles being part of LLM training data, showing a drastic drop in performance for obfuscated versions of well-known puzzles. In addition we conduct a case-study on the first part of the benchmark predicting the position of single items, unveiling that the reasoning abilities of LLMs are mainly limited to a few consecutive steps of reasoning.[1]

1 Introduction

Alongside the general improvement in Large Language Model (LLM) performance across NLP tasks, the trust in their reasoning abilities increases. Their deployment in real-world applications calls for a critically informed view on LLM reasoning since misconceptualising these abilities has severe consequences.



There are five different-coloured houses: red, green, ivory, yellow, blue
There live five residents of a different nationality: English, Spanish, Norwegian, Ukrainian, Japanese
Each resident owns a different pet: Dog, Zebra, Fox, Horse, Snails
Each prefers a different drink: coffee, tea, milk, orange-juice, water
Each one prefers a different brand of cigarettes: old-gold, kools, chesterfields, lucky-strike, parliament
You're given 15 other facts:

1. There are five houses.
2. The Englishman lives in the red house.
3. The Spaniard owns the dog.
4. Coffee is drunk in the green house.
5. The Ukrainian drinks tea.
6. The green house is immediately to the right of the ivory house.
7. The Old Gold smoker owns snails.
8. Kools are smoked in the yellow house.
9. Milk is drunk in the middle house.
10. The Norwegian lives in the first house.
11. The man who smokes Chesterfields lives in the house next to the man with the fox.
12. Kools are smoked in a house next to the house where the horse is kept.
13. The Lucky Strike smoker drinks orange juice.
14. The Japanese smokes Parliaments.
15. The Norwegian lives next to the blue house.

The Norwegian lives next to the blue house
The Englishman lives in the red house

Figure 1: Grid puzzles like the "Zebra" puzzle encode relationships between items in natural language clues. Through deduction and elimination, these clues can be interpreted to populate a unique solution grid.

While classic reasoning benchmarks for LLMs cover basic maths, rule-based, or inferential reasoning (Wan et al., 2024; Mirzadeh et al., 2024; Gui et al., 2024; Wang, 2024), another line of re-

[1] The code used to create obfuscated puzzle variants, run experiments, and evaluate results is available under: https://github.com/arg-tech/MysteryZebra

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13532–13557
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

search assesses LLMs on games like Minecraft, card games, and logic puzzles (e.g. Sudoku) (Wang et al., 2023; Gupta, 2023; Guo et al., 2023; Li et al., 2024a; Shah et al., 2024; Saha et al., 2024).

Nevertheless, the intense scaffolding needed to apply LLMs in games blurs insights into their reasoning abilities. Further, abstract puzzles like Sudoku rely on more abstract non-linguistic representations offering a limited perspective on the link between language and reasoning in LLMs. In contrast, grid puzzles like the popular "Zebra"[2] and "Einstein"[3] puzzle offer a constrained task to evaluate LLM reasoning abilities in a natural-language setup. As illustrated in Figure 1, a typical grid puzzle contains a set of clues describing the relative position of items of the same syntactic domain (e.g. pets). The correct solution can be translated to a grid, where each row represents a domain, where the items are positions relatively to each other in columns. For a successful solution, the model needs to identify clue(s) leading to a valid elimination of positioning options or a correct positioning, keep track of the grid's current status, and ensure a conflict-free solution grid.

Grid-puzzles are rule-based (Giadikiaroglou et al., 2024) and therefore have a strict set of constraints to be satisfied in the solution grid. These constraints can take the form of symbolic logic expressions, turning the puzzles into Constraint Satisfaction Problems (CSPs). This allows to decouple the logical structure of the the clue from its linguistic representation. In the solution process, it is irrelevant whether the clues involve pets or beverages as these are just variables.

Figure 2 details the workflow of this paper, which leverages the similarity to CSPs to propose Mystery Zebra. The two parts of this benchmark challenge the reasoning capabilities of LLMs necessary to solve logic puzzles.

Part one of the benchmark targets the dependence of LLMs on residual lexical clues from the training content when solving a complex logic puzzle. In order to make the logical structure independent from the linguistic representation, we turn each clue into its symbolic logical representation; e.g. The dog-owner drinks coffee → A = B. Similarly to Valmeekam et al. (2023), we apply different obfuscation techniques to manipulate the popular "Zebra" and "Einstein" puzzle instances, which are widely spread on the Internet. We manipulate the phrasing of clues, but also the lexical content of the variables in the clues without changing the underlying symbolic structure. This helps to assess the dependence of LLMs on the exact lexical content of puzzles that are likely contained in their training data. Part two of the benchmark assesses the effect of grid size and formal difficulty of clues. It includes 4,250 newly generated puzzles in 12 formal difficulty classes and 42 sizes. While existing datasets control difficulty mainly based on the grid size or coarse difficulty categories, Mystery-Zebra uses the amount of information conveyed by the clues to judge difficulty on a fine-grained basis.

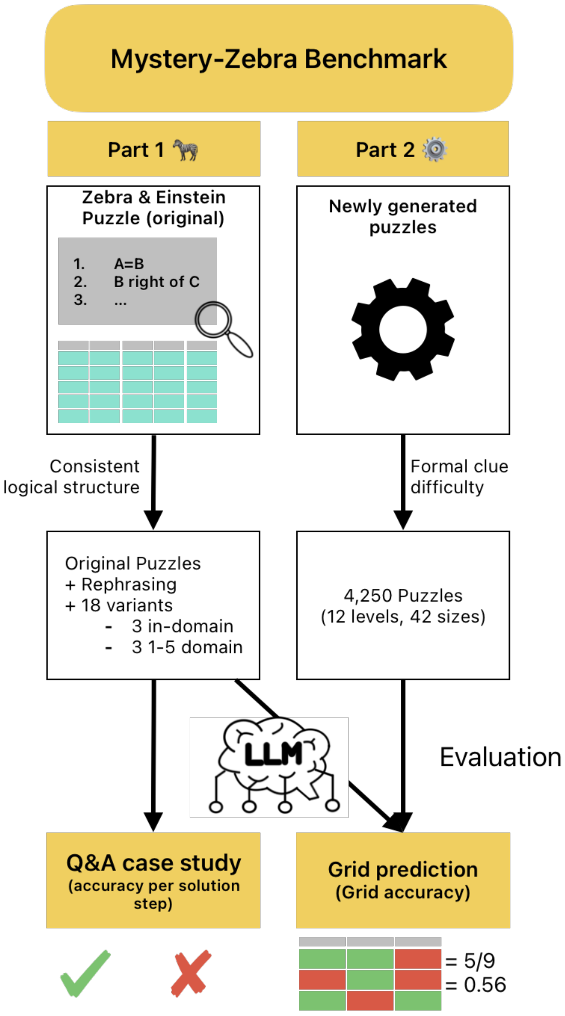


Figure 2: Workflow of creating and testing the Mystery Zebra benchmark. In two distinct parts: (1) contains the two canonical puzzle variants ("Zebra" and "Einstein") with lexical manipulations of them; (2) with 4,250 newly generated puzzles in 42 sizes and with 12 difficulty levels defined through formal difficulty of the puzzle clues. Both parts are tested in a grid prediction setup, while only part 1 is evaluated in a Q&A case-study.

[2] https://en.wikipedia.org/wiki/Zebra_Puzzle

[3] https://www.chessandpoker.com/einsteins-problem-solution.html

A range of open and closed weight LLMs is evaluated on both parts are evaluated via grid-accuracy when predicting the whole solution grid. In addition, part two is used to conduct a case study in a Q&A setup.

The contributions of this paper are 7-fold: (1) we introduce the first adaptable reasoning benchmark enabling multidimensional analysis of LLM reasoning; (2) we evaluate LLMs' logical abstraction abilities from the lexical content of a well-known puzzle; (3) we show that our obfuscation techniques can mitigate data contamination from training; (4) we provide a formally motivated reference point for puzzle difficulty; (5) we find that LLMs struggle primarily with larger puzzles requiring longer reasoning chains; (6) we show that scaling models up to 70B parameters yields only limited gains in performance; and (7) we demonstrate, through a case study, that models can perform only a limited number of consecutive inferential reasoning steps.

2 Related Work

Puzzle solving Formal logic and early machine learning started tackling automatic puzzle solving including CSPs like maths puzzles, Sudoku, Crossword puzzles or grid-puzzles (Wos, 1988; Shazeer et al., 1999; Goldberg et al., 2002; Jones et al.; Shapiro, 2011; Chesani et al., 2017; Valentine and Davis, 1987; Salavati et al., 2009).

Pre-LLM approaches in NLP tried solving grid puzzles with controlled or hybrid architectures (Schwitter, 2013; Jabrayilzade and Tekir, 2020). As LLMs' performance increased, a range of works explored their reasoning abilities with puzzles like Minesweeper (Li et al., 2024b). Results suggest very basic reasoning abilities and issues in performing consistent, coherent longer reasoning chains.

Live-Bench (White et al., 2024) includes Zebra-Puzzles in a Q&A format asking for one single item. Tyagi et al. (2024) develop a LLM grid-puzzle benchmark with 5 sizes and 3 difficulty levels, analysing reasoning chains in depth. They conclude that mistakes are more frequent in the second half of the reasoning process than in the initial part. Lin et al. (2024) provides puzzles in 36 sizes and rank puzzle difficulty based on puzzle size. They find that models between 7 and 10 billion parameters struggle with hard puzzles and show a low accuracy for smaller puzzles. Despite the advancements in these works, they offer only a few sizes and fail to provide fine-grained difficulty levels. Further, potential effects of residual lexical clues from the training data are ignored. This work aims to address this gap by offering a larger variety of sizes and difficulty levels based on a formal perspective and providing systematic obfuscation techniques to mitigate effects for puzzles that might be part of LLM training data.

Task obfuscation Obfuscation techniques originate from adversarial training of image classification models to test the robustness against variations in the input data (Xu et al., 2018; Pezzementi et al., 2018; Woods et al., 2019; Dong et al., 2020; Zhang et al., 2020; Badjie et al., 2024). In language modelling, a wide range of techniques was developed to generate manipulated language input and test the robustness of language models (Xu et al., 2018; Wang et al., 2019; Zhang et al., 2019; Niewinski et al., 2019; Liu et al., 2020).

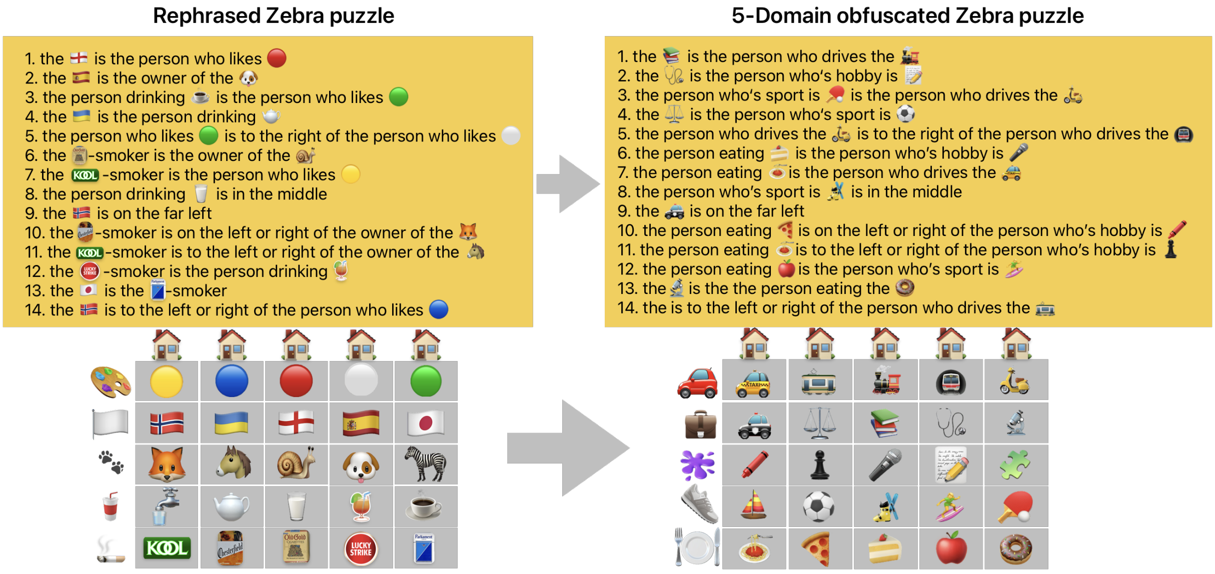
Further, obfuscation is used in adversarial studies to assess LLM-weaknesses (Schwinn et al., 2023; Zou et al., 2024; Kumar, 2024). Steindl et al. (2024); Liu et al. (2024); Yao et al. (2024); Ahmed and Angel Arul Jothi (2024) use this to evaluate the robustness of LLMs against jailbreak attacks that attempt to circumvent safeguarding mechanisms .

