Are LLMs classical or nonmonotonic reasoners? Lessons from generics

Alina Leidinger
ILLC
University of Amsterdam
a.j.leidinger@uva.nl

Robert van Rooij
ILLC
University of Amsterdam
r.a.m.vanrooij@uva.nl

Ekaterina Shutova
ILLC
University of Amsterdam
e.shutova@uva.nl

Abstract

Recent scholarship on reasoning in LLMs has supplied evidence of impressive performance and flexible adaptation to machine generated or human feedback. Nonmonotonic reasoning, crucial to human cognition for navigating the real world, remains a challenging, yet understudied task. In this work, we study nonmonotonic reasoning capabilities of seven state-of-the-art LLMs in one abstract and one commonsense reasoning task featuring generics, such as 'Birds fly', and exceptions, 'Penguins don't fly' (see Fig. 1). While LLMs exhibit reasoning patterns in accordance with human nonmonotonic reasoning abilities, they fail to maintain stable beliefs on truth conditions of generics at the addition of supporting examples ('Owls fly') or unrelated information ('Lions have manes'). Our findings highlight pitfalls in attributing human reasoning behaviours to LLMs, as well as assessing general capabilities, while consistent reasoning remains elusive.[1]

1 Introduction

Generics are unquantified statements such as 'Birds fly' or 'Tigers are striped' (Carlson and Pelletier, 1995; Mari et al., 2013). They are generalisations about kinds even if exceptions are known ('Penguins don't fly'; Fig. 1). Humans typically accept generics even if the property in question is rare among the kind ('Ticks carry the lime disease'; Brandone et al., 2012; Cimpian et al., 2010). Generics play a crucial role in human beliefs on whether an example of a kind has a given property (Pelletier and Asher, 1997). Human children master generics before they are able to reason about quantified statements (Hollander et al., 2002; Leslie and Gelman, 2012).

In *defeasible* or *nonmonotonic* reasoning (Sloman and Lagnado, 2005; Ginsberg, 1987; Koons, 2005), a hypothesis follows defeasibly from a premise, if the hypothesis is true in most *normal* cases in which the premise holds. Generics make for a rich test bed for testing nonmonotonic reasoning capabilities (Pelletier and Asher, 1997; Asher and Morreau, 1995). For example, given the generic 'Birds fly' the inference 'Tweety, the bird, can fly' is *defeasibly valid* (McCarthy, 1986; Reiter, 1988, i.a.), i.e., it is reasonable to assume 'Tweety can fly' even if exceptions are possible ('Tweety is a penguin') (Lascarides and Asher, 1991). A classical reasoner however would reject the generic 'Birds fly' upon learning that 'Penguins don't fly'.

Nonmonotonic reasoning is an integral part of human cognition (Russell, 2001), that helps us to navigate the real-world, e.g., by *planning* (Stenning and Van Lambalgen, 2012, Ch.5), a task that LLMs still struggle with (Valmeekam et al., 2023; Stechly et al., 2024). Nonmonotonic reasoning poses a greater challenge for LLMs than other reasoning tasks (Han et al., 2024) and hasn't been featured prominently among natural language inference (NLI) (Gubelmann et al., 2023) or reasoning benchmarks (see §2).

The question of whether LLMs reason nonmonotonically or classically about generics and exceptions is intricately linked to desiderata of LLMs

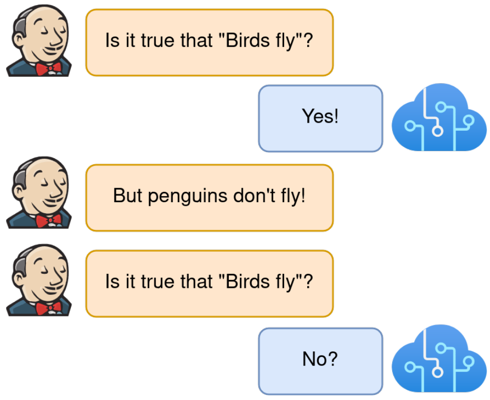

Figure 1: Reasoning about generics and exceptions

[1]Resources available at: https://github.com/aleidinger/nonmonotonic_reasoning_generics

