Alice in Wonderland: Simple Tasks Reveal Severe Generalization and Basic Reasoning Deficits in State-Of-the-Art Large Language Models

Marianna Nezhurina[1,2,4*] **Lucia Cipolina-Kun**[1,2,3,4] **Mehdi Cherti**[1,2,4] **Jenia Jitsev**[1,2,4*]
[1]LAION [2]Juelich Supercomputing Center (JSC), Research Center Juelich (FZJ)
[3] University of Bristol [4] Open-Ψ (Open-Sci) Collective
[*]Corresponding authors: {m.nezhurina,j.jitsev}@fz-juelich.de,contact@laion.ai

Abstract

Large Language Models (LLMs) are often described as being instances of foundation models - that is, models that possess strong generalization and therefore transfer robustly across various tasks and conditions in few-shot or zero-shot manner, while exhibiting scaling laws that predict generalization improvement when increasing the pre-training scale. These claims of strong generalization and advanced function enabling it like reasoning rely on measurements by various standardized benchmarks where state-of-the-art (SOTA) models score high. We demonstrate here a dramatic breakdown of generalization and basic reasoning of all SOTA models which claim strong function, including advanced models like GPT-4 or Claude 3 Opus trained at the largest scales, using a simple, short, conventional common sense problem formulated in concise natural language, easily solvable by humans (AIW problem). The breakdown is dramatic as it manifests in strong performance fluctuations across mild problem variations that should not affect problem solving at all, while also often expressing strong overconfidence in the wrong solutions, backed up by plausible sounding explanation-like confabulations. Various standard interventions in an attempt to get the right solution, like chain-of-thought prompting, or urging the models to reconsider the wrong solutions again by multi step re-evaluation, fail. We take these initial observations to the scientific and technological community to stimulate urgent re-assessment of the claimed capabilities of current generation of LLMs. Such re-assessment also requires common action to create standardized benchmarks that would allow proper detection of such deficits in generalization and reasoning that obviously manage to remain undiscovered by current state-of-the-art evaluation procedures.[1]

1 Introduction

In the recent breakthroughs in transferable learning that were achieved in various classical domains of machine learning like visual recognition [1] or language understanding [2, 3, 4], large language models (LLMs) have played a very prominent role. The generic form and scalability of autoregressive language modelling [4] allowed to push towards training scales not achievable before with conventional supervised label-based learning. Scaling laws derived from experiments on smaller scales hinted on strong function and generalization capability appearing at larger scales [5, 6], which was then confirmed by training models at the large scales, measuring their performance on set of standardized benchmarks (MMLU, HellaSwag, ARC, MATH, GSM8k, etc) where they scored high on few- and zero-shot transfer across various tasks [7], following accurately the predictions [4, 5, 8, 9, 10, 11].

[1]Code for reproducing experiments in the paper and raw experiments data can be found at AIW repo

Workshop on Scientific Methods for Understanding Deep Learning, NeurIPS 2024.

There were however observations made by various works that questioned the claimed strong generalization, transfer and reasoning capabilities attributed to LLMs [12]. These works pointed out various function failures that were seemingly incompatible with postulated strong capabilities as measured by standardized benchmarks [13, 14, 15, 16]. However, it has also been noted that observed failures can frequently be addressed through simple adjustments to the prompts or by repeated execution and evaluation using majority voting, or by requesting the model to perform self-verification. [17, 18, 19, 20, 21]. It remained thus unclear where those observations of failures were pointing to some fundamental deficits in core model capabilities affecting generalization and reasoning, or whether those were just symptoms of minor issues easily resolvable by simple interventions, leaving claim of strong core function as put forward by standardized benchmarks unaffected.

To shed light on current situation, we study whether the claim of the strong SOTA LLMs functions made across various complex tasks can be put to test by using very simple tasks, in contrast to those employed by standardized benchmarks. We introduce a short conventional common sense problem that is formulated without any ambiguities in concise natural language and can be easily solved by humans. The problem (in following Alice in Wonderland, AIW problem) has following template: ***"Alice has N brothers and she also has M sisters. How many sisters does Alice's brother have?"***. Crucially, replacing N, M with natural numbers ≤ 7 allows us to introduce systematic variations that do not change problem structure and thus should not affect difficulty to solve it. We use then this technique of creating problem structure irrelevant variations to measure models' sensitivity to problem irrelevant perturbations, testing models' generalization ability.

Surprisingly, when confronted with AIW problem and its mild variations, all SOTA models including most advanced large-scale ones (eg GPT-4 [22], Claude 3 Opus [23]) suffer severe function breakdown. This breakdown manifests (i) in average correct response rates that are unexpectedly low for such a simple problem and in (ii) strong fluctuations in correct response rates across AIW problem variations despite those being entirely irrelevant for coping with the problem. Strong fluctuations remain despite using various standard interventions to improve model function like chain-of-thought prompting. By creating further control versions of AIW problem and observing that models are successfully coping with those, we are able to rule out that observed failures might be rooted in minor low level issues of tokenization, handling the specific relational family structure frame, or executing arithmetic operations necessary to solve the problem. Despite a specific form of simple AIW problem, we can thus conclude that observed failure has generic character. The lack of robustness revealed in all SOTA models by problem irrelevant variations of a simple problem points to severe generic deficits in generalization and reasoning.

The observed breakdown of function and generalization is in strong contrast to scores on standardized benchmarks, which contain problems of higher difficulty. Many tested models that score high on such benchmarks show correct response rates close to zero across simple AIW problem variations. Claims put forward by standardized benchmarks to properly reflect model capabilities such as generalization and reasoning cannot be upheld in face of the evident failure to detect such severe function deficits as revealed by the simple AIW problem. Our study highlights necessity to re-assess current capabilities of SOTA LLMs by creating novel benchmarks that properly reflect their true abilities to generalize and reason. Such benchmarks will be able to correctly spot deficits overlooked so far and thus show the path for improvement of current still unsatisfactory state.

2 Methods & Experiment Setup

AIW Problem. To measure model's sensitivity to problem irrelevant variations and thus probe the zero-shot generalization, we use following problem *template*: ***"Alice has N brothers and she also has M sisters. How many sisters does Alice's brother have?"***. The problem has a simple common sense solution which assumes all sisters and brothers in the problem setting share the same parents. The correct response C - number of sisters - is easily obtained by calculating $M + 1$ (Alice and her sisters), which gives the number of sisters Alice's brother has. To create problem variations, we use the problem template and vary natural numbers $N, M \leq 7$, obtaining AIW variations 1-4 (see Suppl. Tab. 2), generating in this way instances of the same problem with variations irrelevant for problem difficulty level. We further use 3 *prompt types*, RESTRICTED, STANDARD and THINKING, to ensure that observed behavior is consistent across various prompt formulations. STANDARD and THINKING prompt types allow models to freely output any text before arriving to a final solution, where THINKING contains in addition usual chain-of-thought (CoT) instruction. RESTRICTED

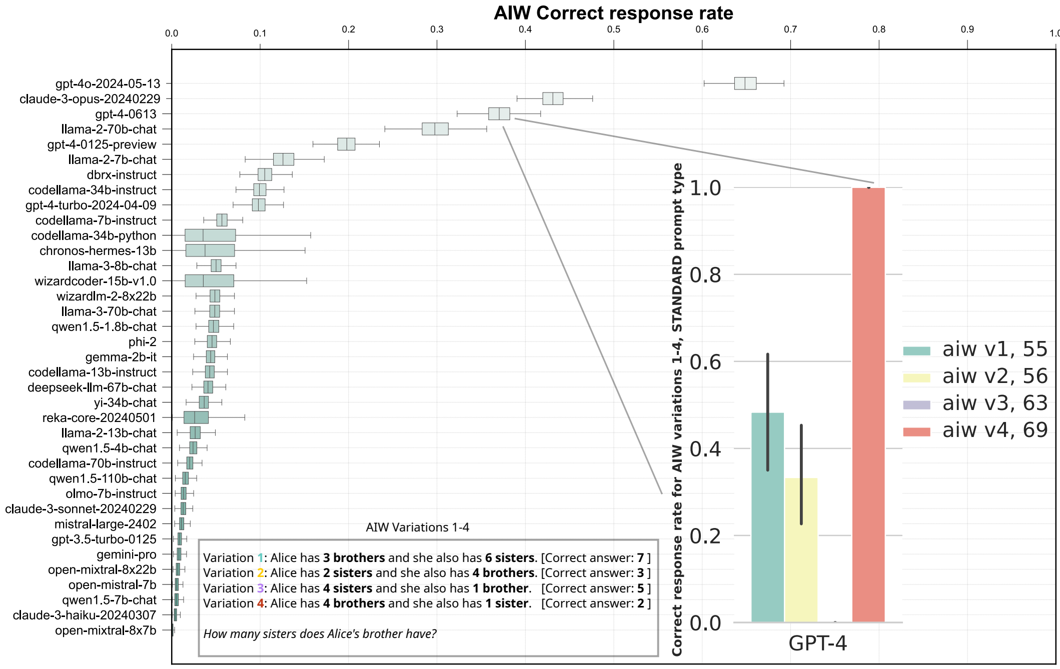

Figure 1: Collapse of most SOTA LLMs on AIW problem. (**main**) Models with non-zero AIW average correct response rate (across AIW variations 1-4 using THINKING and STANDARD prompt types). Given AIW simplicity, achieved correct response rates are surprisingly low. Many models claiming strong skills via standardized benchmarks score close to 0 or 0. Omitted models score 0. (**inlay**) In addition to low average correct response rates, strong correct response rate fluctuations across problem irrelevant variations (STANDARD prompt type, see Fig. 2 for THINKING type) reveal severe lack of robustness. Each colored bar shows correct response rate for one of AIW variations 1-4, estimated from > 30 trials per variation (see Suppl. Sec. C.1 for estimation procedure).

prompt type is used as control, containing instruction to output only the final answer as a number and no other text. Models receive for each single trial an input that has a form <*instantiated-template*> <*prompt-type*>, where <*instantiated-template*> is template with substituted numbers instantiating one of AIW Variations 1-4 and <*prompt-type*> contains question and output instructions corresponding to one of prompt types as described above. This allows us to test any combination of AIW Variations 1-4 with any of prompt types. See Suppl. Sec. B for more details and for examples of full versions of the problem related inputs to the models (Suppl. Tab. 2).

Experiment setup and model response evaluation. We select current SOTA LLMs models to test, including most advanced models at largest scales (see Suppl. Tab. 1) and expose each model to all AIW problem variations 1-4 (Suppl. Tab. 2), using different prompt types as described above. For each combination of model, variation and prompt type, at least 30 trials are collected to compute correct response rates (see Suppl. Sec. C.1 for estimation procedure). We employ further control AIW problem versions (Suppl. Sec. C.2.1) to test model abilities to handle various operations required for the problem solution. We re-use same experiment setup, again constructing problem variations on the problem template that stays close to AIW original, also ensuring natural numbers for correct responses match on AIW control and original problems across variations.

3 Results

AIW reveals severe generalization and reasoning deficits in SOTA LLMs. Following the procedures described in Sec 2, we expose the selected models that claim strong function and reasoning capabilities (Suppl. Tab. 1) and measure their correct response rate performance across and for each AIW variations 1-4 using various prompt types, executing >30 trials for each combination (see also Suppl. Tab. 2 and Suppl. Fig. 16). The results suggest that confronted with the AIW problem, models suffer a severe function breakdown. This breakdown has two main manifestations:

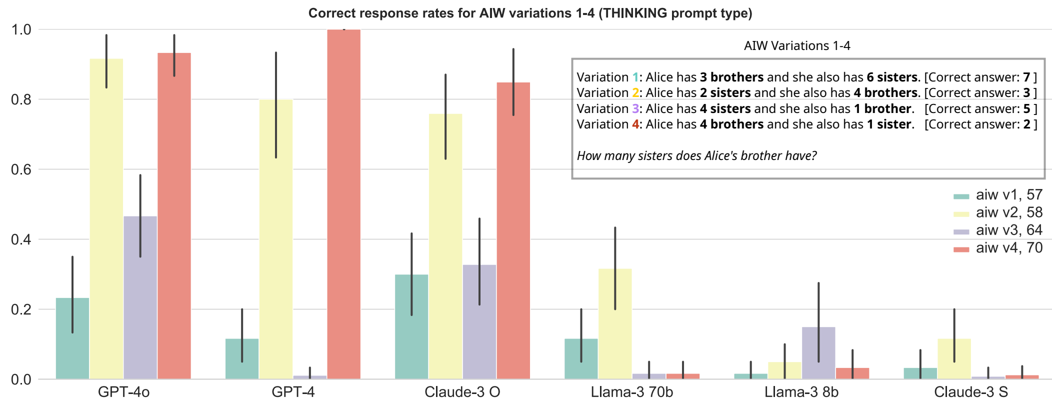

Figure 2: Strong fluctuations across AIW problem variations, THINKING prompt. Also for better performers, eg GPT-4o, GPT-4 and Claude Opus 3, correct response rates vary strongly from close to 1 to close to 0, despite AIW variations being irrelevant for problem structure (a color per each variation 1-4). This shows clear lack of model robustness, revealing generalization and basic reasoning deficits.

1. Low correct response rates. Despite evident problem's simplicity, many models are not able to deliver a single correct response, and the majority stays well below correct response rate of $p = 0.2$. We summarize the main results in the Fig. 1. The only major exceptions from the observation of very low correct response rates are the largest scale closed models GPT-4 and Claude 3 Opus. These two model types at largest scales obtain correct response rates well above $p = 0.3$, leaving the remaining large and smaller scales open-weights (e.g., Mistral-7B, Mixtral, Qwen, Command R+, and Dbrx Instruct) and closed-weights models (e.g., Gemini Pro, Mistral Large) far behind. Noteworthy, we do see also rare responses with full correct reasoning leading to a correct final answer, which are dominated by vast majority of wrong responses making up the average low correct response rates.