Recently, many other planning or reasoning tasks adopted obfuscation methods to probe the reasoning abilities of LLMs. Valmeekam et al. (2023) manipulate the Blocks-World planning task systematically replacing the actions required in the planning process. Nezhurina et al. (2024) propose the Alice in Wonderland (AIW) task, manipulating lexical items in a simple logic puzzle. Xie et al. (2024) use a similar approach on Knights & Knaves puzzles. All works report a breakdown in performance for puzzles from manipulated (training) data.

3 Grid Puzzles as CSPs

A CSP perspective on grid-puzzles helps to determine the difficulty of puzzles. By formulating the rules of the puzzle in pseudo-logical form, it can be determined how many variables are linked through a single clue. In addition, this angle helps to explore how tightly the clues condition the position of items in the grid. These properties are crucial for formal approaches as they influence how many options there are for placing an item in the grid after interpreting a clue. They also determine if a clue can be interpreted on its own or requires combined information from other clues.

This work uses this property to offer a fine-



Rephrased Zebra puzzle

1. the 🏴󠁧󠁢󠁥󠁮󠁧󠁿 is the person who likes 🔴
2. the 🇪🇸 is the owner of the 🐶
3. the person drinking ☕ is the person who likes 🟢
4. the 🇺🇦 is the person drinking 🫖
5. the person who likes 🟢 is to the right of the person who likes ⚪
6. the Old Gold-smoker is the owner of the 🐌
7. the KOOL -smoker is the person who likes 🟡
8. the person drinking 🥛 is in the middle
9. the 🇳🇴 is on the far left
10. the Chesterfield-smoker is on the left or right of the owner of the 🦊
11. the KOOL-smoker is to the left or right of the owner of the 🐴
12. the LUCKY STRIKE -smoker is the person drinking 🍹
13. the 🇯🇵 is the Parliament-smoker
14. the 🇳🇴 is to the left or right of the person who likes 🔵

	🏠	🏠	🏠	🏠	🏠
🎨	🟡	🔵	🔴	⚪	🟢
🏳️	🇳🇴	🇺🇦	🏴󠁧󠁢󠁥󠁮󠁧󠁿	🇪🇸	🇯🇵
🐾	🦊	🐴	🐌	🐶	🦓
🥤	🚰	🫖	🥛	🍹	☕
🚬	KOOL	Chesterfield	Old Gold	LUCKY STRIKE	Parliament

5-Domain obfuscated Zebra puzzle

1. the 📚 is the person who drives the 🚂
2. the 🩺 is the person who's hobby is 📝
3. the person who's sport is 🏓 is the person who drives the 🛵
4. the ⚖️ is the person who's sport is ⚽
5. the person who drives the 🛵 is to the right of the person who drives the 🚇
6. the person eating 🍰 is the person who's hobby is 🎤
7. the person eating 🍝is the person who drives the 🚕
8. the person who's sport is ⛷️ is in the middle
9. the 🚓 is on the far left
10. the person eating 🍕 is on the left or right of the person who's hobby is 🖍️
11. the person eating 🍝is to the left or right of the person who's hobby is ♟️
12. the person eating 🍎is the person who's sport is 🏄
13. the🔬 is the the person eating the 🍩
14. the is to the left or right of the person who drives the 🚃

	🏠	🏠	🏠	🏠	🏠
🚗	🚕	🚃	🚂	🚇	🛵
💼	🚓	⚖️	📚	🩺	🔬
💜	🖍️	♟️	🎤	📝	🧩
👟	⛵	⚽	⛷️	🏄	🏓
🍽️	🍝	🍕	🍰	🍎	🍩

Figure 3: Example of the obfuscation process for the "Zebra" puzzle. On the right is the rephrased "Zebra" puzzle (see Figure 1 for the original phrasing). The items in the puzzle and solution grid are then systematically replaced in the 5-domain obfuscated puzzle so that the logical structure of deductions leading to the solution grid remains intact.

grained range of difficulty levels based on the properties of the clues of the riddle. The lowest level of difficulty uses the commonly employed clue types which link two variables with determined positions. With increasing difficulty, the clues link up to 3 variables in the grid and become less restricting, with the highest levels of difficulty including less strict position clues (A is somewhere to the right of B) and optional conditions (A 'or' B). The difficulty of correctly interpreting the advanced types of clues rises with the size of the grid and especially with the number of items per domain.

4 Dataset creation

Mystery-Zebra contains 4,290 puzzles in two parts: Part one (40 puzzles) contains the original "Einstein" and "Zebra" puzzles with 19 each, preserving the logical structure, while varying the lexical content. Part two includes 4,250 newly generated puzzles not found in any available corpus.[4]

4.1 Manipulated Puzzles

Part one of the benchmark includes the original 'Einstein"and the "Zebra"[5] puzzle (hereafter canonical puzzles; see App. B for the original puzzles) These puzzles are solved in 5×5 grids and their solutions can be found on various webpages, making it likely that they are part of LLM training data.

For each of the canonical puzzles, 19 obfuscated variants are created to test the ability to abstract from the lexical content to the logical structure. We opt for minimal manipulations in order to showcase the effect of minor changes on the model performance. We keep the symbolic structure of the canonical puzzles and vary the lexical representation of variables in the puzzle. in order

We obfuscate the puzzles in a 3-step process. (1) the puzzles are translated to a symbolic representation. (2) optionally, the content of the solution grid and corresponding clues is changed. This step uses the obfuscation strategies detailed below. (3) the puzzle clues are rephrased in a uniform way so that the clues are grammatically valid structures (see App. C for details). Mystery Zebra contains 3 versions per obfuscation strategy.

Rephrasing The canonical puzzles both vary the verb-phrases used for the domains. To allow automatic obfuscation, we adopt uniform verb-phrases within a domain of items (see App. C). In order to test the effect of this, we apply the uniform phrasing to the original puzzles, skipping step (2).

***n-domain* obfuscation** This strategy replaces in step (2) n domains and their corresponding items in the solution grid and clues (e.g. *beverage*→*hobby*). As n increases, the lexical difference to the original puzzle increases. Figure 3 illustrates the dif-

[4] The benchmark is released under BY-NC-SA 4.0 and available here: https://huggingface.co/datasets/arg-tech/MysteryZebra

[5] We use the more commonly known version where the green house is to the right of the ivory house.

ference between the rephrased and the 5-domain obfuscated version of the "Zebra" puzzle.

***in domain* obfuscation** Here, step (2) replaces all items of the solution grid with different items of the same domain, thus keeping the original domains intact (e.g. *yellow*→*azure* in *color*). This creates a smaller semantic difference to the original, also leaving the corresponding verb-phrases unchanged.

4.2 Generated Puzzles

The puzzles for part two are created systematically using a puzzle generator[6]. The difficulty of a puzzle is determined by the size of the solution-grid and the types of clues. The dataset includes puzzles in sizes between 1×2 and 7×7. The generator provides transparent fine-grained metrics for the formal difficulty of a puzzle. We use the first 12 difficulty levels provided. Each level is associated with specific formal clue-structures (e.g. $A = B$; $A = B \vee A = C$) which are added to a growing pool from which the generator can choose (see App. A for all clue-structures). The difficulty of a structure is determined by how restrictive it would be as the only clue. With increasing difficulty, the number of possible solution grids matching the clue increases, meaning that the amount of information contained within the clue decreases.

At least one clue must have a structure from target level (e.g. all Level 6 puzzles have at least one clue in level 6 clue-structure). Since some structures require larger grids (e.g. Level 2: *A is between B and C*), not all grid sizes are available at all levels. Table 5 in App. A contains information on the level size combinations. We create 10 puzzles for each level size combination.

5 Methodology

5.1 Problem formulation

In our experiments, the models predict the solution grid G given a set of clues $C = \{c_1 \dots c_n\}$ and the empty target grid G with d domains and n items per domain. The corresponding prompt is in App. D.

5.2 Grading solutions

We require the model to provide the full solution grid. Because only one misplaced item can lead to a wrong answer, we refrain from a classical accuracy measure. Instead, a point scale (one point per correctly filled cell) is applied, which can be converted to an accuracy score reflecting the percentage of correct cells in the solution grid. This offers a more differentiated and comparable measure across sizes. In the following, this is called grid completion accuracy.

As a baseline, we use a random, uniformly distributed assignment of each item to a random position within the correct domain. This simulates a model that ignores the clues but still places each item exactly once and within the correct domain. Te expected value for correctly guessed cells in a grid with d domains and n items per domain is $E = d$. Transferred to the measure of grid accuracy, this is: $E = \frac{d}{n \cdot d} = \frac{1}{n}$

6 Experimental setup

We assess open and closed-weight models for different puzzle sizes, clue difficulties and obfuscation techniques (see App. E for the hyper-parameters used). The first experiment includes Llama 3.1-8B, 3.3-70B, 3.3-70B (Touvron et al., 2023), Mistral-7B (Jiang et al., 2023), Qwen 2.5-7B, 72B[7], GPT 4o-mini, GPT 4o (Achiam et al., 2023), and R1 (DeepSeek-AI et al., 2025). The second experiment uses only open-weight models between 7 and 72B parameters due to its higher cost. In a pre-study involving three example puzzles from the second part of the benchmark, we determined the models' reactions to the prompt in App. D. No additional prompting strategies are used as Tyagi et al. (2024) report no improvement in puzzle solving over a variety of prompting strategies.

Experiment 1 The first experiment evaluates the open-weight and GPT models on the canonical puzzles and their obfuscated versions. For the original and rephrased versions, three solutions are predicted with different random initialisations. The n*-Domain* and *in domain* obfuscations three different obfuscated versions are tested.

Experiment 2 In this experiment, the models predict a solution for 10 different puzzles from the second part of the benchmark of each level size combination. As we expect no effect from training data in this case, no obfuscated versions are tested.

7 Results

The evaluation of the grid-format experiments presupposes a table format with the domains as rows and the items in the domains as columns. The

[6] https://github.com/quint-t/Puzzle-Generator-and-Solver/tree/master

[7] https://github.com/QwenLM/Qwen

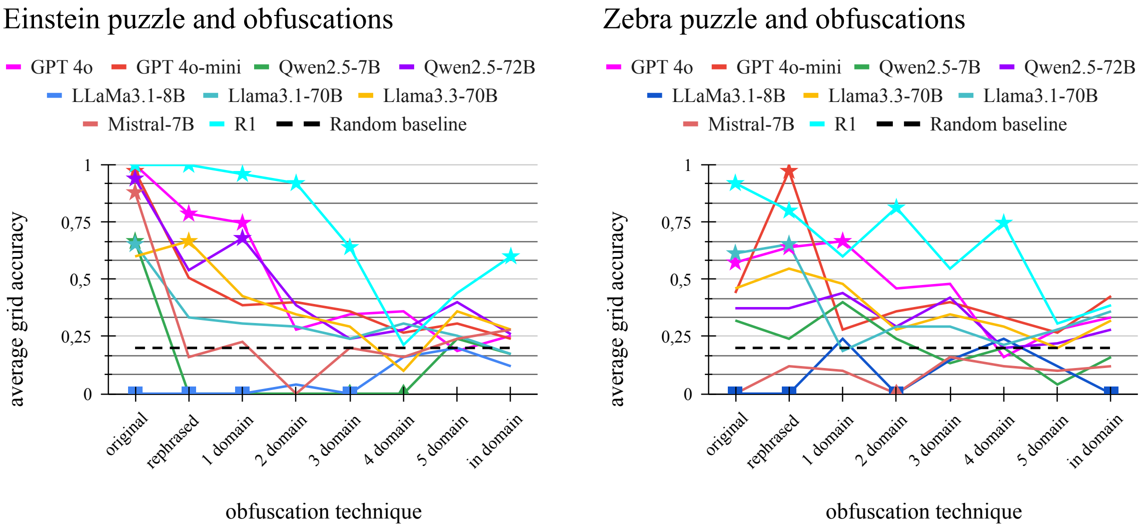


Figure 4: Average grid accuracy on the puzzles and their obfuscated versions of Deepseek R1, GPT 4o, GPT 4o-mini, Llama 3.1-8B, 3.1-70B, 3.3-70B, Qwen 2.5-7B, 2.5-70B, and Mistral 7B. The random baseline for 5 × 5-puzzles of 0.2 is indicated by the dashed black line. Star-shaped points indicate the 1% significance interval. Data points without a correctly formatted prediction are squares or triangles.

Q&A format is evaluated based on the target format defined in the prompt and accounts for minor variations through additional white spaces. All generated answers that do not contain the target formats are excluded from the analysis (see App. F). The significance intervals provided are calculated as described in App. G.

Experiment 1 Figure 4 reports the results of experiment 1. The performance of the GPT models on the "Einstein" puzzle (left) is nearly perfect on the original puzzle and degrades with the degree of obfuscation, being close to the random baseline (0.2) for *in domain* and *5 domain* obfuscation. The performance of R1 is high for up to 2-domain obfuscations but degrades afterwards. For the *in domain* obfuscation, the performance of R1 is significantly above the random baseline. The performance of open-weight models is below the GPT models on the original puzzle. Qwen 2.5-72B is an exception and performs more similarly to the GPT models.