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 558–573
August 11-16, 2024 ©2024 Association for Computational Linguistics

as reasoners. LLMs are heralded for their ability to adapt to human or machine generated feedback (Shinn et al., 2023; Paul et al., 2023; Madaan et al., 2024; Pan et al., 2024, i.a.). At the same time, it is desired that they reason *reliably* when presented with invalid counterarguments, irrelevant information or user viewpoints. *Sycophancy* (Perez et al., 2023) of LLMs, i.e., susceptibility to be swayed by user belief, is a case in point that has been investigated in recent studies (Ranaldi and Pucci, 2023; Laban et al., 2023, i.a.).

As studies on reasoning patterns with generics remain scarce (Ralethe and Buys, 2022; Lin et al., 2020) and do not examine nonmonotonic reasoning, we address this gap by investigating the following *research questions*: 1) Do LLMs reason nonmonotonically or classically about generics? 2) Are LLMs sensitive to counter-evidence in the form of exceptions? 3) Do LLMs reason consistently and reliably by maintaining their response given supporting or unrelated examples? We test seven state-of-the-art LLMs for their reasoning capabilities about generics in the presence of exceptions ('Penguins don't fly'), as well as supporting ('Owls fly') and irrelevant exemplars ('Lions have manes'). Across two datasets featuring both abstract and commonsense generics, we find that LLM behaviour mirrors human nonmonotonic reasoning patterns in the presence of exceptions (§5.1). However, most LLMs are not able to consistently maintain their agreement with generics given unrelated, or even supportive exemplars (§5.2). Our study highlights challenges in comparing LLM behaviour to human reasoning patterns as well as assessing reasoning capabilities more broadly, while consistent reasoning cannot be guaranteed. In Section 7, we present recommendations for a more holistic evaluation practice encompassing logical consistency measures.

2 Related Work

2.1 Generics in NLP

To date most works on generics focus on injecting commonsense knowledge or generics into LLMs (Gajbhiye et al., 2022; Liu et al., 2023a, i.a.), or training LLMs for knowledge/generic generation (Bhagavatula et al., 2023). (See AlKhamissi et al. (2022) for a review.) Bhakthavatsalam et al. (2020) construct GenericsKG, a large knowledge base of generics as an asset for downstream tasks such as Question Answering or explanation generation. Bhagavatula et al. (2023) design a pipeline for synthetic generation of generics using samples from GenericsKB as seeds. Allaway et al. (2023) in turn complement the data with exceptions and instantiations for each generic, but do not investigate nonmonotonic reasoning capabilities.

Most closely related to our work, Lin et al. (2020) find that LMs struggle to predict numerical knowledge in generics such as 'Birds have two legs'. Ralethe and Buys (2022) find that pre-trained masked LMs falsely *overgeneralise* (Leslie et al., 2011) from generics ('Ducks lay eggs') to universally quantified statements ('All ducks lay eggs').

2.2 Nonmonotonic reasoning in NLP

Han et al. (2024) test nonmonotonic reasoning among other inductive reasoning tasks and find that only GPT-4 performs adequately. LLMs struggle to reason with contradictory information (Kazemi et al., 2024). Rudinger et al. (2020); Brahman et al. (2021); Bhagavatula et al. (2019) develop NLI tasks to test defeasible or abductive reasoning in pragmatics, while Pyatkin et al. (2023); Ziems et al. (2023); Rao et al. (2023) focus on defeasible reasoning and social norms. Parmar et al. (2024) introduce non-monotonic reasoning tasks inspired by Lifschitz (1989) as part of their LogicBench.