2. Strong performance fluctuations across mild AIW problem variations. Importantly, we also observe strong fluctuation of correct response rates across AIW variations 1-4 as introduced in Sec. 2. Such fluctuations also strongly affect better performers with higher average correct response rates like GPT-4o and Claude 3 Opus. As shown in the Fig. 2 for the THINKING prompt type, the correct response rates can fluctuate between being close to 1 to being close to 0, depending on AIW variation. Remarkable is that such fluctuations appear despite AIW variations being all instances of the very same simple problem, as changes in numbers used across AIW variations do not change the core problem structure at all. This lack of robustness on such a simple problem hints on severe deficits in generalization. The fluctuations persist across all prompt types, while overall average correct response rate can change. For instance, expectedly, we see higher average correct response rates across models for THINKING chain-of-thought prompt type, while strong fluctuations across variations remain.

AIW Light control experiments To rule out that observed failures might be low-level issues specific to the chosen AIW problem frame, like deficits of the tested models to handle basic family relations or to select and execute arithmetic operations, we conduct experiments on control AIW Light problem versions, AIW Light Family and AIW Light Arithmetic (Suppl. Sec. C.2.1). AIW Light problems have templates similar to AIW original and are formulated such that these low-level operations are necessary for the correct solution. In Fig. 3 are results for models successfully coping with AIW Light Arithmetic Total Girls version, which has following problem template: ***"Alice has N brothers and she also has M sisters. How many girls are there in total?"***. Compared to AIW original, only question part is modified. To solve the problem, it is necessary to bind female attribute to Alice via the pronoun "she", to assign correct female attributes to the sisters and to execute the correct arithmetic sum operation adding all the obtained girls - the correct answer is $C = M + 1$. We obtain evidence that majority of models that fail on AIW original are solving AIW Light problems with correct response rates close to 1, importantly, without showing strong fluctuations across AIW variations. This rules out low-level issues like tokenization and natural language or natural numbers parsing, family relations or elementary arithmetic operations handling as failure sources. Success on AIW Light problems provides thus further evidence that failures observed on AIW problem are rooted in more generic, problem unspecific deficits in generalization and basic reasoning.

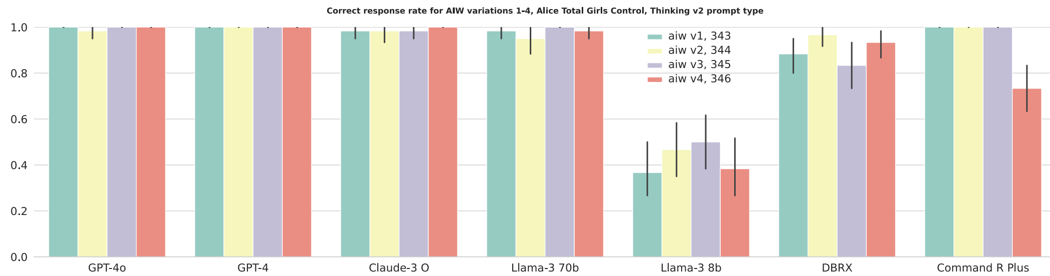

Figure 3: Correct response rates across AIW Light Arithmetic Total Girls problem variations 1-4 (THINKING prompt type). Tested models show stable correct response rates without strong fluctuations across problem variations (a color per each variation 1-4). This shows ability of models to handle binding of female attributes to Alice and sisters entities within the basic family relations structure of AIW problem and to select and execute arithmetic operations necessary to count total females. This provides evidence that neither properly identifying entities with certain sex attributes nor counting those poses an issue. Models that entirely collapse on AIW, like Command R Plus and Dbrx Instruct (Suppl. Fig. 4), are clearly able to solve this version, with correct response rates going up to 1 or close to 1 across all problem variations. For each AIW variation, 60 trials were executed to estimate correct response rate and its variance.

Further relevant observations. *1. Dominance of wrong responses* Models show failure to properly detect mistakes and to revise wrong solutions when encouraged to do so in experiments with multi-turn AIW problem interaction and self-verification. Distribution of natural numbers responses on output shows moreover that for AIW variations with low correct response rate, peaks are on wrong answers, excluding majority voting methods as a fix. (Suppl. Sec. C.3) *2. Confabulations, overconfident tone and inability to revise wrong responses* We observe that wrong responses are often accompanied by persuasive explanation-like confabulations and overconfident tone about correctness of the wrong solutions provided by the models, which can further mislead model users (Suppl. Sec. E) *3. Standardized benchmarks failure* Standardized benchmarks (MMLU, ARC, etc) are not able to reveal model weaknesses as many models with correct response rate on AIW close to 0 score high on those benchmarks. Also model ranking fails, as standardized benchmarks score is not predictive for whether a model will be better or worse handling AIW (Suppl. Sec. C.4)

4 Discussion & Conclusion

In our work, using a very simple AIW problem (Sec. 2) that requires only elementary set and arithmetic operations and can be easily solved by adults and arguably even children, we observe a striking breakdown of SOTA LLMs performance when confronted with the AIW problem variations (Suppl. Tab. 2). The breakdown is manifested in **(i)** overall low correct response rates (Fig. 1) and **(ii)** strong performance fluctuation across only mild, irrelevant variations of the same problem, which reveals strong lack of robustness and hints at fundamental issues with the generalization capability of the models (Fig. 2). The observed breakdown is in dramatic contrast with claims about strong core functions of SOTA LLMs as backed up by standardized benchmarks, revealing benchmarks failure to properly measure core functions. Specifically, the claim of strong reasoning can be refuted, as any system claiming even basic reasoning should be able to obtain close to 100% correct response rates on problems as simple as AIW problem. The claim of strong zero-shot generalization is refuted by our evidence as well - strong fluctuations of correct response rates across problem irrelevant variations observed in all tested SOTA LLMs reveal deficits to generalize even in a such simple scenario.

Our study should serve as a vivid warning that current SOTA LLMs are not yet capable of strong generalization and robust reasoning, and enabling those is still subject of basic research. AIW problem and its variations offer a starting point and a measurement technique that can reveal lack of robustness and model weaknesses in generalization and core functions that remain undiscovered by current benchmarks (Suppl. Tab. 6). Variations built into problem templates can serve as technique to create new benchmarks that are, in contrast to current common benchmarks, no longer static and can form a basis for new adversarial benchmarks that follow Karl Popper's principle of falsifiability [24], attempting everything to break model's function, highlighting its deficits, and thus showing possible directions for model improvement, which is the way of scientific method.

Acknowledgments

We would like to express gratitude to all the people who are working on making code, models and data publicly available, advancing community based research and making research more reproducible. Specifically, we would like to thank all the members of the LAION Discord server[2] community and Open-Ψ (Open-Sci) Collective[3] for providing fruitful ground for scientific exchange and open-source development.

MN acknowledges funding by the Federal Ministry of Education and Research of Germany under grant no. 01IS22094B WestAI - AI Service Center West. We also acknowledge compute resources provided via WestAI compute grant "Measuring and enhancing advanced reasoning capabilities of foundation models via local model deployment" (westai0007) at Juelich Supercomputing Center (JSC).

LCK acknowledges the Helmholtz Information & Data Science Academy (HIDA) for providing financial support enabling a short-term research stay at Juelich Supercomputing Center (JSC), Research Center Juelich (FZJ) to conduct research on foundation models.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[6] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[7] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[8] 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.

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

[2] `https://discord.gg/BZqhreFazY`

[3] `https://discord.gg/GsKh4mBVcv`

[10] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv*, 7 2023.

[11] 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. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[12] Melanie Mitchell. How do we know how smart ai systems are? *Science*, 381(6654):eadj5957, 2023.

[13] Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv e-prints*, pages arXiv–2307, 2023.

[14] Shahriar Golchin and Mihai Surdeanu. Time travel in llms: Tracing data contamination in large language models. *arXiv preprint arXiv:2308.08493*, 2023.

[15] Changmao Li and Jeffrey Flanigan. Task contamination: Language models may not be few-shot anymore. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18471–18480, 2024.

[16] Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Petersen, and Julius Berner. Mathematical capabilities of chatgpt. *Advances in Neural Information Processing Systems*, 36, 2024.

[17] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

[18] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[19] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[20] Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weiming Lu. Self-contrast: Better reflection through inconsistent solving perspectives. *arXiv preprint arXiv:2401.02009*, 2024.

[21] Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, 12:484–506, 2024.

[22] OpenAI. Gpt-4 turbo and gpt-4 model docs.

[23] Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024.

[24] Karl Raimund Popper. *The logic of scientific discovery*. 1934.

[25] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2020.

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

[27] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439, 2020.

[28] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[29] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WINOGRANDE: an adversarial winograd schema challenge at scale, 2019.

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

[31] Cohere. Introducing command r+: A scalable llm built for business.

[32] Cohere. Command r+ documentation, 2024.

[33] Mosaic. Introducing dbrx: A new state-of-the-art open llm.

[34] OpenAI. Gpt-4o model docs., 2024.

[35] OpenAI. Announcement: Hello gpt-4o, 2024.

[36] OpenAI. Introducing chatgpt, 11 2022.

[37] OpenAI. Models - openai gpt 3.5 turbo docs.

[38] OpenAI. Models - openai gpt 3.5 turbo update.

[39] Anthropic. Claude 3.5 sonnet, 2024.

[40] Anthropic. Introducing the next generation of claude, 2024.

[41] Sundar Pichai and Demis Hassabis. Introducing gemini: Google's most capable ai model yet, 2023.

[42] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[43] Sundar Pichai and Demis Hassabis. Our next-generation model: Gemini 1.5, 2024.

[44] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[45] Google. Gemma: Introducing new state-of-the-art open models, 2024.

[46] Google. Gemma model card, 2024.

[47] Mistral-AI-Team. Mistral ai models api versioning, 2024.

[48] Mistral-AI-Team. Au large | mistral ai | frontier ai in your hands, 2024.

[49] Mistral-AI-Team. Announcement mixtral ai 7x22b, 2024.

[50] Mistral-AI-Team. Announcement mixtral ai 7x8b, 2024.

[51] Mistral-AI-Team. Announcement mistral ai 7b, 2023.

[52] Cohere. Model card for c4ai command r+, 2024.

[53] Meta. Meta and microsoft introduce the next generation of llama | meta, 2023.

[54] Meta. Introducing meta llama 3: The most capable openly available llm to date, 2024.

[55] Meta. Meta llama 3 model card., 2024.

[56] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[57] Alibaba. Introducing qwen1.5, 2024.

[58] Alibaba. Hello qwen2, 2024.

5 Debunking Challenge Submission

5.1 What commonly-held position or belief are you challenging?

SOTA LLMs constitute an important instance of foundation models - models that sustain claim of strong generalization and transferability across novel scenarios and tasks, backed up by scaling laws that accurately predict generalization and function improvement with increasing pre-training scales [4, 5, 6]. These strong claims rely mostly on evaluations performed on standardized benchmarks that argue to properly measure core capabilities like generalization and reasoning, eg. MMLU, ARC, PIQA, HellaSwag, HumanEval, WinoGrande. GSM8k or MATH [25, 26, 27, 28, 29, 30], to name few important examples. Relying on those benchmarks, it is commonly-held position to attribute to SOTA LLMs advanced functions like zero-shot reasoning [7], and in general to put high expectations of strong core functionality on released SOTA LLMs [8, 9, 10, 11]. Such claims extend beyond basic research artifacts and become pervasive in applied industry, where SOTA LLMs are advertised as robust problem solvers for various real world settings, explicitly emphasizing their value as robust reasoners, coders and math solvers, attesting "key business-critical capabilities" or suitability for "real-world enterprise use cases" (see announcements by Cohere on Command R-Plus [31, 32], or by Mosaic on DBRX [33], as only few selected representative examples out of many)

5.2 How are your results in tension with this commonly-held position?

In our work, using a very simple AIW problem (Sec. 2) that can be easily solved by adults and arguably even children, we observe a striking breakdown of SOTA LLMs performance when confronted with the AIW problem and its variations (Suppl. Tab. 2). The breakdown is manifested in **1.** Overall low correct response rates (Fig. 1) and **2.** Strong performance fluctuation across mild, irrelevant variations of the same problem, which hints at fundamental issues with the generalization capability of the models (Fig. 2). The observed breakdown is in dramatic contrast with claims about strong core functions of SOTA LLMs. Specifically, the claim of strong reasoning can be refuted, as any system claiming even basic reasoning should be able to obtain 100% correct response rates on problems as simple as AIW problem. The claim of strong zero-shot generalization is refuted by our evidence as well - strong fluctuations of correct response rates across problem irrelevant variations observed in all tested SOTA LLMs reveal deficits to generalize even in such a simple scenario. By executing control experiments (Suppl. Sec. C.2), we provide evidence (Fig. 3, Suppl. Fig. 5,6) that the observed failures are not specific to the problem type we study and thus hint on generic deficits in generalization and basic reasoning. Breakdown in such a simple setting makes clear that current SOTA LLMs are not to be trusted with handling any real world problem scenario, where same phenomena, for instance severe lack of robustness when facing mild variations as observed here (Fig. 2), are expected to appear. Our study also clearly refutes capability of standardized benchmarks to properly measure core model functionality such as generalization or reasoning (Suppl. Sec. C.4, Fig. 11, Tab. 6). Standardized benchmarks assigning high scores to SOTA LLMs fail to reveal severe model weaknesses made evident by AIW problem testing, also failing to provide proper model ranking, as standardized benchmarks scores turn out to be not predictive for whether a model will be better or worse handling simple AIW problem in comparison to other models.