The results for the "Zebra" puzzle show an overall lower performance with the maximum achieved grid completion accuracy being 97% by 4o-mini on the rephrased version. All models except for R1 and Qwen 2.5-7B perform worse on the puzzle as found on the internet compared to the rephrased version. The performance still decreases as the level of obfuscation increases. No model except for R1 outperforms the random baseline significantly after the 2-domain obfuscations. R1's performance is more unstable, bouncing between performances significantly above the random baseline and insignificant performances. Nevertheless, a clear tendency for degrading performance is visible. Mistral and Llama show a poor performance close to or below random baseline across the board of versions. These two models struggle significantly with producing the target format. The larger open-source models achieve a performance more similar to the GPT models with Llama 3.1-70B, even surpassing them for the original and rephrased version.

	Qwen2.5		Llama 3.1		Llama 3.3	Mistral
Md / Lv	7B	72B	8B	70B	70B	7B
Lv1	0.38	0.29	0.30	0.51*	0.57**	0.20
Lv2	0.29	0.23	0.21	0.38	0.43*	0.15
Lv3	0.35	0.27	0.25	0.44*	0.51*	0.18
Lv4	0.37	0.28	0.29	0.45*	0.51*	0.21
Lv5	0.36	0.27	0.30	0.46*	0.55**	0.22
Lv6	0.33	0.27	0.28	0.41	0.48*	0.20
Lv7	0.29	0.22	0.22	0.38	0.43*	0.15
Lv8	0.32	0.28	0.27	0.40	0.47*	0.21
Lv9	0.31	0.26	0.23	0.38	0.43	0.20
Lv10	0.23	0.20	0.21	0.30	0.33	0.15
Lv11	0.23	0.21	0.22	0.28	0.35	0.15
Lv12	0.24	0.19	0.20	0.28	0.31	0.14

Table 1: Average table completion accuracy (for correctly formatted predictions) per level (Lv) for Model (Md) Qwen 2.5-7B, 2.5-72B, Llama 3.1-8B, 3.1-70B, 3.3-70B, and Mistral-7B. With respect to a random baseline, * indicates $p \leq 0.05$ and ** indicates $p \leq 0.01$.

Experiment 2 Table 1 focuses on the results per difficulty level, while averaging over grid sizes and

model	d n	1	2	3	4	5	6	7	baseline
Qwen2-7B	2	0.75**	0.59	0.62**	0.51	0.54	0.54	0.54	0.50
	3	0.41	0.43**	0.41*	0.41**	0.39*	0.37	0.37	0.33
	4	0.35*	0.30	0.29	0.30*	0.26	0.27	0.27	0.25
	5	0.28*	0.23	0.24	0.23	0.20	0.21	0.21	0.20
	6	0.22	0.19	0.20	0.18	0.17	0.16	0.16	0.17
	7	0.20	0.16	0.16	0.16	0.13	0.13	0.13	0.14
Qwen2-72B	2	0.63	0.55	0.57	0.59*	0.43	0.53	0.50	0.50
	3	0.41	0.34	0.35	0.34	0.32	0.34	0.37	0.33
	4	0.27	0.30	0.24	0.23	0.26	0.22	0.24	0.25
	5	0.25	0.21	0.19	0.18	0.20	0.19	0.18	0.20
	6	0.20	0.15	0.17	0.15	0.16	0.16	0.14	0.17
	7	0.15	0.13	0.12	0.14	0.14	0.13	0.13	0.14
Llama3.1-8B	2	0.68**	0.64*	0.68**	0.65**	0.56	0.58*	0.55	0.50
	3	0.44	0.36	0.40	0.39	0.38	0.40*	0.39*	0.33
	4	0.31	0.26	0.26	0.27	0.27	0.28	0.26	0.25
	5	0.28	0.15	0.21	0.24	0.24	0.22	0.21	0.20
	6	0.14	0.08	0.19	0.19	0.19	0.19	0.19	0.17
	7	0.18	0.06	0.15	0.16	0.15	0.15	0.14	0.14
Llama3.1-70B	2	0.95**	0.90**	0.84**	0.82**	0.76**	0.73**	0.74**	0.50
	3	0.73**	0.62**	0.57**	0.54**	0.49**	0.50**	0.48**	0.33
	4	0.54**	0.52**	0.40**	0.35**	0.36**	0.34**	0.33**	0.25
	5	0.47**	0.41**	0.31**	0.31**	0.30**	0.27**	0.27**	0.20
	6	0.39**	0.31**	0.25**	0.25**	0.23**	0.23**	0.22**	0.17
	7	0.35**	0.25**	0.23**	0.20**	0.21**	0.19**	0.18	0.14
Llama3.3-70B	2	0.96**	0.94**	0.93**	0.92**	0.89**	0.85**	0.85**	0.50
	3	0.90**	0.70**	0.65**	0.58**	0.54**	0.54**	0.57**	0.33
	4	0.68**	0.58**	0.47**	0.42**	0.41**	0.39**	0.37**	0.25
	5	0.71**	0.47**	0.38**	0.36**	0.34**	0.29**	0.29**	0.20
	6	0.55**	0.36**	0.32**	0.31**	0.26**	0.25**	0.24**	0.17
	7	0.56**	0.33**	0.27**	0.25**	0.22**	0.21**	0.19*	0.14
Mistral-7B	2	0.77	0.49	0.52	0.50	0.39*	0.44	0.47	0.50
	3	0.43	0.28	0.29	0.32	0.28	0.25*	0.23**	0.33
	4	0.29	0.15	0.25	0.20	0.24	0.18*	0.18*	0.25
	5	0.27	0.14	0.16	0.16	0.14*	0.12*	0.16	0.20
	6	0.22	0.14	0.13	0.15	0.12	0.12	0.14	0.17
	7	0.17	0.11	0.09	0.12	0.13	0.12	0.12	0.14

Table 2: Average table completion accuracy (for correctly formatted predictions) for each grid size (with d (domains) × n (items per domain)) for Qwen 2.5-7B, 72B, Llama 3.1-8B, 70B, Llama 3.3-70B, and Mistral-7B. With respect to a random baseline, * indicates $p \leq 0.05$ and ** indicates $p \leq 0.01$.

shows a slight decrease in performance for higher difficulty levels. Level 2 and 7 stand out with a relatively low performance.

Table 2 shows the results for each grid size when averaged over the difficulty levels. Here, a strong variation in performance across different grid sizes is visible. All models perform best on 1×2-grids. For the 7B models, as d increases, the performance drops to values around the random baseline, while the larger equivalents of Llama present a more stable performance for $n = 2$. Qwen 2.5 72B is an outlier, under-performing its smaller equivalent. Nevertheless, even the performance of best model in this experiment (Llama 3.3-70B) decreases as the grid size increases and reaches values close to the random baseline for larger puzzles.

To reveal underlying effects between the performance and model size, grid size, and level, a correlation analysis is conducted. The results in Table 3 suggest a weak positive effect between model size. In general, the number of items per domain (n) has the strongest negative effect on the performance, while all other factors have medium to low negative effects on the performance.

	n	d	model size	level
all models	-0.70	-0.22	0.25	-0.17
Llama 3.1-8B	-0.83	-0.03	–	-0.10
Llama 3.1-70B	-0.82	-0.29	–	-0.26
Llama 3.3-70B	-0.82	-0.29	–	-0.26
Qwen 2.5-7B	-0.75	-0.38	–	-0.21
Qwen 2.5-72B	-0.82	-0.11	–	-0.15
Mistral-7B	-0.71	-0.19	–	-0.10

Table 3: Correlation analysis of grid accuracy by n (items per domain), d (number of domains), model size (only across all models), difficulty level. Red reflects negative effects, green indicates positive effects. Saturation indicates effect strength.

7.1 Error analysis

To gain a deeper insight into the reasoning chains of the models, a manual qualitative error analysis is conducted. For open weight models, 50 randomly selected, incorrectly solved puzzles up to a size of 5×5 are assessed. For the GPT models and R1, 20 incorrectly solved answers are considered. This analysis also includes outputs that did not adhere to the target format. A list of the error sources and their distribution can be found in App. H.

All models frequently produce pseudo-reasoning (41%); i.e. instances where linguistic markers of reasoning are used to combine premises and conclusions that make no sense. Further, the free-text often contradicts the produced solution grid (37%), while no model is able to detect and correct errors in the predicted solution grid. The only exceptions are Qwen 2.5-7B, Mistral-7B and R1, the former two detect errors in 2% and 4% of the analysed outputs, incorrectly concluding that the puzzle was unsolvable. R1 however, claims to fix detected errors, but fails to accomplish this.

Further, smaller 7-8B parameter models produce the highest amount of (partially) empty or malformed solution grids (44%). This effect is strongest in Llama 3.1-8B (52%) and Mistral-7B (58%). Another prevalent source of errors in these models are contradictions between the free-text reasoning and the solution grid (43%).

70-72B parameter models show a positioning bias that prefers to place items in the leftmost free position (35%) and place items twice in the solution grid (31%). Further, these models ignore clues in the produced free-text reasoning more often than others (17% vs. 0.29%). These models also struggle with clues that contain a negation (13%). Qwen 2.5-72B is the only model that produces grids without free-text reasoning (82%).

The GPT models show the highest rate of pseudo-reasoning (78%) and tend to misinterpret placement clues (28%). This especially includes cases where clues are interpreted more loosely than intended (e.g. “to left” or “to the right” is interpreted as “somewhere to the left/right”). Further, these models are more prone to placing before the drawn conclusions are sufficient for a definitive positioning (23%) and also tend to place items in the leftmost available position (30%).

In *n-domain* obfuscated puzzles, R1 solves not obfuscated rows perfectly, while making mistakes in obfuscated rows. This suggests a strong influence from training material. Further, this model produces the longest free-text reasoning among all tested models, resulting in self-repetitions (75%) and as a result overwrites formerly correct predictions (35%). This model has as well a bias for positioning items in the leftmost position (35%).

7.2 Case study

In addition, we conduct a case study on the canonical puzzles to analyse model-specific behaviours in different reasoning steps of the solution. This is done in a Q&A setup, which asks for the position P of a specific item I in a row R of the solution grid G given a set of clues $C = \{c_1 \dots c_n\}$. The corresponding prompt is shown in App. D. This format is evaluated based on the accuracy of correct answers. Since this setup focuses on the position within a single domain, the random baseline for a puzzle with n items per domain is: $E = \frac{1}{n}$.

We test the GPT and open-weight models from Experiment 1 on the original, rephrased, *in domain*, and *5-domain* obfuscations of the canonical puzzles to avoid cases where only a part of the puzzle is manipulated. We run 10 predictions for each item with open-weight models and 3 predictions per item with GPT models. The items are ordered by the reasoning depth in the canonical puzzles in accordance with solution guides available on the Internet. App. I.1 details the step-by-step solutions.

A per-item analysis shows that 4o, 4o-mini and Llama 3.3 are mostly consistent in their predictions for the same item (see Tables 16-20 in App. I). Since we test only one *in domain* and one *5-domain* obfuscation, the resulting accuracies are close to 0 or 1. Hence, the random baseline is unsuitable and no significance can be reported.

Figures 5 and 6 in App. I report the accuracy at each reasoning step. In all assessed puzzle variations and models, the first two reasoning steps are performed successfully. For smaller models of 7-8B parameters, the performance degrades afterwards regardless of the obfuscation of the puzzle. On the original an rephrased puzzles, Llama 3.3-70B and Qwen 2.5-72B recover from this drop in performance and reach relatively high performance in the last step. The GPT models keep a relatively high performance across the original grid versions with medium drops in performance from the 3rd reasoning step. On the obfuscated versions, the performance drops also for bigger models from the 3rd step onward. Nevertheless, peaks of medium to high accuracy can be found at specific steps in the reasoning process (4th, 7th, and 8th for “Einstein”; 5th and 7th for “Zebra”).

8 Discussion

In both experiments, the reasoning abilities of smaller models (7-8B) are limited. They are successful in one-step reasoning in $d \times 2$ puzzles. For bigger puzzles, their performance approaches the random baseline. The case study in 7.2 and Lin et al. (2024) support this.

In general, models around 70B parameters outperform the random baseline by a statistically significant but very small margin. This indicates that easier clues are solved but not the harder ones.

The correlation analysis reveals only a modest performance gain when scaling models up to 70B parameters, suggesting diminishing returns at even larger scales (e.g. 400B), especially given the substantial computational and environmental costs.The formal difficulty of clues is also a limited predictor. The observed irregularities in Levels 2 and 7 might be related to clues containing the relation *between* introduced in these levels. Further effects may be invisible due to the correlation of difficult clues with more items per domain (see Table 5, App. A), which is the strongest predictor in this analysis.

The correlation between grid size and performance is supported by the case study, which reveals a decrease in performance for longer reasoning chains in fully obfuscated canonical puzzles. Irregular performance peaks in the case study occur in specific later reasoning steps, where the last 1-2 positions in a row are filled (see App. I.1, which increases the likelihood of successful guessing. This effect is further reinforced as the models with these peaks are more deterministic in their predictions.

The error analysis suggests severe difficulties in detecting contradictions in the produced solution grids for all models, while they produce a lot of pseudo-reasoning and are unable to align free-text reasoning and solution grids.