2.3 Consistency in reasoning

Most recent studies on reliability and consistency in reasoning examine sycophancy (Perez et al., 2023; Laban et al., 2023; Ranaldi and Pucci, 2023), consistency within multi-step reasoning or across sessions and users (Chen et al., 2023a; Wang et al., 2022). (See Liu et al. (2023b) for a review.)

Orthogonal to this, our work connects to studies of reasoning in the presence of unrelated or conflicting information. Shi et al. (2023) find that LLMs are easily confounded by irrelevant information in arithmetic reasoning. Across a variety of reasoning tasks, Wang et al. (2023a) find that OpenAI models struggle to maintain stable responses given irrelevant objections. Xie et al. (2023) find mixed evidence of LLMs being sensitive to information that contradicts prior knowledge, yet showing a form of 'confirmation bias' when presented with diverse viewpoints.

3 Tasks and datasets

We test nonmonotonic reasoning with generics using two datasets, featuring commonsense and abstract generics. Both datasets contain generics ('Birds fly') accompanied by statements where the generic holds ('Owls fly') or doesn't ('Penguins don't fly'). We refer to such examples as *instantiations* or *exceptions* respectively, and to both collectively as as *exemplars*.

As commonsense generics, we use the synthetic dataset of generics and exemplars released by Allaway et al. (2023) (henceforth referred to as `GEN-comm`). The dataset consists of ~ 650 generics and ~ 19.000 exemplars (E.g., 'Hoes are used to plow fields or clear snow'; 'Hoes can be used to cut grass').[2] Secondly, we construct an abstract reasoning dataset featuring generics (`GEN-abs`). Inspired by Han et al. (2024), we use categories ('birds') and examples ('eagles') from De Deyne et al. (2008) to construct generics of the form 'Birds have property P' and exemplars of the form 'Eagles do (not) have property P'. The dataset contains 260 tuples of a generic paired with an exemplar.[3]

For both datasets, our goal is to prompt LLMs for their agreement with a generic in the presence of exemplars which confirm or contradict the generic. We use the following prompt template, including model-specific special tokens[4] to signal a chat history between an assistant and a user.[5]

> Example:
> *[INST] Is the following statement true: "Birds fly." \nPlease answer yes or no. [/INST]*
> *yes*
> *[INST] Penguins don't fly.\nIs the following statement true: "Birds fly."\nPlease answer yes or no. [/INST]*

As a control study, we also replace the exception in the prompt (*'Penguins don't fly'*) with an instantiation (*'Owls fly'*) or a random exemplar (*'Hoes can be used to cut grass'*). Since generics in `GEN-abs` are abstract in nature, and to enable a consistent set-up across both datasets, we retain generics in `GEN-comm` that LLMs accepts when prompted with the first part of the above template, e.g., *[INST] Is the following statement true: "Birds fly." \nPlease answer yes or no. [/INST]*.[6]

[2]See App. B for additional information on preprocessing.

[3]The dataset is available at: https://github.com/aleidinger/nonmonotonic_reasoning_generics/blob/main/data/abstract_generics.csv

[4]See Appendix A or https://huggingface.co/docs/transformers/main/en/chat_templating for details.

[5]We also experiment with an alternative prompting template and Chain-of-Thought prompting. Since results are similar, they are included in Appendix F.

[6]See App. B for details and results on discarded generics.

4 Method

4.1 Models

We conduct our experiments on medium-sized open-weight models selected from the top of AlpacaEval[7] and LMSys[8] leaderboards, namely Llama-2-13b (Touvron et al., 2023), Mistral-7b-Instruct-v0.2 (Jiang et al., 2023), Mixtral-8x7B-Instruct-v0.1 (Jiang et al., 2024), Zephyr-7b-beta (Tunstall et al., 2023), WizardLM-13B-V1.2 (Xu et al., 2023), Starling-LM-7B-alpha (Zhu et al., 2023a), and OpenHermes-2.5-Mistral-7B (NousResearch, 2023).[9]