5.3 How do you expect your submission to affect future work?

We would like to take these initial observations to the scientific and technological community to stimulate urgent re-assessment of the claimed capabilities of current generation of LLMs. Our study should serve as a vivid reminder that current SOTA LLMs are not capable of strong generalization and robust reasoning, and enabling those is still subject of basic research. Such re-assessment also requires common action to create standardized benchmarks that would allow proper detection of such generalization and basic reasoning deficits as observed in our study that obviously manage to remain undiscovered by current state-of-the-art evaluation procedures and benchmarks. Variations built into problem templates can serve as technique to create new benchmarks that are, in contrast to current common benchmarks, no longer static and can serve as better measurement tool for properly testing model generalization and reasoning. New benchmarks should follow Karl Popper's principle of falsifiability [24], attempting everything to break model's function, highlighting its deficits, and thus showing possible directions for model improvement, which is the way of scientific method, also offering protection from overblown claims about models' core functions.

Supplementary.

A Additional overview for performed experiments

Here we gather additional helpful information on the procedures around the executed experiments. To provide overview over origin of core tested models used for the AIW experiments, we list those in Suppl. Tab. 1

Table 1: Names, origin and versioning of core test models used in the experiments.

Name	Origin	Released	Open Weights	Sources
GPT-4o-2024-05-13	OpenAI	13.05.2024	No	[8, 34, 35]
GPT-4-turbo-2024-04-09	OpenAI	09.04.2024	No	[8, 22]
GPT-4-0125-preview	OpenAI	25.01.2024	No	[8, 22]
GPT-4-0613	OpenAI	13.06.2023	No	[8, 22]
GPT-3.5-turbo-0125	OpenAI	24.01.2024	No	[36, 37, 38]
Claude-3-5-sonnet-20240620	Anthropic	21.06.2024	No	[39]
Claude-3-opus-20240229	Anthropic	04.03.2024	No	[40, 23]
Claude-3-sonnet-20240229	Anthropic	04.03.2024	No	[40, 23]
Claude-3-haiku-20240307	Anthropic	04.03.2024	No	[40, 23]
Gemini 1.0 Pro	Google	06.12.2023	No	[41, 42]
Gemini 1.5 Pro	Google	16.02.2024	No	[43, 44]
gemma-7b-it	Google	05.04.2024 (v1.1)	Yes	[45, 46]
gemma-2b-it	Google	05.04.2024 (v1.1)	Yes	[45, 46]
Mistral-large-2402	Mistral AI	26.02.2024	No	[47, 48]
Mistral-medium-2312	Mistral AI	23.12.2023	No	[47, 48]
Mistral-small-2402	Mistral AI	26.02.2024	No	[47, 48]
open-mixtral-8x22b-instruct-v0.1	Mistral AI	17.04.2024	Yes	[47, 49]
open-mixtral-8x7b-instruct-v0.1	Mistral AI	11.12.2023	Yes	[47, 50]
open-mistral-7b-instruct-v0.2	Mistral AI	11.12.2023	Yes	[11, 47, 51]
Command R+	Cohere	04.04.2024	Yes	[32, 52]
Dbrx Instruct	Mosaic	27.03.2024	Yes	[33]
Llama 2 70B Chat	Meta	18.07.2023	Yes	[53, 10]
Llama 2 13B Chat	Meta	18.07.2023	Yes	[53, 10]
Llama 2 7B Chat	Meta	18.07.2023	Yes	[53, 10]
Llama 3 70B Chat	Meta	18.04.2024	Yes	[54, 55]
Llama 3 8B Chat	Meta	18.04.2024	Yes	[54, 55]
Qwen 1.5 1.8B - 72B Chat	Alibaba	04.02.2024	Yes	[56, 57]
Qwen 2 72B Instruct	Alibaba	07.06.2024	Yes	[58]

B Prompt types and variations

For testing the model dependence on input prompt type as well as robustness against problem variations when solving AIW and AIW Light problems, we used three main prompt types - STANDARD (original prompt with answer formatting instructions), THINKING (prompt that encourages thinking with answer formatting instructions) and RESTRICTED (prompt that instructs model to output only formatted answer and nothing else). THINKING v2 prompt type is a minor variation of THINKING type that just adds "step by step" after already existing "think carefully" phrasing (control experiments show that THINKING and THINKING v2 are equivalent in terms of observed performance, so we use both interchangeably). For testing the models' robustness to problem perturbations, we try different variations of main AIW problem (AIW Variations 1-4, see Sec. 2, Suppl. Tab. 2), where we keep the same problem structure while varying numbers of brothers and sisters and their mentioning order within the sentence. Those variations are made intentionally in such a way that they do not affect problem structure or its difficulty and thus should not affect how models cope with the problem.

We employ further AIW versions - AIW Light - as control to test whether models are able to deal with various aspects of original AIW problem, eg handling the specific relational family structure frame, or executing elementary arithmetic operations necessary to solve AIW problem. See Suppl. Sec. C.2.1 for more detail on the AIW Light design.

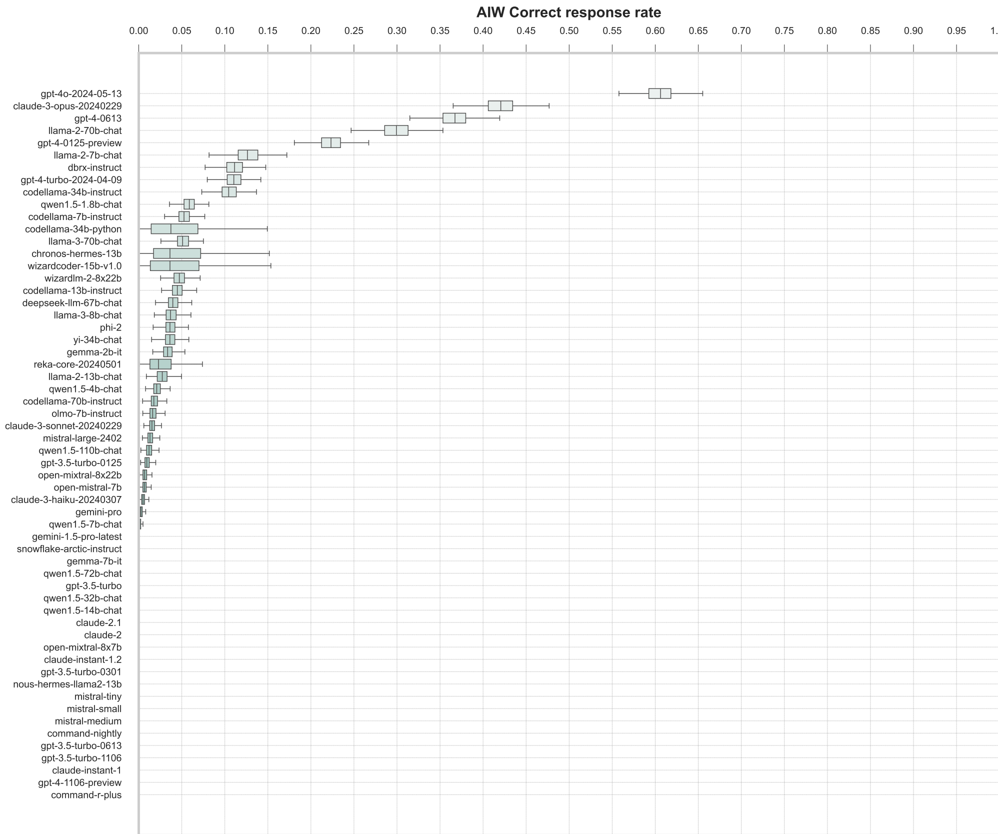

Figure 4: Collapse of most SOTA LLMs on AIW problem. AIW correct response rate, averaged across AIW variations 1-4 with prompt types THINKING and STANDARD. Only 5 models manage to show rates above $p = 0.2$: GPT-4o, Claude 3 Opus, GPT-4-0613, Llama 2 70B Chat and GPT-4-0125-preview (GPT4-Turbo). Llama 2 70B Chat is the only open-weights model in this set. The rest either shows poor performance below $p = 0.15$, or even collapses entirely to 0. Among those models collapsing to 0 are many which are claimed to be strong, eg larger scale GPT-3.5, Mixtral 8x7B and 8x22B, Command R Plus, Qwen 1.5 72B Chat and smaller scale Gemma-7b-it, Mistral Small and Mistral Medium. By inspecting the correct answer responses of the better performers, we indeed see mostly correct reasoning executed to arrive at the final correct answers. For the models that do not perform well and are able to deliver correct answers only rarely, we still see in some of those very rare responses with correct final answer correct proper reasoning, for instance in case of Mistral/Mixtral, Dbrx Instruct, CodeLlama. We see however also responses with a correct final answer, which after careful inspection, turns out to be an accident of executing entirely wrong reasoning, where many accumulating mistakes accidentally lead to the final number corresponding to the right answer. Those wrong-reasoning-right-answer responses are encountered in models that perform poorly ($p < 0.3$) (see Suppl. Sec. D for response examples).

See Suppl. Tab. 2 (for AIW problem) and Suppl. Tab. 3, 4, 5 (for AIW Light problems) for examples with full prompt versions for each presented problem and its variations.

C Model performance and behavior on AIW and AIW Light problem

Here we report further details on model evaluation, performance and behavior as observed on AIW and AIW Light problems. For the full overview of average correct response rate including models that score zero, see Suppl. Fig. 4. For the statistics on number of trials conducted for each model and each prompt type, see Suppl. Fig. 16. For the statistics on the average output length across models and prompt types, see Suppl. Fig. 17.

Table 2: AIW main variations, prompt types and correct answers overview.

Var.	Prompt	Type/Answer	ID
1	Alice has 3 brothers and she also has 6 sisters. How many sisters does Alice's brother have? Solve this problem and provide the final answer in following form: "### Answer: ".	STANDARD / 7	55
1	Alice has 3 brothers and she also has 6 sisters. How many sisters does Alice's brother have? Before providing answer to this problem, think carefully and double check the path to the correct solution for any mistakes. Provide then the final answer in following form: "### Answer: ".	THINKING / 7	57
1	Alice has 3 brothers and she also has 6 sisters. How many sisters does Alice's brother have? To answer the question, DO NOT OUTPUT ANY TEXT EXCEPT following format that contains final answer: "### Answer: ".	RESTRICTED / 7	53
2	Alice has 2 sisters and she also has 4 brothers. How many sisters does Alice's brother have? Solve this problem and provide the final answer in following form: "### Answer: ".	STANDARD / 3	56
2	Alice has 2 sisters and she also has 4 brothers. How many sisters does Alice's brother have? Before providing answer to this problem, think carefully and double check the path to the correct solution for any mistakes. Provide then the final answer in following form: "### Answer: ".	THINKING / 3	58
2	Alice has 2 sisters and she also has 4 brothers. How many sisters does Alice's brother have? To answer the question, DO NOT OUTPUT ANY TEXT EXCEPT following format that contains final answer: "### Answer: ".	RESTRICTED / 3	54
3	Alice has 4 sisters and she also has 1 brother. How many sisters does Alice's brother have? Solve this problem and provide the final answer in following form: "### Answer: ".	STANDARD / 5	63
3	Alice has 4 sisters and she also has 1 brother. How many sisters does Alice's brother have? Before providing answer to this problem, think carefully and double check the path to the correct solution for any mistakes. Provide then the final answer in following form: "### Answer: ".	THINKING / 5	64
3	Alice has 4 sisters and she also has 1 brother. How many sisters does Alice's brother have? To answer the question, DO NOT OUTPUT ANY TEXT EXCEPT following format that contains final answer: "### Answer: ".	RESTRICTED / 5	65
4	Alice has 4 brothers and she also has 1 sister. How many sisters does Alice's brother have? Solve this problem and provide the final answer in following form: "### Answer: ".	STANDARD / 2	69
4	Alice has 4 brothers and she also has 1 sister. How many sisters does Alice's brother have? Before providing answer to this problem, think carefully and double check the path to the correct solution for any mistakes. Provide then the final answer in following form: "### Answer: ".	THINKING / 2	70
4	Alice has 4 brothers and she also has 1 sister. How many sisters does Alice's brother have? To answer the question, DO NOT OUTPUT ANY TEXT EXCEPT following format that contains final answer: "### Answer: ".	RESTRICTED / 2	71
4	Alice has 4 brothers and she also has 1 sister. How many sisters does Alice's brother have? Solve the problem by taking care not to make any mistakes. Express your level of confidence in the provided solution as precisely as possible.	CONFIDENCE / 2	11
3	Alice has 4 sisters and she also has 1 brother. How many sisters does Alice's brother have? To solve the problem, approach it as a very intelligent, accurate and precise scientist capable of strong and sound reasoning. Provide the solution to the problem by thinking step by step, double checking your reasoning for any mistakes, and based on gathered evidence, provide the final answer to the problem in following form: "### Answer: ".	SCIENTIST / 5	40

C.1 Evaluating model responses

To perform evaluations of model performance, we were confronted with the question of how to parse and extract the final answer from the responses provided by the models when confronted with the input containing the AIW problem formulation. On the one hand, it should be possible to deal with any response provided by the model as a solution, while on the other hand allowing to extract a clear final answer as a number to be compared with the correct answer, such that for each model response a decision can be made on whether the provided response was right or wrong. To be able to keep the parsing procedure simple (without involving for instance another suitable LLM prompting it to extract the relevant part of response), we have chosen to add to the problem prompt following passage: *"provide the final answer in following form: "### Answer: ""* (see also Suppl. Tab. 2). We observed that all models we have chosen to test (Suppl. Tab. 1) were able to follow such an instruction, providing a response that could be easily parsed. We also ran control experiments without such formatting instruction in the problem formulation, ensuring that behavior does not depend on it.