For the original "Einstein" puzzle, the open-weight models show performances significantly above the random baseline, suggesting that the existence of this puzzle in the training data is beneficial for the performance on this puzzle. For the original "Zebra" puzzle, only Llama and GPT models show this effect. For any obfuscated version, the open-weight models show similar performances to newly generated puzzles of the same size, indicating that the obfuscation has the intended effect.

The closed weight models show nearly perfect performance on the original version of the "Einstein" puzzle. But their performance degrades drastically with the gradual manipulation of the lexical content of the solution grid, finally not outperforming the random baseline. The error analysis suggests a high degree of pseudo-reasoning for these models and a tendency to place items prematurely in the leftmost available position.

The observed decrease in performance of larger models with the increase of obfuscation suggests that they are able to cope with slightly larger variations than smaller models. Nevertheless, this trend suggests that the dependence on lexical items in the input is not overcome in models with more parameters. R1's relatively high performance in the Einstein-*in domain* is composed of a singular trial with high performance and two low performances, signalling high instability.

On the "Zebra" puzzle, some models improve performance on the rephrased compared to the original puzzle. This could be explained by a dominance of passive voice in the original "Zebra" puzzle (see App. B), which the rephrased version turns into active. This is supported by work reporting that LLMs use passive voice in written text less frequently than humans (Reinhart et al., 2024).

9 Conclusion

This paper develops Mystery-Zebra, the first adaptable systematic benchmark to evaluate on multiple levels. By evaluating a range of open and closed weight LLMs, we demonstrate their dependence on the lexical content from training material for the successful solution of complex grid puzzles.

In comparison to the original puzzles, the performance drops significantly for lexically manipulated puzzles and even reaches the random baseline when all noun phrases in the puzzle are altered. The reasoning abilities of smaller LLMs break down with the slightest variation of the lexical context of a puzzle that is part of their training data.

Our obfuscation strategies can easily be transferred to other tasks to mitigate the effect of logic puzzle benchmarks being part of the training data of the LLMs. This will help to preserve the significance of benchmark results in the future.

An extensive analysis of open weight models on the newly generated puzzles in the second part of the benchmark demonstrated that the grid size has a stronger influence on the performance than the clue difficulty or model size.

In the error analysis, models proved to be incapable to detect and correct contradictions in their output. In addition, our detailed case study of reasoning steps in the solution process demonstrated that the assessed models were only able to perform a very limited number of consecutive reasoning steps correctly.

The results from this paper can direct future research in LLM reasoning by pointing out that improvements through traditional fine-tuning are no clear indicator of improvements in the task itself (see Experiment 1). Hybrid solutions, however, might provide a more sustainable direction to successful LLM reasoning.

Limitations

Due to limits in the compute budget, this work only assesses open-weight models up to a size of 72B parameters on the second part of the benchmark. For this reason, the paper can make only limited claims with regard to the influence of very large parameter sizes on the performance on unknown puzzles of various sizes and difficulty levels.

Due to the substantial compute time of Deepseek R1 (404 seconds on average), this model is only evaluated in the first experiment and the error analysis, but not considered for the case-study. This leads to limited insights into the influence of reasoning depth on the performance of this model. Nevertheless, the error analysis suggests a strong dependence of this model on training material. Future work should consider analysing the reasoning abilities of this model in depth.

Further, the benchmark obfuscates only the nouns and related verb-phrases of the clues and leaves the structures referring to the positions in the solution grid untouched. The influence of such manipulations should be investigated in future research.

Finally, the provided explanation for the improvement in performance observed on the rephrased "Zebra" puzzle relies only on a possible linguistic explanation and evidence from free-text writing studies. Future work should consider the influence of passive voice on performance in more detail.

Ethics statement

The authors didn't identify immediate ethical concerns around the work presented in this paper. Nevertheless, any work around the reasoning abilities of Language Models should be considered with care due to the societal impact of inaccurate assumptions on these abilities.

Acknowledgements

The authors would like to thank Jonathan Pampel for the very helpful discussions on the evaluation and significance testing done in this paper. This work was funded in part by Volkswagen Stiftung Foundation under grant 98 543, "Deliberation Laboratory".

References

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

Sadaf Surur Ahmed and J. Angel Arul Jothi. 2024. Jailbreak attacks on large language models and possible defenses: Present status and future possibilities. In *2024 IEEE International Symposium on Technology and Society (ISTAS)*, pages 1–7.

Bakary Badjie, José Cecílio, and Antonio Casimiro. 2024. Adversarial attacks and countermeasures on image classification-based deep learning models in autonomous driving systems: A systematic review. *ACM Computing Surveys*, 57(1):1–52.

Federico Chesani, Paola Mello, and Michela Milano. 2017. Solving mathematical puzzles: A challenging competition for ai. *AI Magazine*, 38(3):83–96.

DeepSeek-AI et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint*.

Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. 2020. Benchmarking adversarial robustness on image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Panagiotis Giadikiaroglou, Maria Lymperaiou, Giorgos Filandrianos, and Giorgos Stamou. 2024. Puzzle solving using reasoning of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11574–11591. Association for Computational Linguistics.

David Goldberg, Christopher Malon, and Marshall Bern. 2002. A global approach to automatic solution of jigsaw puzzles. In *Proceedings of the Eighteenth Annual Symposium on Computational Geometry*, SCG '02, page 82–87, New York, NY, USA. Association for Computing Machinery.

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

Jiaxian Guo, Bo Yang, Paul Yoo, Bill Yuchen Lin, Yusuke Iwasawa, and Yutaka Matsuo. 2023. Suspicion-agent: Playing imperfect information games with theory of mind aware gpt-4. *arXiv preprint*.

Akshat Gupta. 2023. Are chatgpt and gpt-4 good poker players? – a pre-flop analysis. *arXiv preprint*.

Elgun Jabrayilzade and Selma Tekir. 2020. LGPSolver - solving logic grid puzzles automatically. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1118–1123, Online. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

S. K. Jones, P. A. Roach, and S. Perkins. *Construction of Heuristics for a Search-Based Approach to Solving Sudoku*, pages 37–49. Springer London.

Pranjal Kumar. 2024. Adversarial attacks and defenses for large language models (llms): methods, frameworks & challenges. *International Journal of Multimedia Information Retrieval*, 13(3).

Hao Li, Xue Yang, Zhaokai Wang, Xizhou Zhu, Jie Zhou, Yu Qiao, Xiaogang Wang, Hongsheng Li, Lewei Lu, and Jifeng Dai. 2024a. Auto mc-reward: Automated dense reward design with large language models for minecraft. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16426–16435.

Yinghao Li, Haorui Wang, and Chao Zhang. 2024b. Assessing logical puzzle solving in large language models: Insights from a minesweeper case study. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 59–81. Association for Computational Linguistics.

Bill Yuchen Lin, Ronan Le Bras, and Yejin Choi. 2024. Zebralogic: Benchmarking the logical reasoning ability of language models.

Tong Liu, Yingjie Zhang, Zhe Zhao, Yinpeng Dong, Guozhu Meng, and Kai Chen. 2024. Making them ask and answer: Jailbreaking large language models in few queries via disguise and reconstruction. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 4711–4728, Philadelphia, PA. USENIX Association.

Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. 2020. Adversarial training for large neural language models. *arXiv preprint*.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint*.

Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti, and Jenia Jitsev. 2024. Alice in wonderland: Simple tasks showing complete reasoning breakdown in state-of-the-art large language models. *arXiv preprint*.

Piotr Niewinski, Maria Pszona, and Maria Janicka. 2019. GEM: Generative enhanced model for adversarial attacks. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 20–26, Hong Kong, China. Association for Computational Linguistics.

Zachary Pezzementi, Trenton Tabor, Samuel Yim, Jonathan K. Chang, Bill Drozd, David Guttendorf, Michael Wagner, and Philip Koopman. 2018. Putting image manipulations in context: Robustness testing for safe perception. In *2018 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pages 1–8. IEEE.

Alex Reinhart, David West Brown, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, and Gordon Weinberg. 2024. Do llms write like humans? variation in grammatical and rhetorical styles. *arXiv preprint*.

Soumadeep Saha, Sutanoya Chakraborty, Saptarshi Saha, and Utpal Garain. 2024. Language models are crossword solvers. *arXiv preprint*.

Soroor Salavati, Sahar Hajjarzadeh, and Masoud Mazloom. 2009. An optimized method for solving zebra puzzle.

Leo Schwinn, David Dobre, Stephan Günnemann, and Gauthier Gidel. 2023. Adversarial attacks and defenses in large language models: Old and new threats. In *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops*, volume 239 of *Proceedings of Machine Learning Research*, pages 103–117. PMLR.

Rolf Schwitter. 2013. The jobs puzzle: Taking on the challenge via controlled natural language processing. *Theory and Practice of Logic Programming*, 13(4–5):487–501.

Kulin Shah, Nishanth Dikkala, Xin Wang, and Rina Panigrahy. 2024. Causal language modeling can elicit search and reasoning capabilities on logic puzzles. *arXiv preprint*.

Stuart C. Shapiro. 2011. The jobs puzzle: A challenge for logical expressibility and automated reasoning. In *Papers from the AAAI 2011 Spring Conference*.

Noam M. Shazeer, Michael L. Littmann, and Greg A. Keim. 1999. Solving crossword puzzles as probabilistic constraint satisfaction. In *Proceedings of AAAI 1999*.

Sebastian Steindl, Ulrich Schäfer, Bernd Ludwig, and Patrick Levi. 2024. Linguistic obfuscation attacks and large language model uncertainty. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 35–40, St Julians, Malta. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Nemika Tyagi, Mihir Parmar, Mohith Kulkarni, Aswin Rrv, Nisarg Patel, Mutsumi Nakamura, Arindam Mitra, and Chitta Baral. 2024. Step-by-step reasoning to solve grid puzzles: Where do llms falter? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19898–19915. Association for Computational Linguistics.

Mark Valentine and Robert H. Davis. 1987. The automated solution of logic puzzles. *Information Processing Letters*, 24(5):317–324.

Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. 2023. On the planning abilities of large language models - a critical investigation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yuxuan Wan, Wenxuan Wang, Yiliu Yang, Youliang Yuan, Jen-tse Huang, Pinjia He, Wenxiang Jiao, and Michael Lyu. 2024. LogicAsker: Evaluating and improving the logical reasoning ability of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2124–2155, Miami, Florida, USA. Association for Computational Linguistics.

Dilin Wang, Chengyue Gong, and Qiang Liu. 2019. Improving neural language modeling via adversarial training. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6555–6565. PMLR.

Zeyu Wang. 2024. CausalBench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 143–151, Bangkok, Thailand. Association for Computational Linguistics.

Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian (Shawn) Ma, and Yitao Liang. 2023. Describe, explain, plan and select: Interactive planning with llms enables open-world multi-task agents. In *Advances in Neural Information Processing Systems*, volume 36, pages 34153–34189. Curran Associates, Inc.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. 2024. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint*.

Walt Woods, Jack Chen, and Christof Teuscher. 2019. Adversarial explanations for understanding image classification decisions and improved neural network robustness. *Nature Machine Intelligence*, 1(11):508–516.

Larry Wos. 1988. *Automating Reasoning: How to use a computer to help solve problems requiring logical reasoning*, pages 110–137. Springer New York.

Chulin Xie, Yangsibo Huang, Chiyuan Zhang, Da Yu, Xinyun Chen, Bill Yuchen Lin, Bo Li, Badih Ghazi, and Ravi Kumar. 2024. On memorization of large language models in logical reasoning. *arXiv preprint*.

Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrell, and Dawn Song. 2018. Fooling vision and language models despite localization and attention mechanism. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dongyu Yao, Jianshu Zhang, Ian G. Harris, and Marcel Carlsson. 2024. Fuzzllm: A novel and universal fuzzing framework for proactively discovering jailbreak vulnerabilities in large language models. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4485–4489.

Cheng Zhang, Kun Zhang, and Yingzhen Li. 2020. A causal view on robustness of neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 289–301. Curran Associates, Inc.

Huangzhao Zhang, Hao Zhou, Ning Miao, and Lei Li. 2019. Generating fluent adversarial examples for natural languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5569, Florence, Italy. Association for Computational Linguistics.

Jing Zou, Shungeng Zhang, and Meikang Qiu. 2024. *Adversarial Attacks on Large Language Models*, pages 85–96. Springer Nature Singapore.