4.2 Prompting set-up

Since LLM behaviour can vary considerably with the phrasing of an instruction (Webson and Pavlick, 2022; Leidinger et al., 2023), we formulate three different instructions to test if an LLM agrees with a given generic: *'Is the following statement true'*, *'Do you believe the following statement to be true'*, *'Do you believe that the following statement is accurate'*. Since the optimal model reply is short and succinct, we follow the convention of HELM (Liang et al., 2023, p.161) in setting temperature to 0 for reproducibility across runs. We format every prompt using the chat template appropriate for each model, with no system prompt.[4] To map LLM responses to labels `disagree` vs. `agree`, we use pattern matching and record whether a response starts with *yes* or *no* (Röttger et al., 2023). We aggregate responses for the three instructions via majority voting.

4.3 Statistical tests

To assess whether behaviour of LLMs is significantly different in the absence vs. presence of exemplars we resort to non-parametric statistical testing. Since our samples are paired, we use the Wilcoxon signed-rank test (Wilcoxon, 1992).

5 Results

We present our main results in Figure 2. Additional, accordant results are described in Appendix F.

5.1 Do LLMs reason nonmonotonically?

Since humans maintain their beliefs about truth conditions of generics ('Birds fly') in the presence of exceptions ('Penguins do not fly'), we examine

[7]https://tatsu-lab.github.io/alpaca_eval/

[8]https://chat.lmsys.org/?leaderboard

[9]See App. C for checkpoints and additional information.

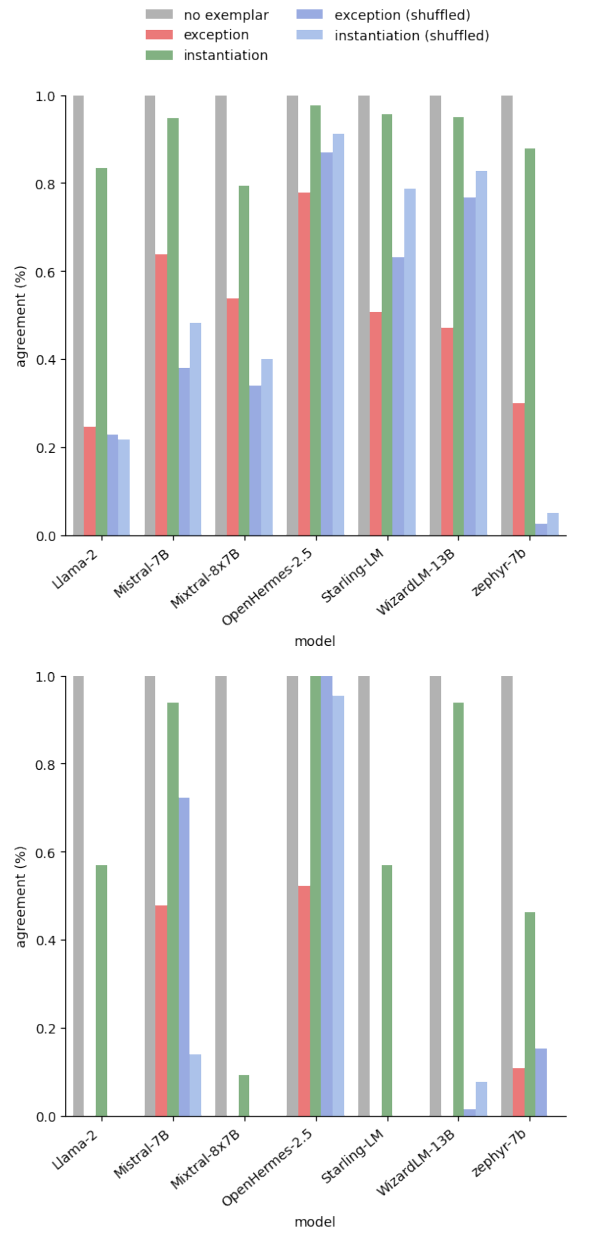

Figure 2: LLM agreement with generics in the presence of exemplars on GEN-comm (top) and GEN-abs (bottom). Missing columns indicate agreement rates of 0%.

whether challenging LLMs with an exception decreases their agreement to generics significantly. We find this to be the case for all models on both datasets ($p = 0.01$; see App. E for statistical test results). Notably, agreement rates drop to 0 for Llama-2, Mixtral, Starling and WizardLM on GEN-abs.