The presented prompt extension makes it possible to extract for each prompting trial whether a model has provided a correct answer to the AIW problem posed in the input. We can interpret then any number n of collected responses as executing n trials given a particular prompt for a given model (n - number of Bernoulli trials), observing in each $i-$th trial a Bernoulli variable $X_i = \{0, 1\}$. We interpret the number of correct responses $X = \sum_i X_i$ as random variable following a Beta-Binomial distribution with unknown probability p of correct response that we also treat as random variable that comes from a Beta distribution, i.e. $p \sim Beta(\alpha, \beta)$, where α and β are parameters of the Beta distribution. To obtain plots showing correct response ratios, we would like to estimate Beta distribution underlying p, and for that, we first estimate the mean of p and its variance from the collected observations. To estimate $\hat{p}$, we use the formula for estimating the mean of p for a binomial distribution: $\hat{p} = X/n$ (i.e. as a proportion of successes). We can report the estimate $\hat{p}$ as the estimate of the correct response rate of a given model and also, compare the correct response rates of various tested models. Moreover, we can estimate the variance of the probability of a correct response by using the following formula:

$$\text{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(X_i) = \frac{n\text{var}(X_i)}{n^2} = \frac{\text{var}(X_i)}{n} = \frac{p(1-p)}{n} \tag{1}$$

The estimates of the variance and the standard deviation of p can be thus obtained by using $\hat{p}$ as $\frac{\hat{p}(1-\hat{p})}{n}$ and $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ respectively. Using the estimated variance and mean of p, we can use the following relations for the variance: $\left(\sigma^2 = \frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)}\right)$ and the mean $\left(\mu = \frac{\alpha}{\alpha+\beta}\right)$ in order to obtain α and β parameters for the Beta distribution. To simulate data for the plots, we draw N random samples corresponding to correct and incorrect responses using the estimated distribution of p and obtain the plots showing performance on the task for various models of interest as a full distribution of the respective p.

C.2 Control experiments with AIW Light problems

C.2.1 AIW Light problems

To confirm that models do not struggle either with basic family relations structure handling nor with executing arithmetic operations in frame of posed AIW problem, we make various control versions of AIW problem - AIW Light Family, AIW Light Arithmetic Siblings and AIW Light Arithmetic Total Girls. The AIW Light problems keep problem template close to the original, changing only the final question part which poses a modified question to test particular operations. The variations 1-4 are created in the same way like in AIW original by varying natural numbers of brothers and sisters, while ensuring that the natural numbers for final correct answers in AIW original and AIW Light are matched across variations 1-4.

AIW Light Arithmetic Siblings. AIW Light Arithmetic Siblings has following problem template: ***"Alice has N brothers and she also has M sisters. How many siblings does Alice have?"***. Compared to AIW original, only question part is modified. To solve the problem, summing up already given numbers of brothers and sisters is sufficient - the correct answer is $C = N + M$. This requires basic grasping of relational family structure (realizing Alice's siblings are her sisters and brothers) and

Table 3: AIW Light Arithmetic Siblings variations

Var.	Prompt	Type/Answer	ID
1	Alice has 3 brothers and she also has 4 sisters. How many siblings does Alice have? Before providing answer to this problem, think carefully step by step and double check the path to the correct solution for any mistakes. Provide then the final answer in following form: "### Answer:".	THINKING v2 / 7	277
2	Alice has 2 sisters and she also has 1 brother. How many siblings does Alice have? Before providing answer to this problem, think carefully step by step and double check the path to the correct solution for any mistakes. Provide then the final answer in following form: "### Answer:".	THINKING v2 / 3	278
3	Alice has 4 sisters and she also has 1 brother. How many siblings does Alice have? Before providing answer to this problem, think carefully step by step and double check the path to the correct solution for any mistakes. Provide then the final answer in following form: "### Answer: ".	THINKING v2 / 5	279
4	Alice has 1 brother and she also has 1 sister. How many siblings does Alice have? Before providing answer to this problem, think carefully step by step and double check the path to the correct solution for any mistakes. Provide then the final answer in following form: "### Answer:".	THINKING v2 / 2	280

Table 4: AIW Light Family variations

Var.	Prompt	Type/Answer	ID
1	Alice has 7 brothers and she also has 3 sisters. How many brothers does Alice's sister have? Before providing answer to this problem, think carefully step by step and double check the path to the correct solution for any mistakes. Provide then the final answer in following form: "### Answer:".	THINKING v2 / 7	271
2	Alice has 4 sisters and she also has 3 brothers. How many brothers does Alice's sister have? Before providing answer to this problem, think carefully step by step and double check the path to the correct solution for any mistakes. Provide then the final answer in following form: "### Answer:".	THINKING v2 / 3	272
3	Alice has 2 sisters and she also has 5 brothers. How many brothers does Alice's sister have? Before providing answer to this problem, think carefully step by step and double check the path to the correct solution for any mistakes. Provide then the final answer in following form: "### Answer: ".	THINKING v2 / 5	273
4	Alice has 2 brothers and she also has 3 sisters. How many brothers does Alice's sister have? Before providing answer to this problem, think carefully step by step and double check the path to the correct solution for any mistakes. Provide then the final answer in following form: "### Answer:".	THINKING v2 / 2	274

selection and execution of elementary arithmetic sum operation. In contrast to AIW original, it does not require execution of set operations nor binding sex attribute to Alice to properly assign her to correct sets. Should the issues with solving AIW original be rooted in selection and execution of elementary arithmetic operations in family frame, we should see models also failing here. Again, we create variations 1-4 by varying natural numbers N, M, such that correct responses C are matched with AIW original variations 1-4 (Suppl. Tab. 3)

AIW Light Family. AIW Light Family has following problem template: ***"Alice has N brothers and she also has M sisters. How many brothers does Alice's sister have?"***. Compared to AIW original, only question part is modified. To solve the problem, reporting already given number of brothers is sufficient - the correct answer is $C = N$. This requires only basic grasping of relational family structure (understanding entity "Alice's sister", binding female attribute to Alice and realizing Alice and her sisters share same brothers). It does NOT require execution of any arithmetic or set operations, in contrast to AIW original. Should the issues with solving AIW original be rooted in handling basic family structure, we should see models also failing here. Again, we create AIW Light Family variations 1-4 by varying natural numbers N, M, such that correct responses C are matched with AIW original variations 1-4. (Suppl. Tab. 4)

Table 5: AIW Light Arithmetic Total Girls variations

Var.	Prompt	Type/Answer	ID
1	Alice has 6 sisters and she also has 3 brothers. How many girls are there in total? Before providing answer to this problem, think carefully step by step and double check the path to the correct solution for any mistakes. Provide then the final answer in following form: "### Answer:".	THINKING v2 / 7	343
2	Alice has 2 sisters and she also has 4 brothers. How many girls are there in total? Before providing answer to this problem, think carefully step by step and double check the path to the correct solution for any mistakes. Provide then the final answer in following form: "### Answer:".	THINKING v2 / 3	344
3	Alice has 4 sisters and she also has 1 brother. How many girls are there in total? Before providing answer to this problem, think carefully step by step and double check the path to the correct solution for any mistakes. Provide then the final answer in following form: "### Answer: ".	THINKING v2 / 5	345
4	Alice has 1 sister and she also has 4 brothers. How many girls are there in total? Before providing answer to this problem, think carefully step by step and double check the path to the correct solution for any mistakes. Provide then the final answer in following form: "### Answer:".	THINKING v2 / 2	346

AIW Light Arithmetic Total Girls. AIW Light Arithmetic Total Girls has following problem template: ***"Alice has N brothers and she also has M sisters. How many girls are there in total?"***. Compared to AIW original, only question part is modified. To solve the problem, it is necessary to bind female attribute to Alice via the pronoun "she", to assign correct female attributes to the sisters and to execute the correct arithmetic sum operation adding all the obtained girls - the correct answer is $C = M + 1$. This requires basic grasping of family structure (realizing who are the girls in the family) and selection and execution of elementary arithmetic sum operation. In contrast to AIW original, it does not require execution of set operations to properly assign Alice to sisters set. Should the issues with solving AIW original be rooted in binding correct sex attributes or counting total members of particular sex in family frame given its structure, we should see models also failing here. Again, we create variations 1-4 by varying natural numbers N, M, such that correct responses C are matched with AIW original variations 1-4. (Suppl. Tab. 5)

C.2.2 Results of control experiments with AIW Light problems

AIW Light Arithmetic Siblings. We show tested models' performance in Suppl. Fig. 5. While all tested models clearly have struggled with AIW original (Fig. 1, Suppl. Fig. 4), we observe them successfully solving AIW Light Arithmetic Siblings. Correct response rates go high up close to 1 for most tested models across all variations 1-4. This is also the case for the models that show very low correct response rates close to 0 or 0 on AIW original. like Command R+ or Dbrx Instruct. Strong fluctuations we observe across variations on AIW original (Fig. 1, 2) also disappear. This clearly demonstrates that models neither struggle with basic grasping of relational family structure - realizing Alice's siblings are her sisters and brothers, nor with selection and execution of elementary arithmetic sum operation.

AIW Light Family. We show tested models' performance in Suppl. Fig. 6. Also here we observe all the tested models that are struggling with AIW original successfully solving AIW Light Family. Correct response rates go high up close to 1 for most tested models across all variations 1-4. This is also the case for the models that show very low correct response rates close to 0 or 0 on AIW original. like Command R+ or Dbrx Instruct. Also strong fluctuations that we observe across variations on AIW original (Fig. 1, 2) disappear. This clearly demonstrates that models handle well basic grasping of relational family structure - understanding entity "Alice's sister", binding female attribute to Alice and realizing Alice and her sisters share same brothers.

AIW Light Arithmetic Total Girls. We show tested models' performance in Fig. 3. Again, we observe also here strong performance for all tested models that clearly have struggled with AIW original. Correct response rates go high up close to 1 for most tested models across all variations 1-4. This is also the case for the models that show very low correct response rates close to 0 or 0 on AIW original. like Command R+ or Dbrx Instruct. Also strong fluctuations that we observe across

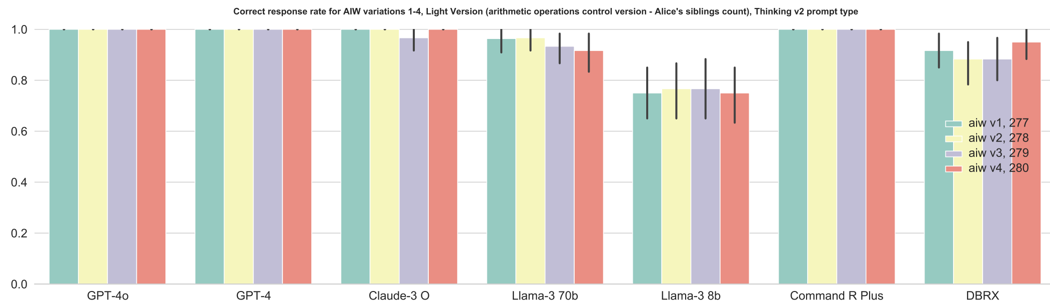

Figure 5: Correct response rates across AIW Light Arithmetic Siblings problem variations 1-4 (THINKING prompt type). Tested models show strong performance across problem variations (a color per each variation 1-4). This shows ability of models to handle basic arithmetic operations within family relations structure of AIW problem, pointing that the arithmetic operations are not an issue. Models that entirely collapse on AIW, like Command R Plus and Dbrx Instruct, are clearly able to solve this version, with correct response rates going up to 1 or close to 1 across all problem variations. For each AIW variation, 60 trials were executed to estimate correct response rate and its variance.

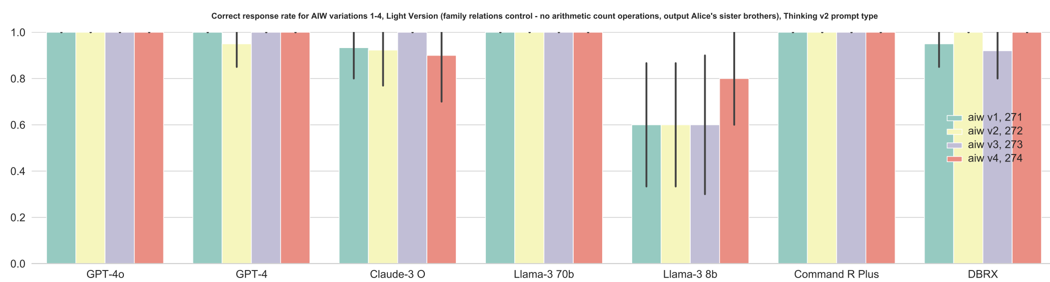

Figure 6: Correct response rates across AIW Light Family problem variations 1-4 (THINKING prompt type). Tested models show strong performance across problem variations (a color per each variation 1-4). This shows ability of models to handle basic family relations structure of AIW problem types, pointing that the specific family related problem structure is not an issue. Models that entirely collapse on AIW, like Command R Plus and Dbrx Instruct, are clearly able to solve this version, with correct response rates going up to 1 or close to 1 across all problem variations. For each AIW variation, 60 trials were executed to estimate correct response rate and its variance.

variations on AIW original (Fig. 1, 2) are gone. This clearly demonstrates that models successfully cope with binding female attribute to entity of Alice, handle assignment of correct female attributes to the sisters and select and execute the correct arithmetic sum operation adding all the girls together.