A Puzzle Levels

Level	Semi-Formal expression	Description
Level 1	A == B	An object that has attribute A has attribute B.
	A is on the left of B	An object with attribute A is next to the left of an object with attribute B (A-B).
	A is on the right of B	An object with attribute A is next to the right of an object with attribute B (B-A).
	A is on the far left	An object with attribute A is on the far left (A-...).
	A is on the far right	An object with attribute A is on the far right (...-A).
	A is in the middle	An object with attribute A is in the middle.
Level 2	A is between B and C	An object with attribute A is between an object with attribute B, and an object with attribute C (any order: B-A-C, C-A-B).
Level 3	A is on the left or right of B	An object with attribute A is next to the left or right of an object with attribute B (A-B or B-A).
	A is on the far left or far right	An object with attribute A is on the far left or far right. (A-... or ...-A)
Level 4	A is in an odd position	An object with attribute A is in an odd position (odd positions : 1, 3, 5, ...).
	A is in an even position	An object with attribute A is in an even position (even positions : 2, 4, 6, ...).
Level 5	A is somewhere to the left of B	An object with attribute A is somewhere to the left of an object with attribute B (any number of intermediates, including 0 : A-...-B).
	A is somewhere to the right of B	An object with attribute A is somewhere to the right of an object with attribute B (any number of intermediates, including 0 : B-...-A).
Level 6	A != B	An object that has attribute A does not have attribute B.
Level 7	A is somewhere between B and C	An object with attribute A is somewhere between an object with attribute B, and an object with attribute C (any number of intermediates, including 0 : B-...-A-...-C, C-...-A-...-B).
Level 8	A is not to the left of B	An object with attribute A is not to the left of an object with attribute B.
	A is not to the right of B	An object with attribute A is not to the right of an object with attribute B.
Level 9	A and B have different parity positions	An object with attribute A and an object with attribute B have different parity positions.
	A and B have the same parity positions	An object with attribute A and an object with attribute B have the same parity positions (positions may be the same or different, but the parity is always the same).
Level 10	A == B or A == C, but not both	An object that has attribute A has attribute B, or an object that has attribute A has attribute C, but not both.
	A == B or B == C, but not both	An object that has attribute A has attribute B, or an object that has attribute B has attribute C, but not both.
Level 11	A == B or A == C or both	An object that has attribute A has attribute B, or an object that has attribute A has attribute C, or both.
	A == B or B == C or both	An object that has attribute A has attribute B, or an object that has attribute B has attribute C, or both.
Level 12	A != B or A != C or both	An object that has attribute A has not attribute B, or an object that has attribute A has not attribute C, or both.
	A != B or B != C or both	An object that has attribute A has not attribute B, or an object that has attribute B has not attribute C, or both.

Table 4: Puzzle clue types per level taken from https://github.com/quint-t/Puzzle-Generator-and-Solver/tree/master

Level	Puzzle sizes (domains × items per domain)	number
Level 1	all sizes from 1 × 2 to 7 × 7	420
Level 2	sizes with at least 3 items per domain	350
Level 3	all sizes from 1 × 2 to 7 × 7	420
Level 4	all sizes from 1 × 2 to 7 × 7	420
Level 5	all sizes from 1 × 2 to 7 × 7	420
Level 6	sizes with at least 2 domains	360
Level 7	sizes with at least 3 items per domain	350
Level 8	all sizes from 1 × 2 to 7 × 7	420
Level 9	sizes with at least 2 domains	360
Level 10	sizes with at least 2 domains and 3 items per domain	250
Level 11	sizes with at least 2 domains and 3 items per domain	250
Level 12	sizes with at least 2 domains and 3 items per domain	250

Table 5: Combinations of difficulty levels and sizes available in the benchmark of newly generated puzzles, since some puzzle sizes are not compatible with some types of clues (e.g. Level 2: A is between B and C.)

B Original Puzzle

Einstein Puzzle	Zebra Puzzle
There are 5 houses (in a row) painted 5 different colors: Blue, Green, Red, White and Yellow. In each house there lives a person of a different nationality: Brit, Dane, German, Norwegian or Swede. These 5 owners each drink a certain beverage: Beer, Coffee, Milk, Tea or Water. They also smoke a certain brand of cigar: Bluemaster, Dunhill, Pall Mall, Prince or Blend. Additionally, they also keep a certain type of pet: Cats, Birds, Dogs, Fish or Horses. The owners DO NOT have the same pet, smoke the same brand of cigar or drink the same beverage.	There are five different-colored houses: red, green, ivory, yellow, blue There live five resident of a different nationality: english, spanish, ukranian, norwegian, japanese Each resident owns a different pet: dog, fox, zebra, horse, snails Each one prefers a different drink: coffee, tea, milk, orange-juice, water Each one smokes a different brand of cigarettes: old-gold, kools, chesterfields, lucky-strike, parliaments
1. The Brit lives in a Red house. 2. The Swede keeps Dogs as pets. 3. The Dane drinks Tea. 4. The Green house is on the left of the White house. 5. The Green house owner drinks Coffee. 6. The owner who smokes Pall Mall rears Birds. 7. The owner of the Yellow house smokes Dunhill. 8. The owner living in the center house drinks Milk. 9. The Norwegian lives in the first house. 10. The owner who smokes Blend lives next to the one who keeps Cats. 11. The owner who keeps horses lives next to the man who smokes Dunhill. 12. The owner who smokes Bluemaster drinks Beer. 13. The German smokes Prince. 14. The Norwegian lives next to the Blue house. 15. The owner who smokes Blend has a neighbor who drinks Water.	1. There are five houses. 2. The Englishman lives in the red house. 3. The Spaniard owns the dog. 4. Coffee is drunk in the green house. 5. The Ukrainian drinks tea. 6. The green house is immediately to the right of the ivory house. 7. The Old Gold smoker owns snails. 8. Kools are smoked in the yellow house. 9. Milk is drunk in the middle house. 10. The Norwegian lives in the first house. 11. The man who smokes Chesterfields lives in the house next to the man with the fox. 12. Kools are smoked in the house next to the house where the horse is kept. 13. The Lucky Strike smoker drinks orange juice. 14. The Japanese smokes Parliaments. 15. The Norwegian lives next to the blue house. Now, who drinks water? Who owns the zebra?

Table 6: Original “Einstein” and “Zebra” puzzle

s

Domain	Phrasing in obfuscated / generator puzzles	Phrasing in canonical Einstein	Phrasing in canonical Zebra
Color	the person who likes [color]	[color] house	[color] house
Beverage	the person drinking [beverage]	drinks [beverage]	[beverage] is drunk / driniks [beverage]
Pet	the owner of the [pet]	keeps [pet] as pets / rears [pet] / keeps [pet] /	owns the [pet] / owns [pet] / the [pet] is kept / the man with the [pet]
Nationality	the [nationality]	the [nationality]	the [nationality]
Cigar	the [cigar]-smoker	the owner|man who smokes [cigar] / smokes [cigar]	[cigar] are smoked / the [cigar] smoker / smokes [cigar]
Food	the person eating [food]	–	–
House	the [house] house	–	–
Job	the [job]	–	–
Transport	the person driving the [transport name]	–	–
Music-Genre	the fan of [music-genre]	–	–
Movie-Genre	the person watching [movie-genre]	–	–
Sport	the person who’s sport is [sport]	–	–
Hobby	the person who’s hobby is [hobby]	–	–
Flower	the person who grows [flower]	–	–
Birthday	the person who’s birthday is in [birthday month]	–	–
Game	the person playing [game]	–	–

Table 7: Phrasing patterns used in the obfuscated versions of the canonical puzzles and the generator-puzzles in the second part of Mystery-Zebra. For reproducibility of the rephrasing obfuscation, the equivalents in the canonical puzzles are detailed in the two rightmost columns.

C Obfuscation Details

To facilitate the obfuscation process, the canonical puzzles are first translated to the same pseudo-logical language as used by the puzzle-generator, where items are stated as `domain:item`, equivalence is expressed as "=" and negation is expressed as "!=". Any plural forms are replaced with their singular equivalents for this step.
Once the obfuscation is complete, the items are translated back to natural language using the patterns in Table 7. Tables 9 and 10 show the full rephrased versions in comparison to the original phrasing. The same phrasing patterns are applied as well to translate the puzzles from the generator to natural language input. For simplicity, only one phrasing pattern is used for each domain. The pseudo-logical formulas from the generator are translated to natural language according to Table 8.

pseudo-logical sign	phrasing in obfuscated / generator-puzzles
!=	is not
==	is

Table 8: Translation patterns for the pseudo-logical structures created by the puzzle generator and used to obfuscate canonical puzzles.

original phrasing	rephrased
1. The Brit lives in a Red house.	1. The british is the person who likes red
2. The Swede keeps Dogs as pets.	2. The swedish is the owner of the dog
3. The Dane drinks Tea.	3. the danish is the person drinking tea
4. The Green house is on the left of the White house.	4. The person who likes green is on the left of the person who likes white
5. The Green house owner drinks Coffee.	5. The person who likes green is the person drinking coffee
6. The owner who smokes Pall Mall rears Birds.	6. The pall-mall-smoker is the owner of the bird
7. The owner of the Yellow house smokes Dunhill.	7. The person who likes yellow is the dunhill-smoker
8. The owner living in the center house drinks Milk.	8. The person drinking milk is in the middle
9. The Norwegian lives in the first house.	9. The norwegian is on the far left
10. The owner who smokes Blend lives next to the one who keeps Cats.	10. The blend-smoker is on the left or right of the owner of the cat
11. The owner who keeps horses lives next to the man who smokes Dunhill.	11. The owner of the horses is on the left or right of the dunhill-smoker
12. The owner who smokes Bluemaster drinks Beer.	12. The bluemaster-smoker is the person drinking beer
13. The German smokes Prince.	13. The german is the prince-smoker
14. The Norwegian lives next to the Blue house.	14. The norwegian is on the left or right of the person who likes blue
15. The owner who smokes Blend has a neighbor who drinks Water.	15. The blend-smoker is on the left or right of the person drinking water

Table 9: Original "Einstein" puzzle in comparison to the *rephrased* obfuscation.

original phrasing	rephrased
1. There are five houses.	
2. The Englishman lives in the red house.	1. the english is the person who likes red
3. The Spaniard owns the dog.	2. the spanish is the owner of the dog
4. Coffee is drunk in the green house.	3. the person drinking coffee is the person who likes green
5. The Ukrainian drinks tea.	4. the ukrainian is the person drinking tea
6. The green house is immediately to the right of the ivory house (to your right as you stand facing the row of five houses).	5. the person who likes green is to the right of the person who likes ivory
7. The Old Gold smoker owns snails.	6. the old-gold-smoker is the owner of the snails
8. Kools are smoked in the yellow house.	7. the kools-smoker is the person who likes yellow
9. Milk is drunk in the middle house.	8. the person drinking milk is in the middle
10. The Norwegian lives in the first house.	9. the norwegian is on the far left
11. The man who smokes Chesterfields lives in the house next to the man with the fox.	10. the chesterfield-smoker is on the left or right of the owner of the fox
12. Kools are smoked in a house next to the house where the horse is kept.	11. the kools-smoker is to the left or right of the owner of the horse
13. The Lucky Strike smoker drinks orange juice.	12. the lucky-strike-smoker is the person drinking orange-juice
14. The Japanese smokes Parliaments.	13. the japanese is the parliament-smoker
15. The Norwegian lives next to the blue house.	14. the norwegian is to the left or right of the person who likes blue

Table 10: Original "Zebra" puzzle in comparison to the *rephrased* obfuscation.

D Model Prompting

This work uses a basic prompting strategy since the work by Tyagi et al. (2024) suggests that elaborate prompting strategies like self-consistency have no beneficial effects on the grid-puzzle solving abilities of LLMs.

Grid-setup The grid-setup prompt specifies an empty solution grid to improve the consistency in formatting in the outputs.

```
> Please solve the following logic
puzzle in the following table:

[EMPTY SOLUTION GRID]

Puzzle:
[PUZZLE CLUES].

Think step by step when solving the
puzzle. Please put '############'
around the final solution table.
```

Q&A-setup The Q&A-setup prompts the model first to solve the puzzle and then to indicate the position of a specific item in the grid in the form of *item:Num*.

```
> Solve the following logic puzzle:
```

```
[PUZZLE CLUES]

After solving tell me where is **[TARGET
ITEM]**. Give the answer in the format
**[TARGET ITEM]:Num**.
```

E Hyper-Parameters

The hyper parameters of the models were left at their default if not mentioned explicitly.

model	hyper-parameter	setting
LLama 3.1-8B	max_new_tokens	10000
	temperature	0.1
	top_p	0.9
LLama 3.1-70B	max_new_tokens	10000
	temperature	0.1
	top_p	0.9
LLama 3.3-70B	max_new_tokens	10000
	temperature	0.1
	top_p	0.9
Qwen 2.5-7B	max_new_tokens	10000
Qwen 2.5-72B	max_new_tokens	10000
Mistral-7B	max_new_tokens	10000

Table 11: Hyper-parameters for open weight models

F Pre-processing

Grid-setup The open-ended generations of the models were pre-processed according to this requirement. Since this study focuses on the logic capabilities and not on formatting abilities non-adhering answers were excluded from the evaluation process. This filtering was done checking for two criteria: First, the result needs to be a table with boundaries marked by "|". Second, the leftmost column needs to contain all expected domain names for the respective puzzle.[8]. This method does not account for all possible formatting issues (see the error analysis in 7.1), but to in order to ensure reproducibility of this pre-processing, no more fine-grained case distinctions were included. Table 12 reports the percentage of wrong output format for the evaluated models, indicating that especially some of the smaller open-weight models struggle significantly with the target output format.