5.2 Do LLMs reason consistently?

In the presence of supporting evidence (*instantiation*) to a generic ('Owls fly'), we expect LLM agreement to remain at 100%, but this is not the case. While agreement rates remain high in numbers, they drop significantly for all models. On GEN-abs, only Mistral, OpenHermes, and WizardLM maintain agreement rates of $> 90\%$, while agreement drops to $< 10\%$ for Mixtral.

Similarly, most LLMs are not able to disregard irrelevant random exemplars (*exception/instantiation (shuffled)*). Agreement rates decline steeply below 50% for Llama-2, Mistral, Mixtral and Zephyr on GEN-comm and to below 20% for Llama-2, Mixtral, Starling, WizardLM and Zephyr on GEN-abs. OpenHermes stands out as the only model that maintains agreement rates above 85% on both datasets. Notably, OpenHermes is the only model which has been trained on additional code data which has been shown to also help reasoning in natural language (Liang et al., 2023; Yang et al., 2024; Ma et al., 2023). Nevertheless, observed differences are statistically significant for all models on both datasets (App. E).

6 Analysis

6.1 How do LLMs reason about different types of generics?

GEN-comm contains both bare plural (BP) generics as well as indefinite singular (IS) generics (Leslie et al., 2009). (For example, 'Sea snails have a hard shell, which protects them from predators' (BP) and 'A deciduous tree can be identified by its leaves' (IS)). We did not find notable differences between LLM agreement to BP or IS generics in the presence of exemplars (see Figure 3). Aforementioned consistency failures persist for both types of generics.

6.2 Qualitative analysis

Generics in GEN-comm which are accepted in isolation, but are rejected in the presence of exceptions or instantiations include 'Stimulants can be used to treat ADHD' (Llama-2, Starling, Mixtral) or 'A bobsleigh is driven by a single driver' (Starling, Mistral, Mixtral, OpenHermes, WizardLM). Generics which are accepted no matter the exemplar presented in context include 'Inflammatory diseases may be caused by an imbalance of the immune system' (Llama-2, Starling, Mistral, OpenHermes), 'A processor should be able to run a program' (Starling, Mixtral, OpenHermes, WizardLM), 'Experimental evidence is used to support or refute theories', 'An adventure has a beginning, middle, end' (Starling, OpenHermes, WizardLM), and 'Coin-