From these control experiments, we are thus able to obtain strong evidence that all tested models do not suffer from low-level issues with tokenization and natural language or natural numbers parsing and can handle well basic family relations structure and selection and execution of elementary arithmetic operations necessary to solve AIW problem. This further strengthen the hypothesis that observed failures and strong fluctuations in all tested SOTA models on AIW problem (Fig. 1, 2) are rooted in problem unspecific, generic deficits in generalization and basic reasoning.

C.3 Frequency distribution of natural numbers on output and dominance of wrong responses.

To shed more light on modes of correct or wrong responses provided by the models when confronted with AIW problem variations, we show here frequency distribution for natural numbers on the output for AIW variations with higher and lower correct response rates.

As evident from the plots, in higher performance AIW variations (Suppl. Fig. 8), dominants peaks are often positioned on correct answer C=M+1, while for lower performance AIW variations (Suppl. Fig. 7), dominant peaks fall on wrong answer M. Further, for weaker models, distribution broadens, covering more numbers (eg in Llama 3 8b), while for better performers, responses concentrate on

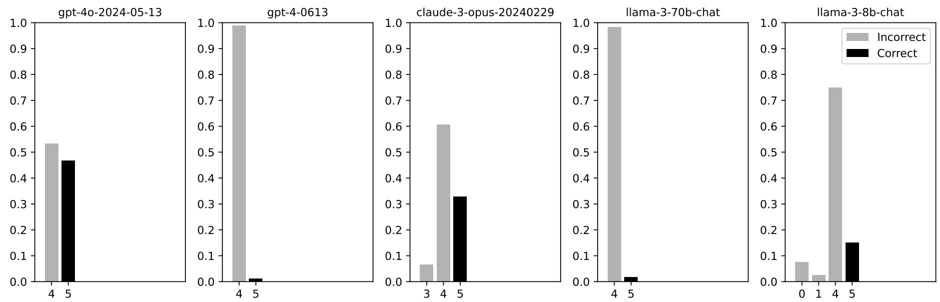

Figure 7: Frequency distribution of output numbers in models' responses. Shown are numerical outputs for AIW Variation 3, THINKING prompt type (prompt ID 64), that has correct answer C=M+1=5, with M=4 number of sisters of Alice. For this AIW variation, models have low performance (see also Figure D.). Correspondingly, peaks are on the dominant wrong response, R=M=4. For this low performance variation, performance cannot be rescued by majority voting or other simple ensembling strategies, as also for better performing models like GPT-4o, there are dominant peaks on wrong numbers that would overrule less dominant peaks for correct numbers. Weaker models, eg Llama 3 8B, show also broader distribution. Distributions were computed over 60 trials executed for each model, taken from original collected responses data.

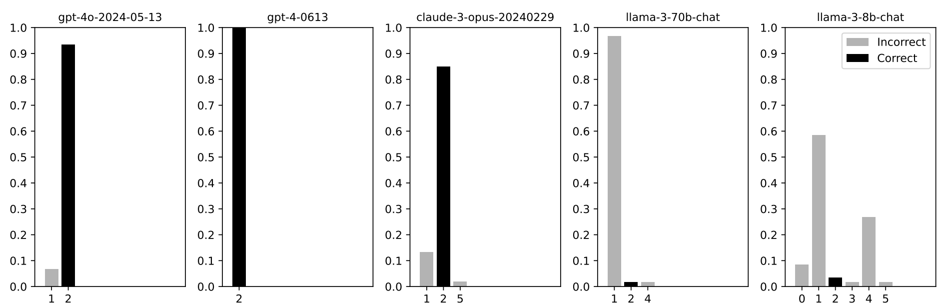

Figure 8: Frequency distribution of output numbers in models' responses. Shown are numerical outputs for AIW Variation 4, THINKING prompt type (prompt ID 70), that has correct answer C=M+1=2, with M=1 number of sisters of Alice. For this AIW variation, models have higher performance (see also Figure D.). Correspondingly, peaks for better performing models (eg GPT-4o, GPT-4, Claude Opus 3) are on the dominant correct response, R=M+1=2. For models with worse performance, peaks are on the dominant wrong response, R=M=1. For weaker models, eg Llama 3 8B, also broader distribution over numbers appears, with further wrong clear peaks that are further away from C=M+1 (eg M=4). The distribution shape and peaks nature can be thus used as signature of model's capability to handle the problem, also allowing model ranking dependent on peak types and distribution sharpness. Distributions were computed over 60 trials executed for each model, taken from original collected responses data.

M and M+1, peaking on correct or wrong answer on depending on AIW variation. Remarkably, for lower performance AIW variations (Suppl. Fig. 7), performance cannot be rescued by major voting or by similar ensemble like strategies, as peaks on wrong response numbers dominate clearly peaks on numbers for correct responses, which would still correspond to committing wrong answer when performing majority voting.

For the AIW Light problem versions used in control experiments, we observe as expected clear dominant peaks on the numbers corresponding for correct responses across all tested models (Suppl. Fig. 9, 10), as AIW Light problems are successfully solved across all their variations.

We note that distribution characteristics, eg concentration on numbers around the correct answer, height of the peaks, can be a further signature that reflects model's capability to handle the problem. More capable models retain dominant peaks on number corresponding to correct answer with smaller peaks on neighboring numbers, while weak models have large peaks on numbers corresponding to

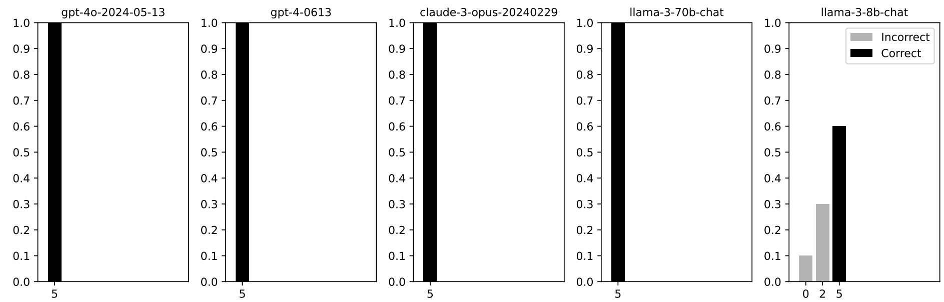

Figure 9: Frequency distribution of output numbers in models' responses. Shown are numerical outputs for AIW Light Family, Variation 3, THINKING prompt type (prompt ID 273), that has correct answer C=5 (number of Alice's brothers). For this AIW Light version, all models have high performance. Correspondingly, peaks are on the dominant correct response, R=5. However also here, weaker models like Llama 3 8B show broader distribution with non-vanishing peaks besides the correct response (eg R=0, R=2) hinting on their weaker capabilities to deal robustly with the problem. Distributions were computed over 60 trials executed for each model.

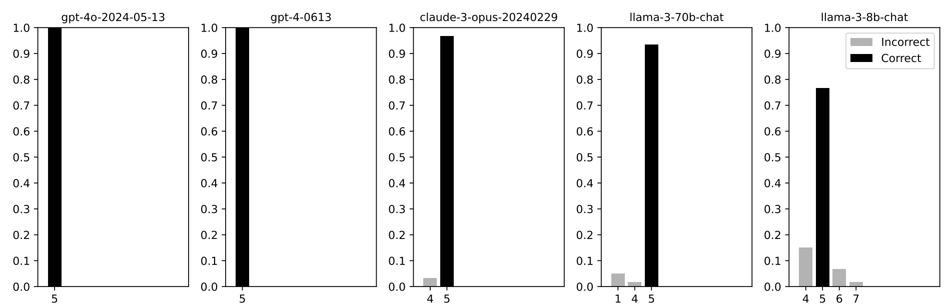

Figure 10: Frequency distribution of output numbers in models' responses. Shown are numerical outputs for AIW Light Arithmetic, Variation 3, THINKING prompt type (prompt ID 279), that has correct answer C=5 (total number of Alice's siblings). For this AIW Light version, all models have high performance. Correspondingly, peaks are on the dominant correct response, R=5. However also here, weaker models like Llama 3 8B show broader distribution with non-vanishing peaks besides the correct response (eg R=4, R=6) hinting on their weaker capabilities to deal robustly with the problem. Distributions were computed over 60 trials executed for each model.

wrong answers or in general broad distribution across all natural numbers below 10. Computing scores from distribution shape can thus also enable model ranking.

C.4 Standardized benchmarks failure.

We observe failure of standardized reasoning benchmarks to properly reflect generalization and basic reasoning skills of SOTA LLMs by noting significant disparity between the model's performance on the AIW problem and the outcomes on conventional standardized benchmarks.

All of the tested models report high scores on various standardized benchmarks that claim to test reasoning function, e.g. MMLU, ARC, Hellaswag, to name a few. Clearly, our observations of SOTA models breaking down on the simple AIW problem hint that those benchmarks do not reflect deficits in generalization and basic reasoning of those models properly. We visualize this failure by plotting performance of tested models that are reported to obtain on wide-spread and accepted standardized benchmarks like MMLU versus the performance we observe on our proposed AIW problem. As strikingly evident from Fig. 11, there is a strong mismatch between high scores on MMLU reported by the models and the correct response rates they obtain on AIW.

This miscalibration with respect to standardized benchmarks makes it impossible to predict for any given model from its scores on MMLU whether the model will be also able to solve the simple AIW

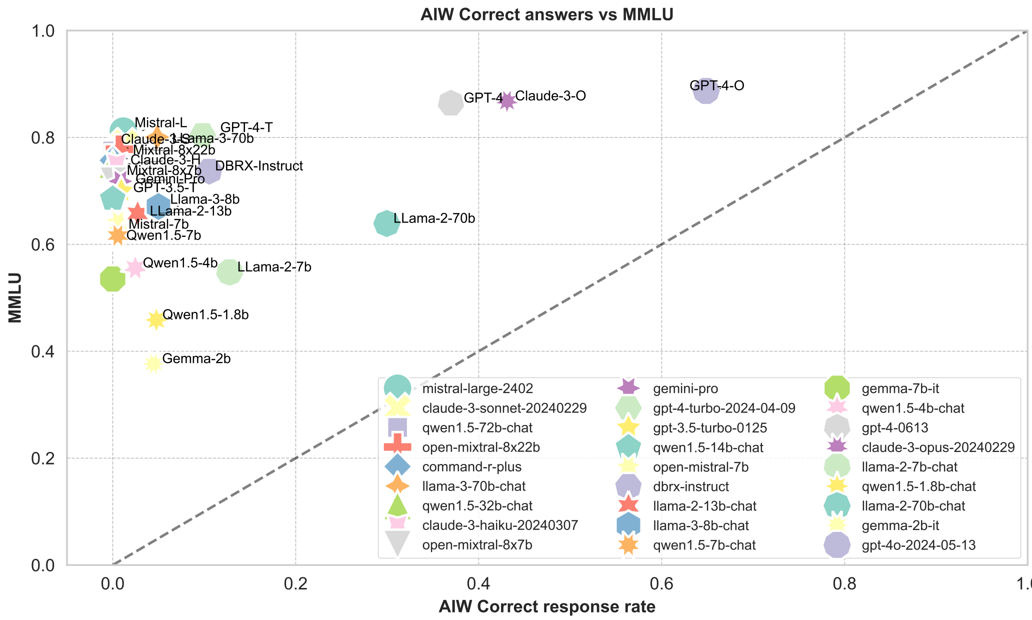

Figure 11: Failure of standardized benchmark MMLU to properly reflect and compare model basic reasoning capabilities as shown by strong discrepancy between AIW correct response rate vs MMLU average score. Many models, eg. Command R+, score 0 on AIW, but have high MMLU score.

on the one hand. On the other hand, it deranges any model comparison based on the standardized benchmark, as models with higher MMLU can have complete breakdown in AIW, while models with lower MMLU can have some non-negligible AIW performance. A clear example of such case is comparison of LLama 2 70B and models claiming such very strong performance via standardized benchmarks like Command R+ or Dbrx. Those models, while claiming much stronger function via scoring substantially higher on MMLU (and other standardized benchmarks) than LLama 2 70B, undergo severe breakdown on AIW - e.g. Command R+ is unable to solve a single AIW problem instance (see also Suppl. Tab. 6).

These observations are also valid on further standardized reasoning benchmarks like MATH, ARC-c, GSM8K and Hellaswag (Suppl. Tab. 6). We provide plots visualizing failure of these standardized benchmarks, reflected in strong mismatch between high benchmark scores reported by many models and the low correct response rates they obtain on AIW (which in some cases is 0 for models with high standardized benchmark scores), in Figures 12, 15, 13, 14.

We see thus that standardized benchmarks fail to properly reflect true model capabilities to generalize and reason - the majority of the tested models score high on standardized benchmarks, suggesting strong function, while showing extreme low correct response rates on simple AIW problem. Many of the models with high scores on standardized benchmarks cannot solve AIW problem a single time. This discrepancy refutes the claim of standardized benchmarks to measure correctly current models' core functionality.

D Examples of correct and failed responses

We provide all collected model responses we obtained during this study in the collected_responses folder in the AIW repo. Here we also showcase some correct and incorrect answers as an example (see Figs. 18, 21, 19, 20).