Model	Excluded from evaluation
R1	0.00%
GPT 4o	0.00%
GPT 4o-mini	0.00%
Qwen 2.5-7B	5.72%
Qwen 2.5-72B	13.55%
Llama 3.1-8B	14.39%
Llama 3.1-70B	5.30%
Llama 3.3-70B	3.80%
Mistral	43.28%

Table 12: Percentage of outputs per model excluded from the evaluation in the grid-setup. The evaluation requires tables with domains in the rows and items in the columns in the model output.

Q&A-setup The prompt for the Q&A specifies the target format used for the pre-processing as *item:Num*. The pre-processing filters the open-ended generations by the models for this format and slight variations containing additional white-space characters, the word *position* or *house*, or one letter in front of the integer indicating the position in the grid. All other versions or formats are excluded from the analysis in Section 7. The percentage of outputs excluded from the analysis is reported in Table 13.

Model	Excluded from evaluation
GPT 4o	0.00%
GPT 4o-mini	0.00%
Llama 3.1 8B	40.75%
Llama 3,1-70B	14.45%
Llama 3,3-70B	3.15%
Mistral 7B	39.85%
Qwen 2.5-72B	13.40%
Qwen 2.5 7B	5.55%

Table 13: Percentage of outputs per model excluded from the evaluation in the Q&A-setup. The evaluation requires the format *item:num*.

G Significance Testing

Grid-setup The grid-setup in Experiments 1 and 3 does not allow to calculate significance values straightforward since the grid-score is not a binomially distributed variable. For this reason, we use the following approximation to calculate significances. The expected values E and Variances V can be calculated for the case $d = 1$ as follows:

$$E = \frac{1}{n}, V = \frac{2n-1}{n^2}$$

By approximating the case for $d > 1$ as a

[8] Any row with an invalid domain name (e.g. a header row) is not considered in the evaluation.

Gauss-Distribution, we get:

$$E = \frac{1}{n}, V = \frac{2n-1}{dn^2}$$

As the experiments are repeated k-times, we get:

$$E = \frac{1}{n}, V = \frac{2n-1}{dkn^2}$$

The average Variance over a grid size or level is then:

$$V\left(\frac{X_1 + X_2 + \ldots + X_j}{j}\right) = \frac{V(X_1) + \ldots + V(X_j)}{j^2}$$

The significance for the measured average normalized grid-score m then calculated in a one-sided T-test:

$$p = 2 \cdot P(X > m) = 2 \cdot P\left(\mathcal{N} > \frac{(m-E)}{\sqrt{V}}\right)$$

H Error Analysis

model	R1	4o	4o-mini	Qwen2.5		Llama3.1		Llama3.3	Mistral	Average
				7B	72B	8B	70B	70B	7B	
E1	0.00%	0.00%	15.00%	2.00%	0.00%	0.00%	14.00%	8.00%	0.00%	4.17%
E2	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	4.00%	2.00%	4.00%	1.39%
E3	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	6.00%	0.83%
E4	0.00%	0.00%	0.00%	6.00%	0.00%	4.00%	4.00%	8.00%	6.00%	3.89%
E5	0.00%	0.00%	0.00%	8.00%	2.00%	2.00%	2.00%	0.00%	16.00%	4.17%
E6	0.00%	25.00%	20.00%	0.00%	2.00%	6.00%	20.00%	6.00%	16.00%	9.44%
E7	0.00%	0.00%	0.00%	22.00%	4.00%	52.00%	0.00%	0.00%	58.00%	18.89%
E8	0.00%	0.00%	0.00%	2.00%	0.00%	0.00%	0.00%	0.00%	4.00%	0.83%
E9	60.00%	60.00%	95.00%	42.00%	8.00%	22.00%	52.00%	46.00%	40.00%	41.11%
E10	35.00%	40.00%	20.00%	2.00%	2.00%	6.00%	2.00%	4.00%	14.00%	9.44%
E12	50.00%	30.00%	25.00%	44.00%	4.00%	40.00%	30.00%	58.00%	46.00%	36.67%
E13	0.00%	0.00%	15.00%	0.00%	0.00%	2.00%	0.00%	0.00%	8.00%	2.22%
E13	0.00%	50.00%	15.00%	2.00%	2.00%	12.00%	14.00%	36.00%	10.00%	14.17%
E14	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	2.00%	0.00%	0.28%
E15	0.00%	0.00%	0.00%	0.00%	82.00%	0.00%	12.00%	2.00%	0.00%	13.33%
E16	75.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	4.17%
E17	35.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1.94%

Table 14: Distribution of Error sources (Err) per Model (Md) in the Error-Analysis of Experiments 1 and 2. The column average gives the average of each error source for each model. For an explanation of the error sources, see Table 15.

Error code	Explanation
E1	Issues with placing items in inter-domain relations
E2	Negations in clues
E3	Grid size not respected
E4	Abbreviations
E5	Item placed in the wrong domain
E6	Item placed prematurely (before it was possible to deduce the correct position)
E7	(partially) empty grid / columns and rows swapped / other issues with the grid
E8	Puzzle classified as not solvable
E9	Pseudo-Reasoning
E10	Position bias (items placed from left to right without justifiable indication)
E11	Inconsistency between produced free-text reasoning and solution grid
E12	Items positoned twice in the grid
E13	Clue misinterpreted
E14	Not all clues used
E15	No reasoning provided
E16	Repeats itself more than necessary
E17	Overwrites already correct solutions

Table 15: Reference table for error codes and explanation.

I Case Study

I.1 Solution paths

The solution paths used to analyse the stepwise accuracy of in the solution process are based on solution paths widely found on the internet [9]. The solution processes are relatively constrained for both puzzles due to their setup and refers to the clues as given in Table 6 in App. B. In the following the two solution paths will be described referring to the original item names. For the equivalents in the *in domain* and *in domain* obfuscations, refer to Tables 17 and 20.

I.1.1 Einstein solution path

Step 1: Norwegian, Milk This step places "Norwegian" and "Milk" directly to positions 1 and 3 based on clues 8) and 9). Hence, this step does not require further deduction and only basic understanding of the clues' semantics.

Step 2: Blue In this step, "blue" is placed to position 2 via a 1-step inference based on clue 14) stating that the Norwegian lives next to the blue house.

Step 3: Green, Coffee, White This step places "green", "white", and "coffee" based on clues 4) and 5). Restricted by the deduction from clue 5) which places Milk in the centre and from clue 14), the only remaining positions are 4 for "green" and "coffee" and 5 for "white".

Step 4: British, Red, Yellow Here, we place "British" and "Red" based on clue 1). Given the former deductions, the only remaining position for a colour and a nationality is 3. "yellow" fits then in the remaining spot for a colour: 1. Note that it is irrelevant for the correct position of yellow, whether "green", "white" and "red" were placed in the correct positions, as long as "blue" and "Norwegian" were placed correctly.

Step 5: Dunhill, Horses This step fills in "Dunhill" to position 1 and "Horses" to 2 based on the position of "yellow" and clues 7) and 11).

Step 6: Dane, Tea, Cat, Beer, Bluemaster, Water, Blend This step considers the possible positions for the groups [Dane + Tea] → 2 or 5 (clue 3), [Bluemaster + Beer] → 2 or 5 (clue 12), [Blend next to Cats and next to Water] → [2, 1, 1], [4, 3, 5], [5, 5] (clues 10 and 15) in the context of the former deductions in the respective domains. By elimination, the positions can be determined as follows. Blend:2, Water:1, Cat:1, Dane:2, Tea:2, Bluemaster:5, Beer:5.

Step 7: German, Prince, Swede, Dog Based on the former deductions and clues 2) and 13), there remains only one option to place the pairs [German + Prince] → 4 and [Swede + Dog] → 5. Also note that a former wrong placing of the other items into potential other positions still allows to place at least German and Prince correctly.

Step 8: Birds, Pall-Mall The only remaining position based on clue 6) and former deductions for "Pall-Mall" and "Birds" is 3. Also for these items, there is an increased probability of correct positioning since there is only one potential positioning before that would make position 3 unavailable.

Step 9: Fish In this step, the last remaining spot for a pet is filled, which is 4.

I.1.2 Zebra solution path

An important property of the Zebra puzzle is its branched reasoning path. The reasoning path reflects the order in which items can be placed without a doubt. For this reason, the path places the possible correct deductions during the branching before the deductions that can be made only after determining the correct branch.

Step 1: Norwegian, Milk This step places "Norwegian" and "Milk" directly to positions 1 and 3 based on clues 9) and 10). Hence, this step does not require further deduction and only basic understanding of the clues' semantics.

Step 2: Blue In this step, "blue" is placed to position 2 via a 1-step inference based on clue 15) stating that the Norwegian lives next to the blue house.

Step 3: Yellow, Kools, Horse From the third step, based on clues 6) and 4), there are two possible positions for "green"+"coffee", "ivory" and "red", which are either 4, 3, 5 or 5, 4, 3. This cannot be resolved until step 6, but the options leave only one option to position "yellow " correctly, which is 1 since blue occupies 2. Based on this and clue 8) and 12), "Kools" goes to position 1 and "Horse" goes to 1.

[9] For Einstein: https://www.chessandpoker.com/einsteins-problem-solution.html; For Zebra: https://waitbutwhy.com/table/zebra-puzzle

Step 4: Ukrainian, Tea Another conclusion that is independent from the branched reasoning is the positioning of “Ukrainian” and “Tea” based on clue 5). Independently from former conclusions, 2 is the only valid position for this. This is entirely independent from the deductions made in step 3 and the branch of the reasoning path that considers “green”, “ivory” and “red”.

Step 5: Water Based on Step 4 and clue 13), which demands to place “orange-juice” and “Lucky-Strike” together and blocks positions 4 or 5 depending on the branch taken for “green”+“coffee”. This means that only one option remains free for “water”, which is 1.

Step 6: Ivory, Green, Coffee At this point, it can be observed that only the reasoning path, where “green”+“coffee”, “ivory” and “red” are at 4, 5, 3 allows further valid deductions.

Step 7: Orange-juice, Lucky-Strike Following from the step before, the only remaining position for “orange-juice” and “lucky-strike” is 4.

Step 8: Japanese, Spanish, Red, English, Snails, Dog, Old-Gold, Parliaments In the next step clues 1), 3), 14), and 7) are evaluated. From the former steps “Spanish”+“Red” go into position 3. Since [Parliaments + Japanese] can only go to 5, the further possible options are limited to [Spanish + Dog] → 5, [Old-Gold + Snails] → 3.

Step 9: Fox, Chesterfields Finally, based on clue 11), there is only one option left to place “Fox” and “Chesterfields”, which are “fox”:1 and “Chesterfields”:2.

Step 10: Zebra This step fills the last remaining position for a pet, which is 4, with “Zebra”.

I.2 Per-step accuracy plots

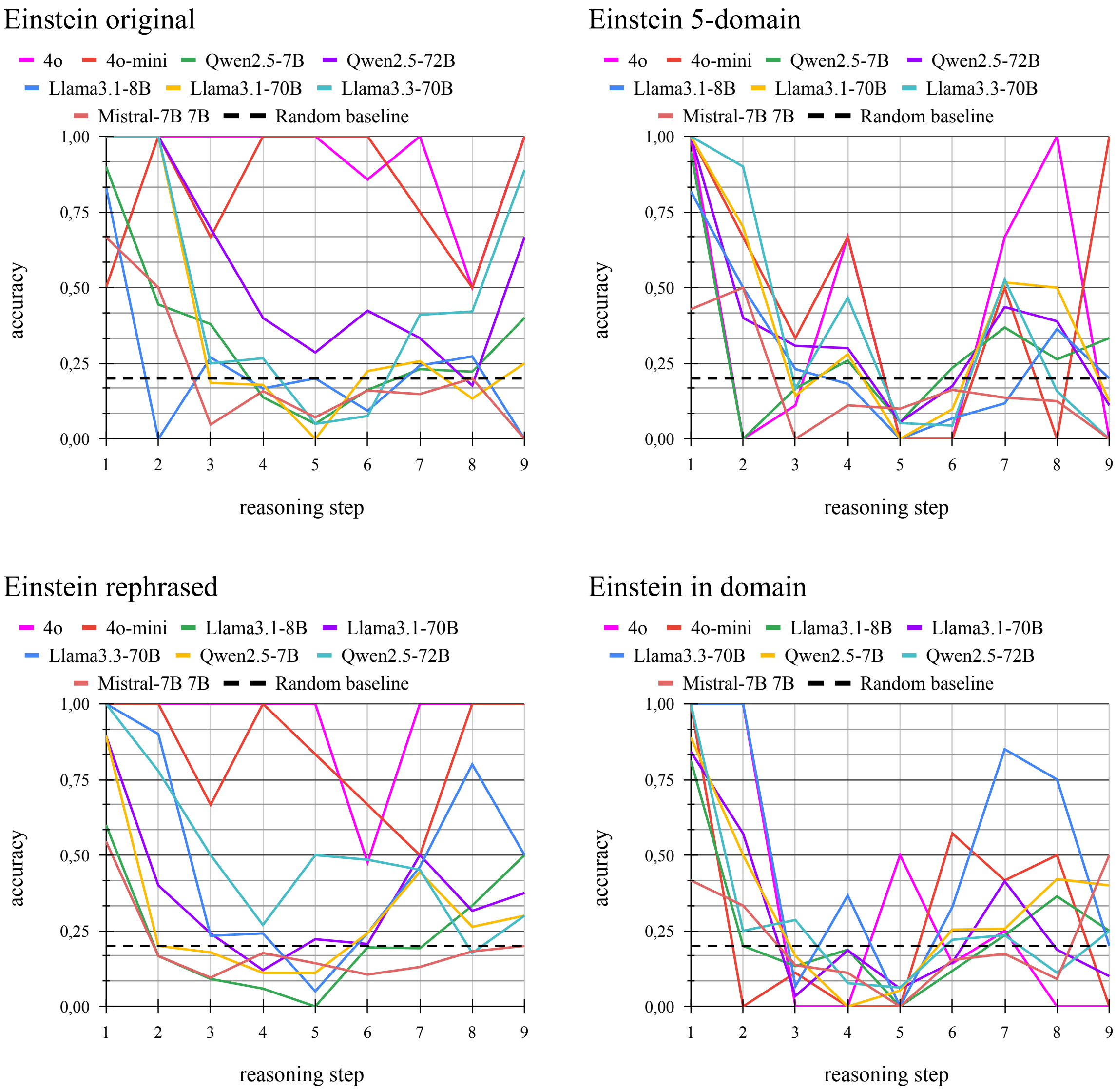


Figure 5: Q&A accuracy per reasoning step on Einstein puzzles with unchanged grids (*original*, *rephrased*) and manipulated grids (*5-domain*, *in domain*). The dashed line indicates the random baseline.