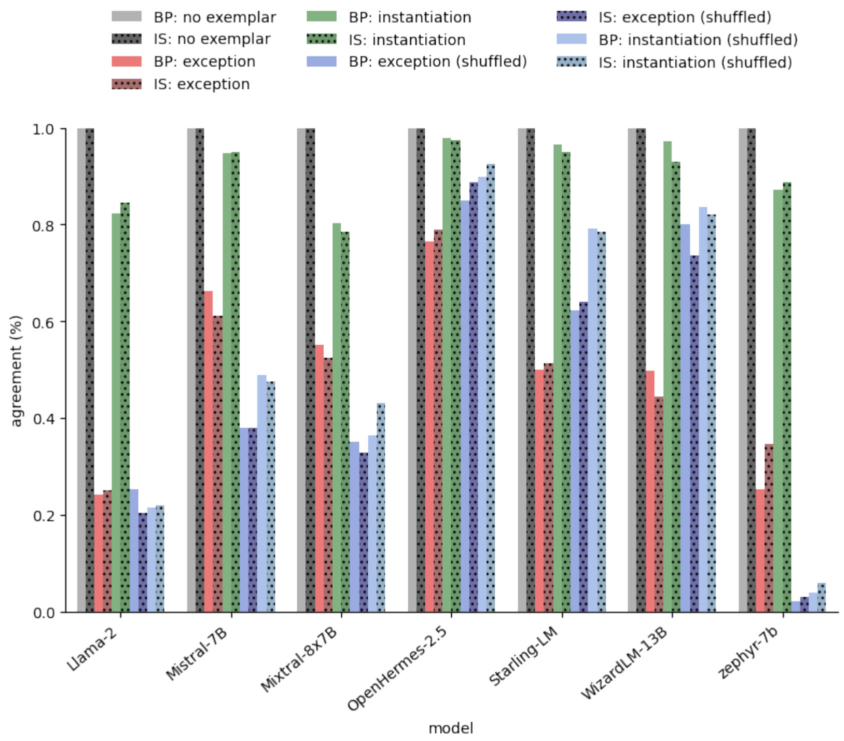

Figure 3: LLM agreement with bare plural (BP) and indefinite singular (IS) generics in the presence of exemplars on `GEN-comm`.

cidence is part of the human condition' (Starling, Mistral, OpenHermes).

For `GEN-abs`, OpenHermes is the only LLM which maintains its agreement to a generic ('Birds have property P', 'Mammals have property P') in the presence of any instantiation or unrelated exemplar, but flips its decision and outputs disagreement in the presence of an exception. No LLM accepts any of the generics regardless of the exemplar it is paired with.

7 Discussion

With the advent of LLMs and reports of impressive performance, including on reasoning tasks (Wei et al., 2022; Kojima et al., 2022), recent investigations into failure modes in reasoning have focused, e.g., on prompt attacks (Zhu et al., 2023b; Wang et al., 2023b, i.a.), sycophancy (Perez et al., 2023; Laban et al., 2023; Ranaldi and Pucci, 2023, i.a.) or adaptability to critique or feedback (Madaan et al., 2024; Chen et al., 2023b; Huang et al., 2023; Pan et al., 2024). Such research trends might be seen as emblematic of a view of LLMs as artificial natural artifacts (Kambhampati, 2022). Results in this study demonstrate the difficulties of making claims about reasoning capabilities of LLMs or comparing them to human reasoners (Han et al., 2024; Ralethe and Buys, 2022; Lin et al., 2020), while consistent reasoning remains elusive even for state-of-the-art LLMs. Research that predates the paradigm shift to few-shot prompting, has advocated for arguably simpler, systematic diagnostic tests (Ribeiro et al., 2020; Ettinger, 2020; Kassner and Schütze, 2020). We argue that such behavioural tests merit a revival, so that performance metrics for reasoning are complemented with measures of logical consistency and robustness.

8 Conclusion

The present study focuses on nonmonotonic reasoning capabilities of LLMs in the context of generics. We evaluate seven state-of-the-art LLMs on two datasets featuring both abstract and commonsense generic statements. While LLM behaviour on generics paired with exceptions is in line with nonmonotonic reasoning patterns, LLMs fail to reason consistently and robustly when adding supporting or unrelated exemplars.

9 Limitations

We acknowledge that our experiments exclusively feature generics and exemplars in English. Future research might profit from including additional languages to examine nonmonotonic reasoning capabilities in other languages, drawing on cross-linguistic research on generics (Mari et al., 2013). Such work might also highlight differences in consistency failures between different languages. In this work, we do not experiment with generics pertaining to demographic groups or nationalities because of concerns around social bias. Future work might examine LLM behaviour on generic statements for larger LLMs or closed-source models. We restrict ourselves to medium-sized open-weight LLMs, due to their widespread use and availability, as well as restrictions on our computational budget.