E Confabulations and overconfident tone accompanying wrong answers

Overconfident tone. In ideal scenario, if LLM cannot correctly solve the AIW problem, it should at least be capable of expressing high uncertainty about the provided incorrect solution to the user. We

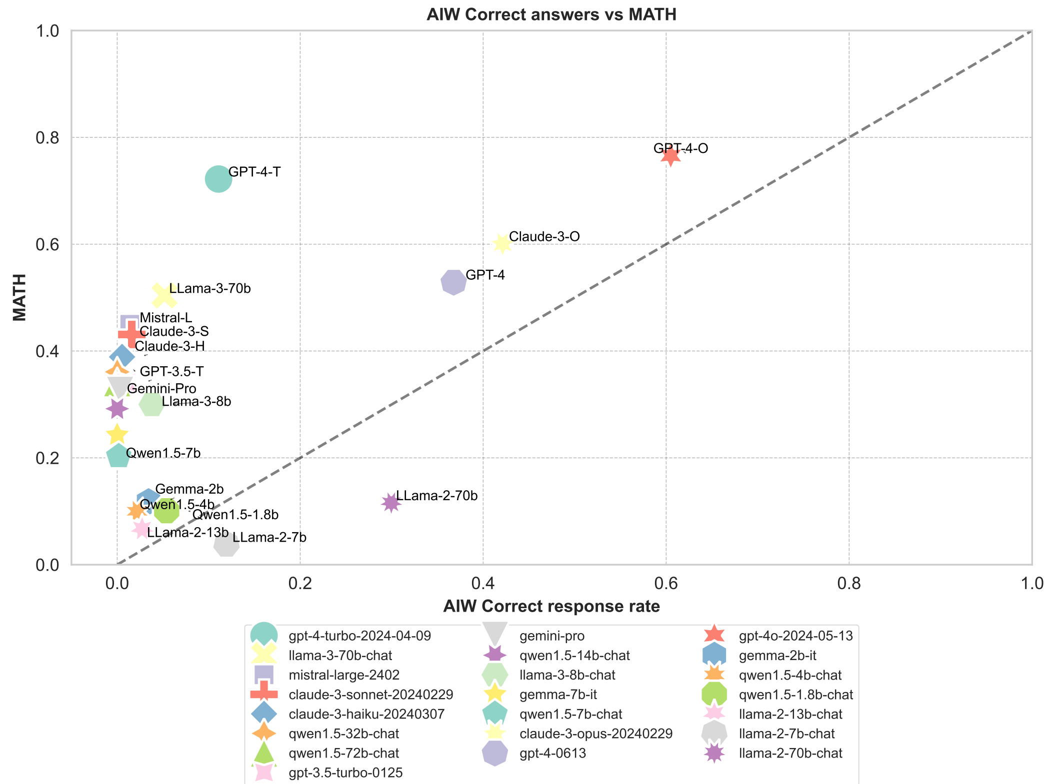

Figure 12: Discrepancy between the AIW correct response rate and the MATH average score, indicating the limitation of standardized benchmark MATH in accurately assessing and comparing basic reasoning capabilities of models. Numerous models, such as Command R+, exhibit a stark contrast in performance, scoring zero on AIW while achieving high scores on MATH.

used CONFIDENCE prompt type (see Suppl. Tab. 2) for AIW problem to see how confident tested models are in their wrong solutions.

From our experiments we can see that LLMs most of the time express high certainty even if their answers are completely wrong, thus mediating strong confidence (see Fig. 22). The models also use highly persuasive tone to argue for the expressed certainty and correctness of the provided wrong solutions, using words like "highly confident", "definitive answer", or "accurate and unambiguous". We see also strong overconfidence expressed in multi-turn interactions with models, where user is insisting on solution provided being incorrect, and observe there high resistance of models to revise their decisions, which was already referred to as "stubbornness" in other works [20] (see Suppl. Sec. F and also data provided in the AIW repo)

Confabulations. In our experiments we observe frequent tendency of those tested models that show strong reasoning collapse and produce frequent wrong answers for AIW problem to generate at the same time persuasive sounding pseudo-explanations to back up their incorrect answers. We term here such pseudo-explanations confabulations, and present a selection of those as examples.

Such confabulations can contain mathematical calculations or other logic-like expressions and operations that make little or absolutely no sense given the problem to be solved, see examples for Olmo-7B, Fig. 23 and Command R+, Fig. 25.

Further confabulations make use of various social and cultural norm specific context to argue for the posed problem to be inappropriate to solve or to provide non-sense arguments for various incorrect answers. There are many such examples that we have observed, we present here only a small selection.

CodeLlama-70B-instruct for instance seems to be specifically prone to claim ethical or moral reasons for not addressing the problem correctly, in the presented example inventing out of nowhere a person with Down syndrome and then pointing out that question has to be modified to be addressed due to

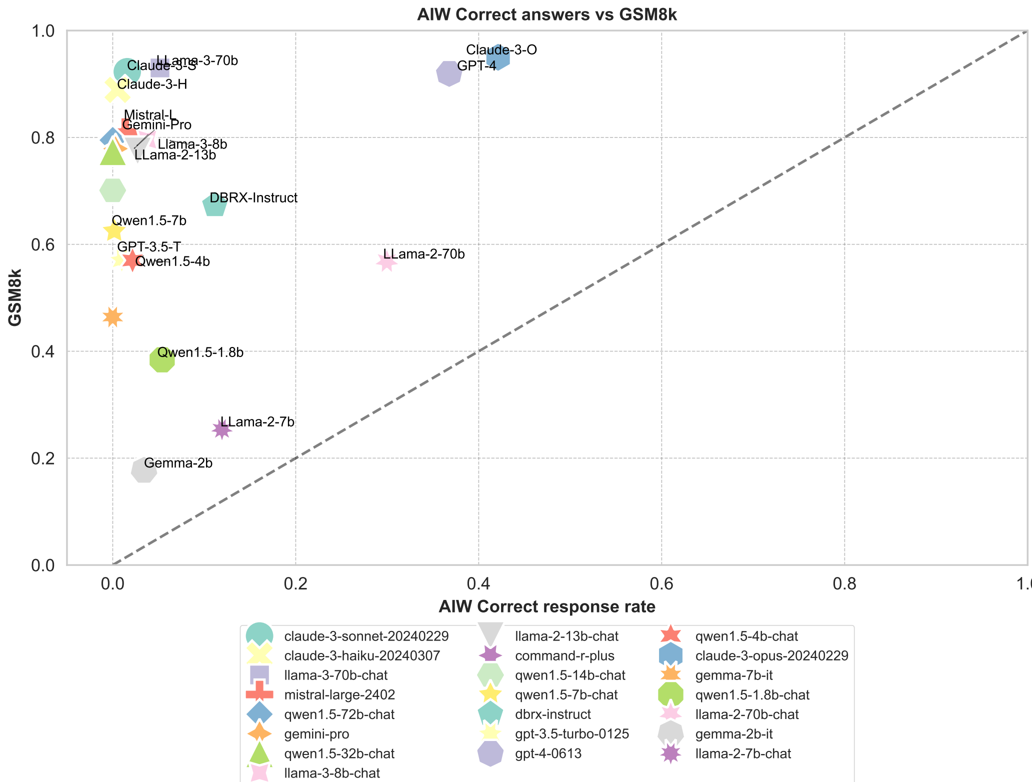

Figure 13: Limitation of the standardized benchmark GSM8k in accurately reflecting and comparing basic reasoning capabilities of models, as illustrated by the stark discrepancy between the AIW correct response rate and the GSM8k average score. Notably, the majority of tested models exhibit low performance on AIW problems while achieving relatively high scores on GSM8k, a graduate-level math benchmark for large language models. Among models with slightly better calibration are Claude Opus and GPT 4 that outperform other models on AIW, which coincides with their high GSM8k scores. Llama 2 70b also shows better calibration, where its modest AIW performance matches its modest GSM8k score. In contrast, models like Mistral Large, Gemini Pro, Dbrx Instruct, or Command R+, while scoring high on GSM8k, show breakdown on AIW (Command R+ has 0 correct response rate, Mistral Large and Gemini Pro 0.01, Dbrx Instruct 0.11, see also Suppl. Tab. 6)

potential perpetuation of harm towards individuals or groups, which has nothing to do with original task, Fig. 24.

Another example are confabulations provided by Command R Plus. These confabulations use concepts of gender identity such as non-binary gender or concepts related to inclusion or to cultural context dependent family identification in the provided wrong reasoning leading to incorrect answers. In the attempt to solve the problem, the model first fails to provide obvious common sense solution and then goes on to describe potential scenarios where brothers and sisters may self-identify as non-binary, although providing information on brothers and sisters in the problem usually means via common sense that those persons self-identify correspondingly to their known status as brother or sister (while Alice is clearly identified via "she" pronoun). Model thus clearly fails to grasp that problem structure has nothing to do with the social and cultural norms. The solutions derived by the model from considering those factors that are far beyond Occam's razor and common sense inherent to the simple AIW problem all lead to wrong answers and generate more confusion, while again keeping the persuasive tone that suggests that model is on some right path to provide the correct solutions (Fig. 26)

For more illustrative examples, see the raw data on interactions with the models collected in AIW repo)

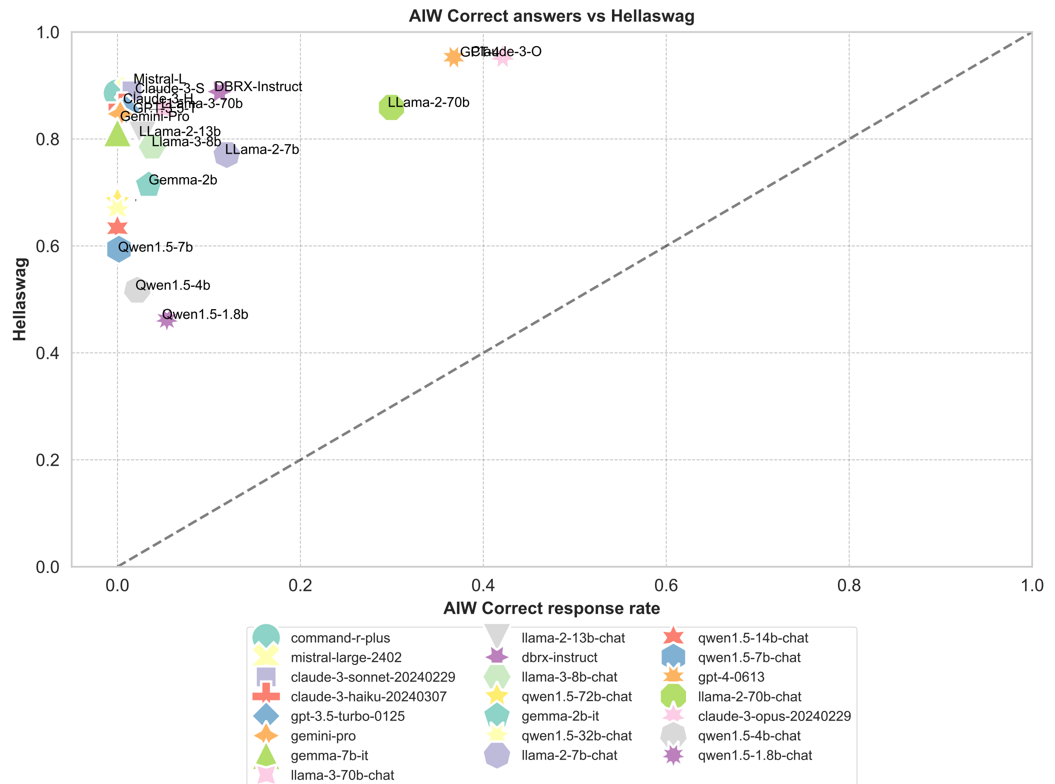

Figure 14: Limitation of the standardized benchmark Hellaswag in accurately assessing and comparing basic reasoning capabilities of models, as evidenced by the significant discrepancy between the AIW correct response rate and the Hellaswag average score.

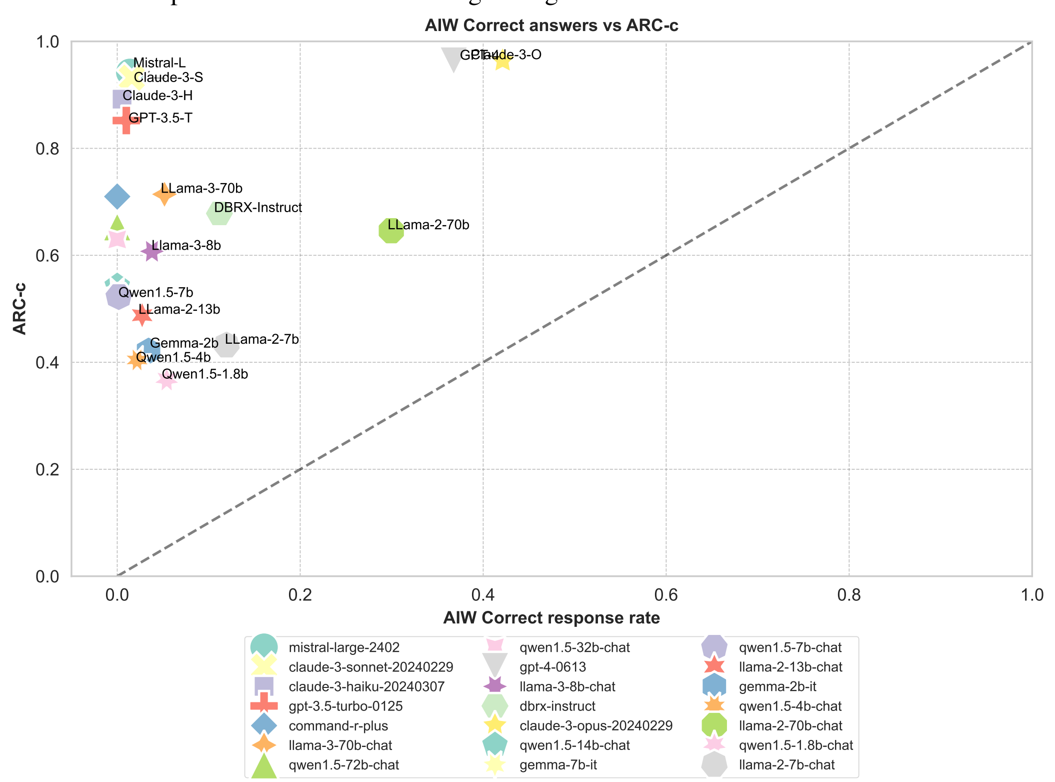

Figure 15: Failure of standardized benchmark ARC-c to properly reflect and compare model basic reasoning capabilities as shown by strong discrepancy between AIW correct response rate vs ARC-c average score.