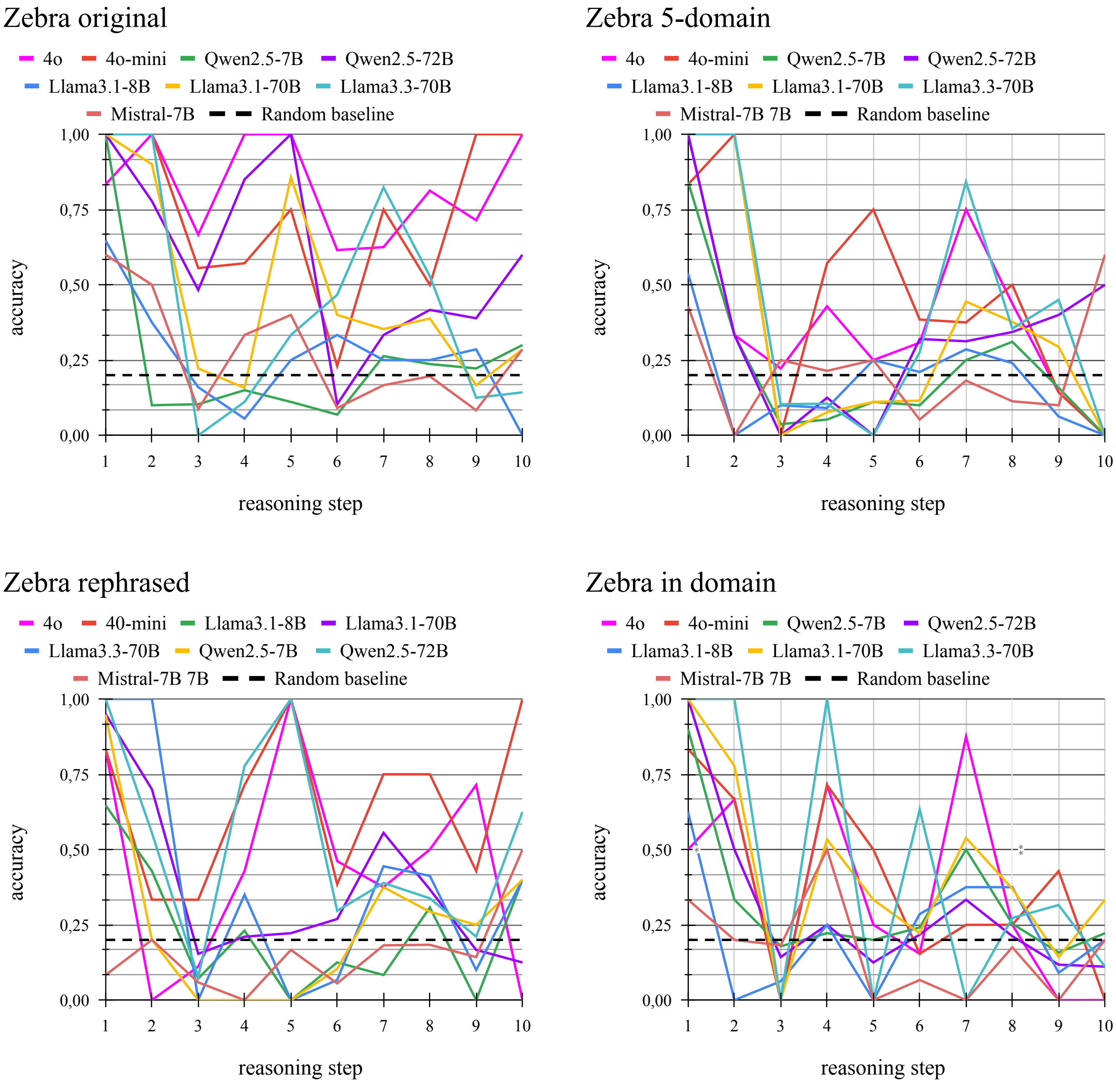


Figure 6: Q&A accuracy per reasoning step on Zebra puzzles with unchanged grids (*original*, *rephrased*) and manipulated grids (*5-domain*, *in domain*). The dashed line indicates the random baseline.

I.3 Per-item accuracy tables

Einstein original Q&A per item

		4o	4o-mini	Qwen2.5		Llama3.1		Llama3.3	Mistral
item	step			7B	72B	8B	70B	70B	7B
milk	1	1.00	0.00	0.80	1.00	0.67	1.00	1.00	0.56
norwegian	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.83
blue	2	1.00	1.00	0.44	1.00	0.00	1.00	1.00	0.50
coffee	3	1.00	0.00	0.30	0.43	0.33	0.11	0.30	0.00
green	3	1.00	1.00	0.44	0.75	0.25	0.30	0.20	0.00
white	3	1.00	1.00	0.40	0.88	0.22	0.13	0.25	0.14
brit	4	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
red	4	1.00	1.00	0.30	0.50	0.22	0.33	0.60	0.33
yellow	4	1.00	1.00	0.11	0.86	0.22	0.20	0.20	0.20
dunhill	5	1.00	1.00	0.10	0.67	0.38	0.00	0.10	0.25
horses	5	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
beer	6	1.00	1.00	0.30	0.44	0.17	0.40	0.30	0.56
blend	6	1.00	1.00	0.40	0.33	0.13	0.00	0.00	0.00
bluemaster	6	1.00	1.00	0.10	0.75	0.25	0.57	0.22	0.43
cats	6	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
dane	6	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
tea	6	1.00	1.00	0.22	0.57	0.00	0.22	0.00	0.00
water	6	1.00	1.00	0.10	1.00	0.13	0.43	0.00	0.00
dogs	7	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
german	7	1.00	1.00	0.60	0.78	0.56	0.56	0.70	0.38
prince	7	1.00	1.00	0.30	0.44	0.43	0.44	0.90	0.33
swede	7	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
birds	8	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
pall mall	8	0.00	0.00	0.50	0.38	0.60	0.29	0.80	0.33
fish	9	1.00	1.00	0.40	0.67	0.00	0.25	0.89	0.00

Einstein rephrased Q&A per item

		4o	4o-mini	Qwen2.5		Llama3.1		Llama3.3	Mistral
item	step			7B	72B	8B	70B	70B	7B
milk	1	1.00	1.00	0.80	1.00	0.38	0.90	1.00	0.43
norwegian	1	1.00	1.00	1.00	1.00	0.86	0.89	1.00	0.75
blue	2	1.00	1.00	0.20	0.78	0.17	0.40	0.90	0.17
coffee	3	1.00	0.00	0.11	0.63	0.20	0.11	0.20	0.00
green	3	1.00	1.00	0.11	0.33	0.00	0.20	0.10	0.00
white	3	1.00	1.00	0.30	0.56	0.13	0.40	0.40	0.29
british	4	1.00	1.00	0.22	0.00	0.00	0.33	0.60	0.00
red	4	1.00	1.00	0.00	0.11	0.14	0.10	0.10	0.29
yellow	4	1.00	1.00	0.10	0.67	0.00	0.00	0.00	0.20
dunhill	5	1.00	0.67	0.00	0.75	0.00	0.33	0.00	0.00
horse	5	1.00	1.00	0.22	0.25	0.00	0.11	0.10	0.25
beer	6	0.00	0.00	0.30	0.56	0.29	0.30	0.00	0.13
blend	6	1.00	1.00	0.20	0.40	0.14	0.30	0.20	0.00
bluemaster	6	0.00	0.67	0.56	0.10	0.40	0.00	0.10	0.00
cat	6	0.00	1.00	0.10	0.29	0.00	0.20	0.00	0.00
danish	6	1.00	1.00	0.30	0.80	0.40	0.14	0.80	0.00
tea	6	1.00	0.00	0.22	0.40	0.00	0.00	0.60	0.40
water	6	0.33	1.00	0.00	0.80	0.17	0.44	0.00	0.14
dog	7	1.00	1.00	0.33	0.56	0.25	0.50	0.22	0.17
german	7	1.00	0.00	0.75	0.50	0.25	0.30	0.60	0.00
prince	7	1.00	1.00	0.33	0.11	0.00	0.63	0.70	0.00
swedish	7	1.00	0.00	0.40	0.80	0.29	0.63	0.30	0.40
bird	8	1.00	1.00	0.44	0.20	0.33	0.40	0.80	0.00
pall-mall	8	1.00	1.00	0.10	0.14	0.33	0.22	0.80	0.29
fish	9	1.00	1.00	0.30	0.30	0.50	0.38	0.50	0.20

Table 16: Q&A results on “Einstein original” and “Einstein rephrased” as per-item accuracy.

Einstein *in domain* Q&A per item

			4o	4o-mini	Qwen2.5		Llama3.1		Llama3.3	Mistral
item	orig item	step			7B	72B	8B	70B	70B	7B
dutch	norwegian	1	1.00	1.00	1.00	1.00	0.89	1.00	1.00	0.50
mirinda	milk	1	1.00	1.00	0.75	1.00	0.71	0.67	1.00	0.38
chestnut	blue	2	1.00	0.00	0.50	0.25	0.20	0.57	1.00	0.33
7up	coffee	3	0.00	0.00	0.10	0.29	0.00	0.00	0.00	0.33
aquamarine	white	3	0.00	0.00	0.30	0.50	0.17	0.11	0.20	0.14
grey	green	3	0.00	0.33	0.10	0.13	0.20	0.00	0.00	0.00
azure	yellow	4	0.00	0.00	0.00	0.00	0.14	0.00	0.00	0.20
black	red	4	0.00	0.00	0.00	0.00	0.00	0.44	0.40	0.00
japanese	english	4	0.00	0.00	0.00	0.25	0.29	0.13	0.70	0.13
game	dunhill	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
turtle	horse	5	1.00	0.00	0.11	0.10	0.00	0.13	0.00	0.00
almond-milk	tea	6	0.00	0.00	0.57	0.11	0.00	0.00	0.00	0.00
australian	dane	6	0.00	0.00	0.30	0.10	0.20	0.00	0.20	0.40
chaman	blend	6	1.00	1.00	0.00	0.13	0.00	0.13	0.10	0.33
davidoff	bluemaster	6	0.00	1.00	0.33	0.38	0.00	0.33	1.00	0.60
fanta	beer	6	0.00	1.00	0.56	0.30	0.50	0.25	1.00	0.00
ferret	cat	6	0.00	0.00	0.11	0.11	0.00	0.29	0.00	0.00
hot-chocolate	water	6	0.00	1.00	0.00	0.60	0.14	0.00	0.00	0.00
french	german	7	1.00	0.67	0.30	0.00	0.00	0.13	0.80	0.00
guinea-pig	dog	7	0.00	0.00	0.20	0.29	0.50	0.80	0.60	0.43
havana	prince	7	0.00	1.00	0.44	0.38	0.60	0.57	1.00	0.00
spanish	swedish	7	0.00	0.00	0.10	0.30	0.00	0.33	1.00	0.20
baccarat	pall-mall	8	0.00	0.00	0.33	0.00	0.33	0.00	0.80	0.00
chinchilla	bird	8	0.00	1.00	0.50	0.22	0.40	0.43	0.70	0.11
lizard	fish	9	0.00	0.00	0.40	0.25	0.25	0.10	0.20	0.50
Variance			0.44	0.47	0.26	0.27	0.25	0.28	0.43	0.20