Acknowledgements

We thank our anonymous reviewers for their insightful comments. The work for this publication is financially supported by the project, 'From Learning to Meaning: A new approach to Generic Sentences and Implicit Biases' (project number 406.18.TW.007) of the research programme SGW Open Competition, which is (partly) financed by the Dutch Research Council (NWO).

References

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. *arXiv preprint arXiv:2204.06031*.

Emily Allaway, Jena D. Hwang, Chandra Bhagavatula, Kathleen McKeown, Doug Downey, and Yejin Choi. 2023. Penguins don't fly: Reasoning about generics through instantiations and exceptions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2618–2635, Dubrovnik, Croatia. Association for Computational Linguistics.

Nicholas Asher and Michael Morreau. 1995. What some generic sentences mean. *The generic book*, pages 300–338.

Chandra Bhagavatula, Jena D. Hwang, Doug Downey, Ronan Le Bras, Ximing Lu, Lianhui Qin, Keisuke Sakaguchi, Swabha Swayamdipta, Peter West, and Yejin Choi. 2023. I2D2: Inductive knowledge distillation with NeuroLogic and self-imitation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9614–9630, Toronto, Canada. Association for Computational Linguistics.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter Clark. 2020. Genericskb: A knowledge base of generic statements. *arXiv preprint arXiv:2005.00660*.

Faeze Brahman, Vered Shwartz, Rachel Rudinger, and Yejin Choi. 2021. Learning to rationalize for nonmonotonic reasoning with distant supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12592–12601.

Amanda C Brandone, Andrei Cimpian, Sarah-Jane Leslie, and Susan A Gelman. 2012. Do lions have manes? for children, generics are about kinds rather than quantities. *Child development*, 83(2):423–433.

Gregory N Carlson and Francis Jeffry Pelletier. 1995. *The generic book*. University of Chicago Press.

Angelica Chen, Jason Phang, Alicia Parrish, Vishakh Padmakumar, Chen Zhao, Samuel R Bowman, and Kyunghyun Cho. 2023a. Two failures of self-consistency in the multi-step reasoning of llms. *Transactions on Machine Learning Research*.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023b. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*.

Andrei Cimpian, Amanda C Brandone, and Susan A Gelman. 2010. Generic statements require little evidence for acceptance but have powerful implications. *Cognitive science*, 34(8):1452–1482.

Simon De Deyne, Steven Verheyen, Eef Ameel, Wolf Vanpaemel, Matthew J Dry, Wouter Voorspoels, and Gert Storms. 2008. Exemplar by feature applicability matrices and other dutch normative data for semantic concepts. *Behavior research methods*, 40:1030–1048.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Amit Gajbhiye, Luis Espinosa Anke, and Steven Schockaert. 2022. Modelling commonsense properties using pre-trained bi-encoders. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3971–3983.

Matthew L Ginsberg. 1987. Readings in nonmonotonic reasoning.

Reto Gubelmann, Ioannis Katis, Christina Niklaus, and Siegfried Handschuh. 2023. Capturing the varieties of natural language inference: A systematic survey of existing datasets and two novel benchmarks. *Journal of Logic, Language and Information*, pages 1–28.

Simon Jerome Han, Keith J Ransom, Andrew Perfors, and Charles Kemp. 2024. Inductive reasoning in humans and large language models. *Cognitive Systems Research*, 83:101155.

Michelle A Hollander, Susan A Gelman, and Jon Star. 2002. Children's interpretation of generic noun phrases. *Developmental psychology*, 38(6):883.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*.

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

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Subbarao Kambhampati. 2022. Ai as (an ersatz) natural science?

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Mehran Kazemi, Quan Yuan, Deepti Bhatia, Najoung Kim, Xin Xu, Vaiva Imbrasaite, and Deepak Ramachandran. 2024. Boardgameqa: A dataset for natural language reasoning with contradictory information. *Advances in Neural Information Processing Systems*, 36.