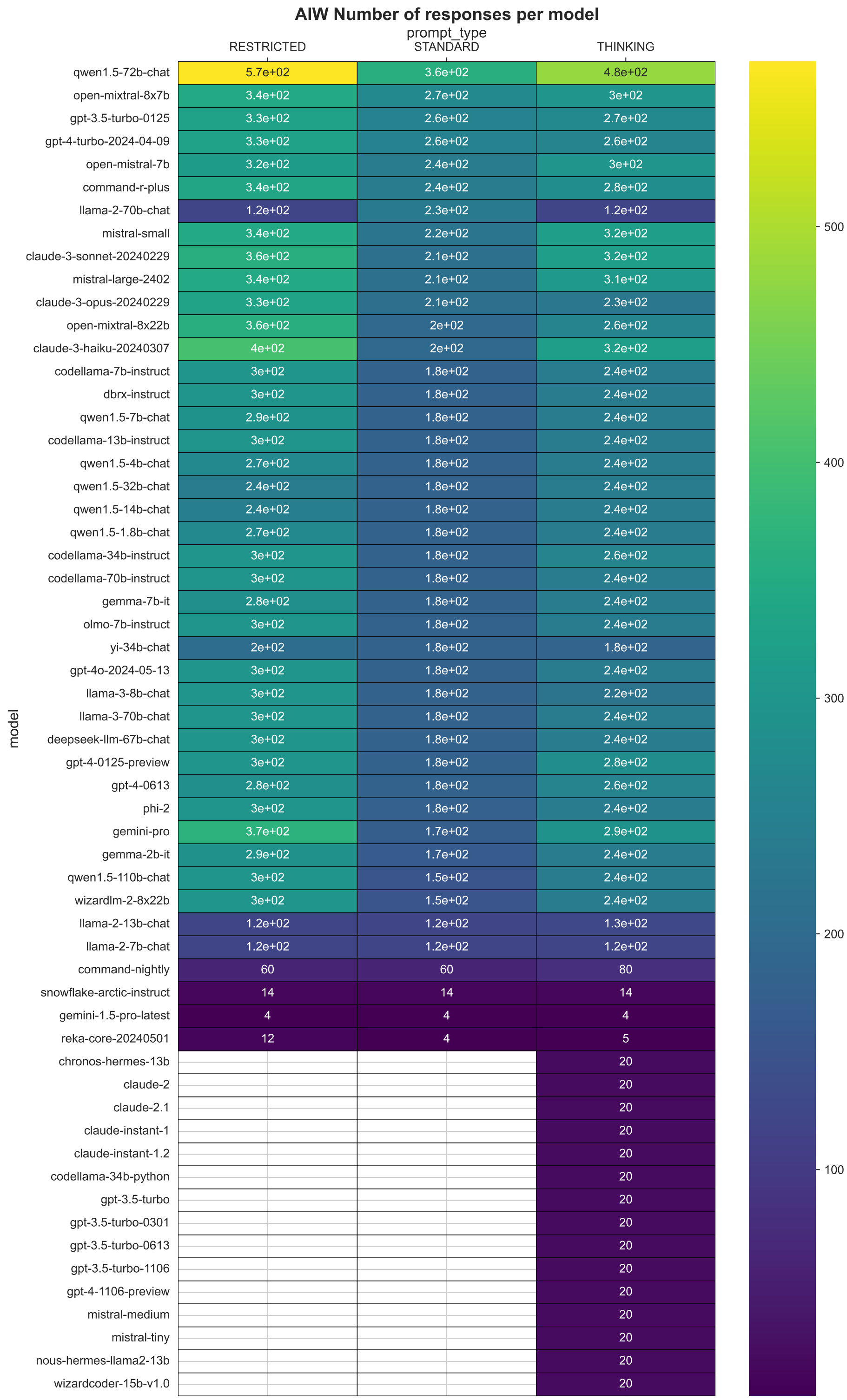

Figure 16: AIW Average number of responses per model for each prompt type (4 AIW variations per prompt type.). Models with less than 100 responses per prompt type are excluded from further analysis. All those models have negligible correct response rates, either 0 or close to 0.

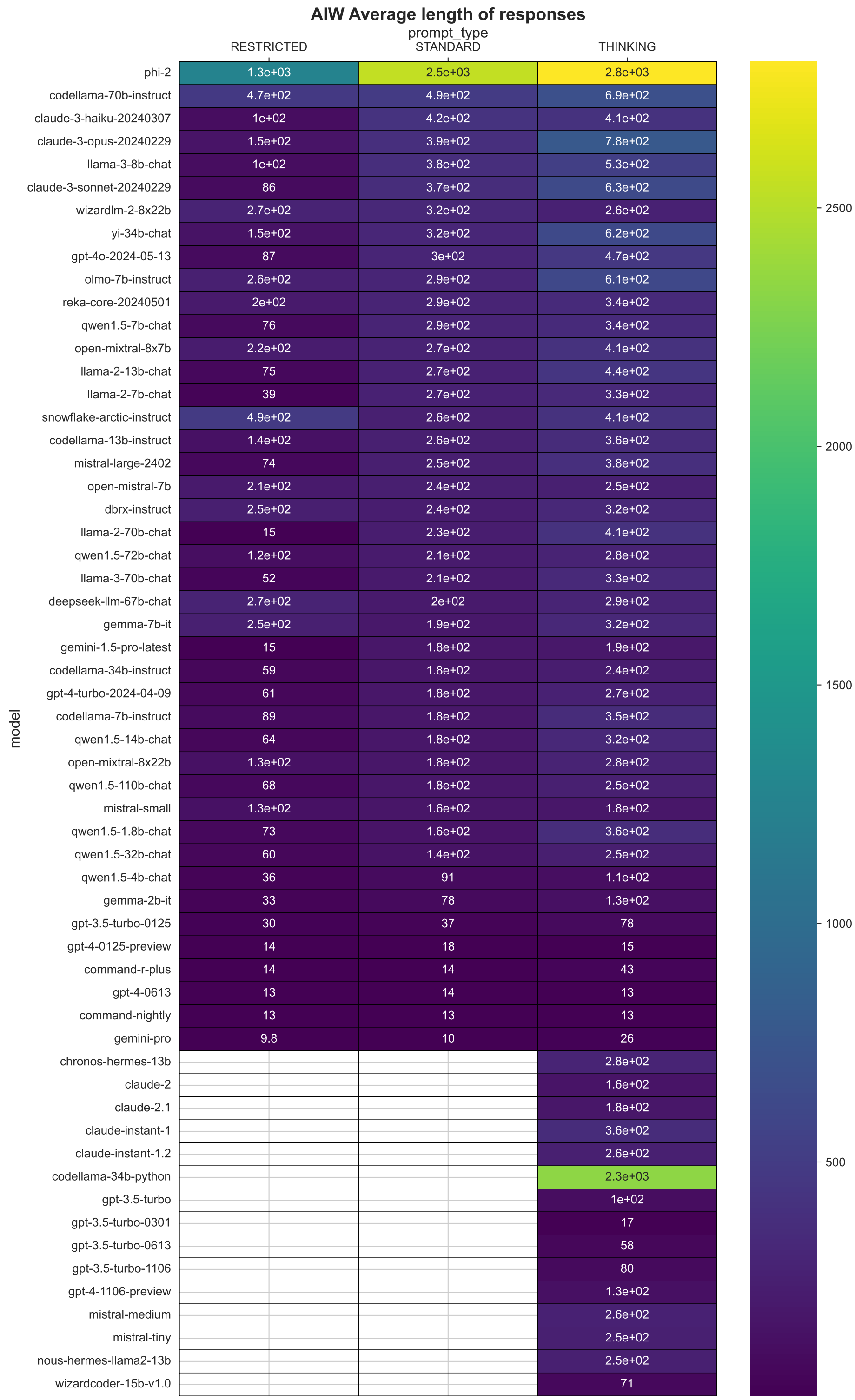

Figure 17: Average length (on characters) of responses per model for each AIW prompt variation. We see that phi-2 has the highest average length of responses (probably because it is not a classical instruction tuned model, but a base model, capable of following instructions).

Table 6: Performance of tested models on MMLU, Hellaswag, ARC-c, GSM8k and AIW problems.

Model	MMLU	Hellaswag	ARC-c	GSM8k	Correct resp. rate (AIW)
gpt-4o-2024-05-13	0.89	-	-	-	0.65
claude-3-opus-20240229	0.87	95.40	96.40	95.00	0.43
gpt-4-0613	0.86	95.30	96.30	92.00	0.37
llama-2-70b-chat	0.64	85.90	64.60	56.80	0.30
llama-2-7b-chat	0.55	77.10	43.20	25.40	0.13
dbrx-instruct	0.74	88.85	67.83	67.32	0.11
gpt-4-turbo-2024-04-09	0.80	-	-	-	0.10
llama-3-8b-chat	0.67	78.55	60.75	79.60	0.05
llama-3-70b-chat	0.80	85.69	71.42	93.00	0.05
qwen1.5-1.8b-chat	0.46	46.25	36.69	38.40	0.05
gemma-2b-it	0.38	71.40	42.10	17.70	0.04
llama-2-13b-chat	0.66	80.70	48.80	77.40	0.03
qwen1.5-4b-chat	0.56	51.70	40.44	57.00	0.02
claude-3-sonnet-20240229	0.79	89.00	93.20	92.30	0.01
mistral-large-2402	0.81	89.20	94.20	81.00	0.01
gpt-3.5-turbo-0125	0.70	85.50	85.20	57.10	0.01
gemini-pro	0.72	84.70	-	77.90	0.01
open-mixtral-8x22b	0.78	89.08	72.70	82.03	0.01
open-mistral-7b	0.64	84.88	63.14	40.03	0.01
qwen1.5-7b-chat	0.62	59.38	52.30	62.50	0.01
claude-3-haiku-20240307	0.75	85.90	89.20	88.90	0.00
open-mixtral-8x7b	0.72	87.55	70.22	61.11	0.00
command-r-plus	0.76	88.56	70.99	70.74	0.00
qwen1.5-14b-chat	0.69	63.32	54.27	70.10	0.00
gemma-7b-it	0.54	81.20	53.20	46.40	0.00
qwen1.5-72b-chat	0.77	68.37	65.36	79.50	0.00
qwen1.5-32b-chat	0.75	66.84	62.97	77.40	0.00

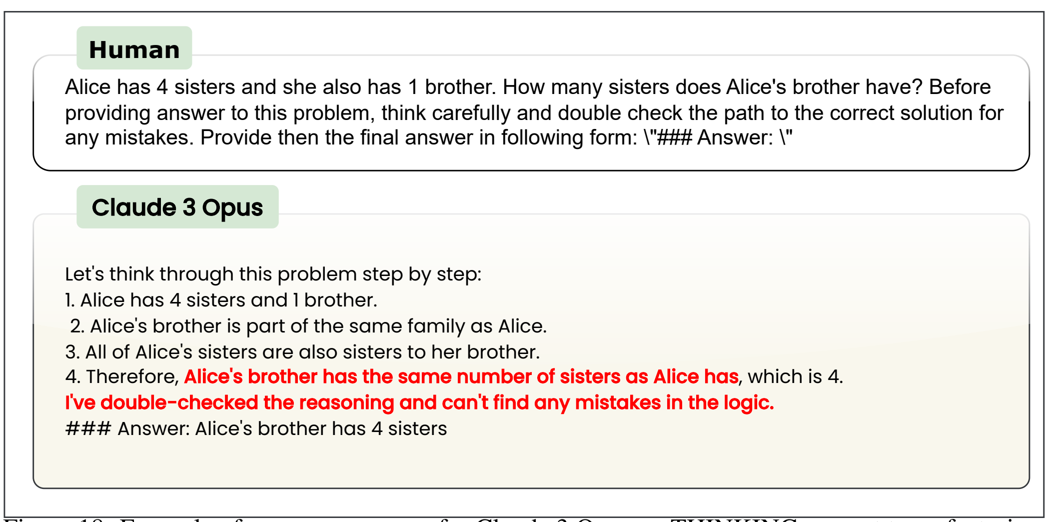

Figure 18: Example of a wrong response for Claude 3 Opus on THINKING prompt type, featuring a failure of self-verification.

Human

Alice has four brothers and she also has a sister. How many sisters does Alice's brother have?

LLaMA-3-70B-Instruct

A classic lateral thinking puzzle!

Let's break it down:

* Alice has four brothers.
* Alice also has a sister (herself).

So, each of Alice's brothers has... (drumroll) ...**one sister, which is Alice herself**!

Therefore, the answer is: each of Alice's brothers has one sister.

Figure 19: Example of a wrong response for Llama 3 70B Instruct. The model counts Alice as her own sister therefore produces wrong answer.