Einstein *in domain* Q&A per item

			4o	4o-mini	Qwen2.5		Llama3.1		Llama3.3	Mistral
item	orig item	step			7B	72B	8B	70B	70B	7B
folk	norwegian	1	1.00	1.00	0.90	1.00	1.00	1.00	1.00	0.50
gothic-revival	milk	1	1.00	1.00	1.00	1.00	0.60	1.00	1.00	0.38
motorbike	blue	2	0.00	0.67	0.00	0.40	0.50	0.70	0.90	0.50
snowmobile	white	3	0.33	1.00	0.30	0.63	0.00	0.11	0.30	0.00
subway	green	3	0.00	0.00	0.00	0.20	0.25	0.10	0.10	0.00
wooden	coffee	3	0.00	0.00	0.20	0.13	0.29	0.22	0.10	0.00
quad-bike	yellow	4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
train	red	4	1.00	1.00	0.22	0.50	0.25	0.50	0.80	0.00
trance	english	4	1.00	1.00	0.63	0.40	0.25	0.38	0.60	0.25
july	horse	5	0.00	0.00	0.10	0.00	0.00	0.00	0.11	0.00
mechanic	dunhill	5	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.25
futuristic	water	6	0.00	0.00	0.00	0.13	0.50	0.13	0.11	0.17
june	cat	6	0.00	0.00	0.30	0.00	0.33	0.00	0.00	0.00
librarian	blend	6	0.00	0.00	0.10	0.20	0.00	0.25	0.20	0.40
motorbikemaster	bluemaster	6	0.00	0.00	0.50	0.13	0.00	0.13	0.00	0.00
palace	tea	6	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.17
ranch-style	beer	6	0.00	0.00	0.50	0.40	0.00	0.11	0.00	0.00
rock	dane	6	0.00	0.00	0.13	0.30	0.00	0.13	0.00	0.29
ambient	german	7	1.00	0.00	0.30	0.40	0.00	0.20	0.70	0.20
bartender	prince	7	0.00	0.00	0.50	0.30	0.00	0.13	0.10	0.33
reggae	swedish	7	0.67	1.00	0.22	0.56	0.00	0.86	0.60	0.00
september	dog	7	1.00	1.00	0.44	0.50	0.33	0.78	0.75	0.14
architect	pall-mall	8	1.00	0.00	0.44	0.50	0.17	0.44	0.00	0.33
may	bird	8	1.00	0.00	0.10	0.30	0.60	0.56	0.33	0.00
february	fish	9	0.00	1.00	0.33	0.11	0.20	0.13	0.00	0.00
Variance			0.47	0.48	0.27	0.28	0.26	0.33	0.36	0.18

Table 17: Q&A results on “Einstein” *in domain* and Einstein *in domain* as per-item accuracy.

Zebra original Q&A per item

		4o	4o-mini	Qwen2.5		Llama3.1		Llama3.3	Mistral
item	step			7B	72B	8B	70B	70B	7B
milk	1	1.00	1.00	1.00	1.00	0.56	1.00	1.00	0.67
norwegian	1	1.00	1.00	1.00	1.00	0.75	1.00	1.00	0.50
blue	2	1.00	1.00	0.10	0.78	0.38	0.90	1.00	0.50
fox	3	0.00	0.00	0.11	0.20	0.38	0.63	0.00	0.00
lucky-strike	3	1.00	1.00	0.10	0.60	0.13	0.00	0.00	0.11
yellow	3	1.00	1.00	0.10	0.67	0.00	0.11	0.00	0.13
tea	4	1.00	0.00	0.20	0.90	0.13	0.20	0.11	0.38
ukrainian	4	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
water	5	1.00	1.00	0.11	1.00	0.25	0.86	0.33	0.40
coffee	6	0.00	0.00	0.00	0.00	0.13	0.10	0.10	0.13
green	6	1.00	0.00	0.00	0.20	0.33	0.30	0.50	0.00
ivory	6	1.00	0.00	0.22	0.11	0.57	0.80	0.80	0.13
horse	7	0.00	0.33	0.44	0.25	0.38	0.50	1.00	0.25
orange-juice	7	0.67	1.00	0.10	0.43	0.00	0.22	0.57	0.13
dog	8	0.00	0.00	0.10	0.29	0.17	0.33	0.44	0.00
english	8	1.00	1.00	0.00	0.63	0.17	0.11	0.67	0.00
japanese	8	1.00	0.00	0.56	0.63	0.38	0.30	0.80	0.71
old-gold	8	1.00	1.00	0.22	0.20	0.00	0.33	0.40	0.25
parliaments	8	1.00	1.00	0.40	0.80	0.38	0.70	0.89	0.67
red	8	1.00	0.33	0.50	0.20	0.13	0.22	0.10	0.00
snails	8	1.00	1.00	0.00	0.10	0.29	0.50	0.00	0.00
spanish	8	0.67	0.00	0.10	0.56	0.33	0.63	0.89	0.00
kools	9	1.00	1.00	0.33	0.63	0.33	0.20	0.00	0.00
zebra	9	0.00	1.00	0.11	0.20	0.00	0.13	0.33	0.14
chesterfields	10	1.00	1.00	0.30	0.60	0.00	0.29	0.14	0.29

Zebra rephrased Q&A per item

		4o	4o-mini	Qwen2.5		Llama3.1		Llama3.3	Mistral
item	step			7B	72B	8B	70B	70B	7B
milk	1	1.00	1.00	0.50	0.90	1.00	0.89	1.00	0.17
norwegian	1	1.00	1.00	0.78	1.00	1.00	1.00	1.00	0.00
blue	2	0.00	0.33	0.43	0.70	1.00	0.20	0.56	0.20
horse	3	0.00	0.00	0.00	0.22	0.00	0.00	0.00	0.20
kools	3	0.00	1.00	0.13	0.00	0.00	0.00	0.25	0.00
yellow	3	0.00	0.00	0.00	0.22	0.00	0.00	0.00	0.00
tea	4	1.00	1.00	0.40	0.22	0.30	0.00	0.89	0.00
ukrainian	4	0.00	1.00	0.13	0.20	0.40	0.00	0.67	0.00
water	5	1.00	1.00	0.00	0.22	0.00	0.00	1.00	0.17
coffee	6	0.67	0.33	0.00	0.25	0.00	0.10	0.30	0.20
green	6	0.33	0.00	0.00	0.22	0.00	0.10	0.40	0.00
ivory	6	0.00	0.00	0.29	0.33	0.20	0.11	0.14	0.00
lucky-strike	7	1.00	0.67	0.17	0.50	0.78	0.29	0.33	0.00
orange-juice	7	0.00	1.00	0.00	0.60	0.11	0.44	0.44	0.29
dog	8	0.00	0.00	0.00	0.22	0.30	0.00	0.30	0.13
english	8	1.00	1.00	0.57	0.13	0.20	0.20	0.25	0.00
japanese	8	1.00	1.00	0.29	0.88	1.00	0.78	0.67	0.20
old-gold	8	1.00	1.00	0.25	0.25	0.00	0.10	0.14	0.25
parliament	8	0.33	1.00	0.38	0.75	1.00	0.80	0.60	0.75
red	8	0.00	0.67	0.57	0.20	0.10	0.30	0.25	0.00
snails	8	1.00	1.00	0.13	0.10	0.00	0.00	0.22	0.17
spanish	8	0.00	0.00	0.20	0.57	0.70	0.10	0.20	0.29
chesterfield	9	0.33	0.00	0.00	0.11	0.20	0.50	0.33	0.33
fox	9	1.00	0.67	0.00	0.22	0.00	0.00	0.10	0.00
zebra	10	0.0	0.67	0.40	0.13	0.40	0.40	0.63	0.50

Table 18: Q&A results on “Zebra original” and “Zebra rephrased” as per-item accuracy.

Table 19: Generated by Spread-LaTeX

Zebra *in domain* Q&A per item										
			4o	4o-mini	Qwen2.5		Llama3.1		Llama3.3	Mistral
item	orig item	step			7B	72B	8B	70B	70B	7B
german	norwegian	1	1.00	1.00	1.00	1.00	0.83	1.00	1.00	0.14
lemonade	milk	1	0.00	1.00	0.80	1.00	0.00	1.00	1.00	0.50
coral	blue	2	1.00	1.00	0.33	0.50	0.00	0.78	1.00	0.20
chestnut	yellow	3	0.00	0.00	0.22	0.10	0.00	0.00	0.00	0.00
havana	kools	3	0.00	0.00	0.10	0.11	0.00	0.00	0.00	0.00
mouse	horse	3	0.00	0.00	0.22	0.22	0.20	0.00	0.00	0.44
almond-lemonade	tea	4	1.00	0.67	0.22	0.13	0.33	0.83	1.00	0.00
dutch	ukrainian	4	1.00	1.00	0.22	0.38	0.20	0.33	1.00	0.89
hot-chocolate	water	5	0.00	1.00	0.20	0.13	0.00	0.33	0.00	0.00
aquamarine	ivory	6	0.00	0.00	0.20	0.11	0.20	0.22	0.30	0.13
black	green	6	0.00	0.00	0.40	0.29	0.33	0.25	0.60	0.00
cola	coffee	6	0.00	0.00	0.11	0.29	0.33	0.20	1.00	0.00
baccarat	lucky-strike	7	1.00	0.00	0.56	0.38	0.33	0.57	0.00	0.00
iced-tea	orange-juice	7	1.00	0.00	0.44	0.29	0.40	0.50	0.00	0.00
bird	snails	8	0.00	0.00	0.30	0.13	0.40	0.11	0.00	0.00
fonseca	old-gold	8	1.00	0.00	0.00	0.11	0.25	0.22	0.00	0.13
goldfish	dog	8	0.00	0.00	0.00	0.00	0.33	0.25	0.10	0.25
italian	spanish	8	0.00	0.00	0.22	0.40	0.22	0.30	0.00	0.13
malaysian	english	8	0.00	0.00	0.10	0.00	0.14	0.22	0.10	0.20
mexican	japanese	8	1.00	0.67	0.70	0.29	0.38	0.89	1.00	0.33
orange	red	8	0.00	0.33	0.00	0.00	0.50	0.11	0.00	0.14
tiparillo	parliaments	8	0.00	1.00	0.60	0.67	1.00	0.80	1.00	0.25
lizard	fox	9	0.00	0.67	0.00	0.13	0.00	0.00	0.00	0.00
Pall-mall	chesterfields	9	0.00	0.00	0.30	0.11	0.20	0.25	0.60	0.00
cat	zebra	10	0.00	0.00	0.22	0.11	0.20	0.33	0.11	0.20
Variance			0.48	0.44	0.26	0.27	0.25	0.32	0.46	0.21

Zebra *in domain* Q&A per item										
			4o	4o-mini	Qwen2.5		Llama3.1		Llama3.3	Mistral
item	orig item	step			7B	72B	8B	70B	70B	7B
electronic	norwegian	1	1.00	1.00	0.89	1.00	0.57	1.00	1.00	0.57
monopoly	milk	1	1.00	1.00	0.80	1.00	0.50	1.00	1.00	0.29
lettuce	blue	2	0.67	1.00	0.33	0.33	0.00	1.00	1.00	0.00
orchid	horse	3	0.67	0.00	0.00	0.00	0.00	0.00	0.33	0.29
radish	yellow	3	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.33
victorian	kools	3	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00
domino	tea	4	0.33	1.00	0.00	0.13	0.20	0.00	0.20	0.33
techno	ukrainian	4	0.00	0.33	0.11	0.13	0.00	0.20	0.00	0.13
chess	water	5	0.33	1.00	0.11	0.00	0.25	0.11	0.00	0.25
carrot	ivory	6	0.67	0.33	0.20	0.33	0.25	0.20	0.40	0.00
go	coffee	6	0.00	0.67	0.10	0.33	0.14	0.14	0.22	0.00
zucchini	green	6	0.00	0.00	0.00	0.29	0.25	0.00	0.20	0.17
mah-jongg	orange-juice	7	1.00	0.00	0.30	0.33	0.14	0.56	0.78	0.33
townhouse	lucky-strike	7	1.00	0.33	0.20	0.29	0.43	0.33	0.90	0.13
azalea	dog	8	0.00	1.00	0.11	0.11	0.25	0.40	0.50	0.13
bellflower	snails	8	0.33	0.00	0.14	0.00	0.00	0.25	0.00	0.00
colonial	old-gold	8	0.00	0.00	0.20	0.00	0.14	0.00	0.30	0.17
futuristic	parliaments	8	1.00	1.00	0.70	0.78	0.71	1.00	1.00	0.00
indie	english	8	0.00	0.00	0.11	0.44	0.14	0.00	0.00	0.00
onion	red	8	0.00	1.00	0.00	0.13	0.17	0.00	0.00	0.11
pop	spanish	8	1.00	0.00	0.20	0.11	0.29	0.25	0.10	0.00
trance	japanese	8	1.00	1.00	1.00	1.00	0.14	1.00	1.00	0.43
gothic-revival	chesterfields	9	0.33	0.00	0.20	0.60	0.10	0.44	0.90	0.00
marigold	fox	9	0.00	0.00	0.11	0.20	0.00	0.13	0.00	0.14
dahlia	zebra	10	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.60
Variance			0.43	0.47	0.29	0.33	0.19	0.38	0.41	0.18

Table 20: Q&A results on Zebra *in domain* and Zebra *in domain* as per-item accuracy.