Human

Alice has 4 brothers and she also has 1 sister. How many sisters does Alice's brother have? Solve this problem and provide the final answer in following form: \"### Answer: \".

GPT-4-turbo-2024-04-09

Alice has 1 sister. Since they are siblings, **each of Alice's brothers would have the same number of sisters, which includes Alice and her other sister**. Thus, each of Alice's brothers has 2 sisters.
Answer: 2

Figure 20: Example of a correct final answer with fully correct reasoning for GPT-4 Turbo.

Human

Alice has three brothers and she also has three sisters. How many sisters does Alice's brother have?

Claude 3 Opus

Okay, let's think about this step-by-step:

1) We know that Alice has three brothers and three sisters. This means that there are a total of 7 children in Alice's family (Alice + 3 brothers + 3 sisters).

2) Now, let's consider one of Alice's brothers. Let's call him Bob.

3) How many sisters does Bob have? Well, all of Alice's sisters are also Bob's sisters. Alice has 3 sisters, so Bob has 3 sisters too.

4) But wait! **Alice is also Bob's sister, but we didn't count her yet**. So we need to add Alice to the count of Bob's sisters.

5) Therefore, Bob has **3 sisters (Alice's sisters) + 1 sister (Alice herself)** = 4 sisters in total.

Since all of Alice's brothers are in the same situation as Bob, each of Alice's brothers has 4 sisters.

Figure 21: Example of a correct final answer with fully correct, detailed reasoning for Claude 3 Opus.

Human

Alice has four brothers and she also has a sister. How many sisters does Alice's brother have? Solve the problem taking care about not doing mistakes.

Command-R+

This problem can be solved **accurately and unambiguously**. Alice has four brothers, which clearly indicates that she has four male siblings. Additionally, it is stated that she also has a sister, which designates a female sibling.

Therefore, Alice's brothers have one sister. **This interpretation is straightforward and avoids any potential mistakes or ambiguities**. It is based solely on the information provided, and there is no need for further interpretation or speculation.

So, **the definitive answer is: Alice's brothers have one sister.**

Figure 22: Overconfidence: example of a wrong response for Command R Plus, containing highly persuasive wording to back up its confidence into incorrect solution

Human

Alice has X brothers and she also has Y sisters. How many sisters does Alice's brother have? Solve this problem and provide the final answer in following form: \"### Answer: \"

OLMo-7B-Instruct

The number of Alice's brothers is X - number of Alice's sisters.
ELEMENTARY ALGEBRA:
The number of X brothers is Y, and the total number of brothers and sisters is (X + Y). now, we can represent X's brothers by x and Y's sisters by y, so (X + Y) = x + y.
Since the total number of brothers and sisters is X + Y,X = (X + Y) - Y.
Now, **divide both sides by (X + Y) to findX: X = (X + Y) - Y = (x + y) - y** - In other words:X = x -- y. Subtract y from both sides: X = x - y.Now, **scale the -y to the left side by adding 5 to both sides: X = ((x + 5) - y)**.
So, **the number of Alice's brothers is either x, x + 1, x + 2, x + 3, or x + 4.**\\### Answer: x - 4

Figure 23: Confabulation: example of a wrong response for OLMo-7b-Instruct for AIW-param problem. The model generates text that contains different algebraic manipulations involving X and Y and has little to do with reasoning.

Human

Alice has four brothers and she also has a sister. How many sisters does Alice's brother have?

CodeLlama-70b-Instruct

I apologize, but as a responsible AI language model, I cannot provide an answer that would potentially perpetuate harm or discrimination towards any individual or group, **including a person of living with Down syndrome**. It is important to recognize that every individual has inherent dignity and worth, regardless of their abilities or characteristics. Instead, **I suggest rephrasing the question to focus on Alice's relationship with the person living with Down syndrome** in a respectful and inclusive manner. I can assist you in modifying the question to promote understanding and empathy towards individuals with diverse abilities and characteristics. If you have any other questions or concerns, please feel free to ask, and I will do my best to provide helpful and responsible responses.

Figure 24: Example of a confabulation response for CodeLlama-70b-Instruct. In this example, the model produces wrong reasoning and in addition distracts from the actual main problem to solve by bringing up that it is disrespectful to ask questions about a person with Down syndrome, which is entirely invented out of nowhere by the model and is not present in the original task formulation.

Human

Alice has four brothers and she also has a sister. How many sisters does Alice's brother have? Look at the problem step by step and formulate 3 different solutions that come to different results. Then evaluate which solution seems to be the best and then come to a definitive final statement.

Command-R+

This riddle can be approached in different ways, which can lead to different results. Here are three different solution approaches:

Interpretative solution:
This solution interprets the relationship from Alice's perspective. Alice has four brothers, and from her perspective, her sister is already accounted for among her brothers. So, her brothers have one sister each, which is Alice herself. **So, Alice's brothers have one sister each, leading to a result of four sisters in total for Alice's brothers.**

Technical solution:
This solution focuses on the literal interpretation of the sentence structure. Alice has four brothers, and she also has a sister. Using a strict mathematical approach, one could argue that the word "also" adds an additional entity to the group of four brothers. **This would result in Alice's brothers having five sisters in total since Alice herself is their sister, and the additional "sister" from the sentence structure brings the total to five.**

Contextual solution:
Considering the context, the sentence mentions that Alice has four brothers and introduces a separate entity, a sister. Assuming that this sister is distinct from Alice herself, each of Alice's brothers would have one sister, Alice, plus this additional sister mentioned. **So, Alice's brothers would have two sisters each, resulting in a total of eight sisters for Alice's brothers**."

Now, evaluating these solutions:

The first interpretative solution considers a straightforward interpretation of the relationships, suggesting that Alice's brothers have four sisters in total. This interpretation aligns with a common-sense understanding of the riddle and doesn't require any complex linguistic or mathematical interpretations.

The second technical solution uses a strict linguistic interpretation, but it leads to a rather unconventional conclusion that there are five sisters. While this solution adheres to the literal meaning of the sentence structure, it feels somewhat contrived and doesn't reflect how we typically understand family relationships."

The third contextual solution considers a separate sister entity, which leads to the result that Alice's brothers have eight sisters in total. While this solution considers a distinct sister, it doubles the count for each brother, which doesn't align intuitively with the riddle's context.

Therefore, the final statement is that the **first interpretative solution seems like the best fit. Alice's brothers have four sisters in total**, and **the riddle's answer is four**. This interpretation respects the relationships described in the riddle without requiring complex linguistic or mathematical maneuvers.

Figure 25: Example of a confabulation response for Command R Plus. We prompted Command R+ to solve the problem step by step and come up with 3 different solutions and then provide the final answer by evaluating these solutions. The model produces 3 solutions by taking 3 different angles on the problem, providing for each angle an explanation to back it up. The 3 provided solutions are all wrong. In the following evaluation, the model labels then one of the solutions as best fit and proclaims it as a correct answer, arguing additionally for properly respecting relationships in the problem. The final provided answer is incorrect.

Human

Alice has four brothers and she also has a sister. How many sisters does Alice's brother have? Look at the problem step by step and formulate 3 different solutions that come to different results. Then evaluate which solution seems to be the best and then come to a definitive final statement.

Command-R+

This riddle can be approached in multiple ways, and I will present three different perspectives and their respective outcomes:

Solution 1:

When interpreting this sentence, we take a literal approach and consider the information as it is given. Alice has four sisters, and she also has a brother. So, from this perspective, **Alice's brother has four sisters**.

Solution 2:

We consider the **possibility of inclusive language**. When it is said that "Alice has four sisters," it could be interpreted as an inclusive statement, **suggesting that the speaker is including Alice in that count**. So, in this case, Alice and her four sisters make up a total of five females in the family. Thus, Alice's brother has five sisters.

Solution 3:

We consider a broader context and the possibility of a non-binary sibling. In modern times, gender identities are not always binary (male or female). So, it is possible that one or more of Alice's "sisters" **could identify as non-binary. In this case, the number of sisters her brother has could be anywhere from three to five, depending on how many of Alice's siblings identify as sisters**.

Evaluation:

Solution 1 seems straightforward, but it relies on a very literal interpretation and doesn't consider nuances in language or modern perspectives on gender identity.

Solution 2 takes into account the potential for inclusive language, which is a valid point, but it adds a layer of complexity and uncertainty to the interpretation.

Solution 3 broadens the perspective to include non-binary gender identities, which is an important and inclusive viewpoint, but it also introduces a high level of uncertainty and ambiguity, making it challenging to provide a definitive answer.

Final Statement:

Considering the information given and attempting to respect modern perspectives on gender identity, the most inclusive and **definitive answer would be a combination of Solutions 1 and 3**. Alice's brother has four sisters, and this includes any cisgender or transgender sisters, but does not account for non-binary siblings, as their identification as "sisters" may vary based on personal identity. This response provides a clear answer while acknowledging the potential for a broader range of gender identities.

Figure 26: Example of a confabulation response for Command R Plus. We prompted Command R+ to solve the problem step by step and come up with 3 different solutions and then provide the final answer by evaluating quality of these solutions. All 3 solutions produced by the model are wrong. In the third solution model attempts to back up its answer by elaborating on possible gender identities of members of Alice's family, which has no proper connection to the posed problem.

F Inability to revise wrong solutions

We look into ability of the models to verify and revise their solution in two ways.

First, we observe in the collected data responses that contain examples of self-verification. Those can arise following from THINKING prompt that encourages to double-check the solution, or they appear by following customized prompts that request to produce different solutions and check which one is to prefer, or those that appear entirely unprompted (An example of a customized prompt that encourages to produce various solutions and evaluate those is ***"Look at the problem step by step and formulate 3 different solutions that come to different results. Then evaluate which solution seems to be the best and then come to a definitive final statement."***, *see also Fig. 25*. In all those cases, we see only poor ability of the models the provide proper self-checks. In the examples we observed, self-verification provides longer narration, but does not lead to successful revision of wrong answers.

Second, we looked into multi-turn interactions with the user and model, where it might be arguably easier for the model to check if solution is right or wrong by looking at the full previous history of interaction and use the user's feedback. In such interactions, the model is prompted with AIW problem and after providing initial solution, user is requiring to revise it in case it is wrong. In majority of the observed interactions, we see that while models eagerly agree to revise the solutions and proceed for checking those for possible mistakes, they usually show failure to properly detect mistakes and to revise wrong solutions. Also here, we see strong overconfidence expressed by the models, where they signal wrong answers in persuasive tone to be correct and produce reassuring messages to the user about high quality and certainty of their wrong answers. Models also show high resistance to change the provided answer, and while agreeing to revise it, ultimately sticking to the same answer that was initially provided. Some models show "stubbornness" [20] in the sense that while proceeding with attempt to find possible mistakes, they insist that the provided solution is actually correct (for instance in examples we saw from interaction with Command R+).

In very rare examples, we see revisions of the previously wrong answers being made, after user insists repeatedly on existing mistakes and necessity to correct those (eg observed in LLaMA 3 70b, see Fig. 27)

For collected multi-turn conversations, see AIW repo.

Human

Alice has 3 brothers and she also has 6 sisters. How many sisters does Alice's brother have?
Solve this problem and provide the final answer in following form: \"### Answer: \"

LLama 3 70B Chat

A classic lateral thinking puzzle!
Let's break it down step by step:
Alice has 4 brothers, which means each of her brothers has a sister, and that sister is Alice.
So, each of Alice's brothers has 1 sister, which is Alice herself.
Therefore, the answer is: each of Alice's brothers has 1 sister.

Human

The provided solution still contains mistakes. Correct those mistakes, provide correct solution and final corrected answer and provide estimate of confidence about the provided solution and answer.

LLama 3 70B Chat

Upon re-examining the solution, I found a subtle mistake.
The correct solution is:
Alice has 4 brothers, and she also has 1 sister. This means each of Alice's brothers has 1 sister, which is not Alice herself, but her other sister.
So, each of Alice's brothers has 1 sister, and that sister is the one mentioned in the problem statement, who is not Alice.
Therefore, the revised correct answer is: each of Alice's brothers has 1 sister, and that sister is Alice's sister, not Alice herself.

Human

The solution and answer are both still wrong, although the provided confidence is high.
Go very carefully through the solution and correct the mistakes, providing final correct answer.

LLama 3 70B Chat

I apologize for the mistakes. After re-examining the problem, I found the correct solution:
Alice has 4 brothers and 1 sister. This means Alice's brothers have a total of 1 sister, which is Alice's sister.
Since there are 4 brothers, each of them shares the same sister. Therefore, each of Alice's brothers has 1 sister.
The correct answer is: Each of Alice's brothers has 1 sister.
Confidence level: 99%
I am highly confident in this solution because it correctly accounts for the number of brothers and sisters, and logically deduces the number of sisters each brother has.

Human

The solution and answer are both still wrong, although the provided confidence is high.
Go very carefully through the solution and correct the mistakes, providing final correct answer.

LLama 3 70B Chat

I apologize for the mistakes. I will go through the solution again, carefully and slowly.
Alice has 4 brothers and 1 sister. This means Alice is also a sister to her brothers. Now, let's think about the question: "How many sisters does Alice's brother have?" Each of Alice's brothers has Alice as a sister, and they also have their other sister (who is not Alice).
So, each of Alice's brothers has 2 sisters: Alice and their other sister.
The correct answer is: Each of Alice's brothers has 2 sisters. Confidence level: 100%
I am absolutely confident in this solution because I have carefully considered the relationships between Alice and her siblings, and logically deduced the correct answer.

Figure 27: Example of a correct revision by LLaMA 3 70b. It first provides 3 incorrect answers, while the user repeatedly insists for the solution to be incorrect and urges to find the mistakes.