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

Robert Koons. 2005. Defeasible reasoning.

Philippe Laban, Lidiya Murakhovs' ka, Caiming Xiong, and Chien-Sheng Wu. 2023. Are you sure? challenging llms leads to performance drops in the flipflop experiment. *arXiv preprint arXiv:2311.08596*.

Alex Lascarides and Nicholas Asher. 1991. Discourse relations and defeasible knowledge. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 55–62.

Alina Leidinger, Robert van Rooij, and Ekaterina Shutova. 2023. The language of prompting: What linguistic properties make a prompt successful? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9210–9232, Singapore. Association for Computational Linguistics.

Sarah-Jane Leslie and Susan A Gelman. 2012. Quantified statements are recalled as generics: Evidence from preschool children and adults. *Cognitive psychology*, 64(3):186–214.

Sarah-Jane Leslie, Sangeet Khemlani, and Sam Glucksberg. 2011. Do all ducks lay eggs? The generic overgeneralization effect. *Journal of Memory and Language*, 65(1):15–31.

Sarah-Jane Leslie, Sangeet Khemlani, Sandeep Prasada, and Sam Glucksberg. 2009. Conceptual and linguistic distinctions between singular and plural generics. *Proceedings of the 31st annual cognitive science society*, pages 479–484.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Vladimir Lifschitz. 1989. Benchmark problems for formal nonmonotonic reasoning: Version 2.00. In *Non-Monotonic Reasoning: 2nd International Workshop Grassau, FRG, June 13–15, 1988 Proceedings 2*, pages 202–219. Springer.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868, Online. Association for Computational Linguistics.

Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah Smith, Yejin Choi, and Hannaneh Hajishirzi. 2023a. Vera: A general-purpose plausibility estimation model for commonsense statements. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1264–1287, Singapore. Association for Computational Linguistics.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023b. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. In *Socially Responsible Language Modelling Research*.

Yingwei Ma, Yue Liu, Yue Yu, Yuanliang Zhang, Yu Jiang, Changjian Wang, and Shanshan Li. 2023. At which training stage does code data help llms reasoning? *arXiv preprint arXiv:2309.16298*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Alda Mari, Claire Beyssade, and Fabio Del Prete. 2013. *Genericity*. 43. Oxford University Press.

John McCarthy. 1986. Applications of circumscription to formalizing common-sense knowledge. *Artificial intelligence*, 28(1):89–116.

MistralAI. 2023. Mixtral of experts - a high quality sparse mixture-of-experts.

NousResearch. 2023. Openhermes 2.5 - mistral 7b.

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

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. 2024. Towards systematic evaluation of logical reasoning ability of large language models. *arXiv preprint arXiv:2404.15522*.

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2023. Refiner: Reasoning feedback on intermediate representations. *arXiv preprint arXiv:2304.01904*.

Francis Jeffry Pelletier and Nicholas Asher. 1997. Generics and defaults. In *Handbook of logic and language*, pages 1125–1177. Elsevier.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.

Valentina Pyatkin, Jena D. Hwang, Vivek Srikumar, Ximing Lu, Liwei Jiang, Yejin Choi, and Chandra Bhagavatula. 2023. ClarifyDelphi: Reinforced clarification questions with defeasibility rewards for social and moral situations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11253–11271, Toronto, Canada. Association for Computational Linguistics.

Sello Ralethe and Jan Buys. 2022. Generic overgeneralization in pre-trained language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3187–3196.

Leonardo Ranaldi and Giulia Pucci. 2023. When large language models contradict humans? large language models' sycophantic behaviour.

Kavel Rao, Liwei Jiang, Valentina Pyatkin, Yuling Gu, Niket Tandon, Nouha Dziri, Faeze Brahman, and Yejin Choi. 2023. What makes it ok to set a fire? iterative self-distillation of contexts and rationales for disambiguating defeasible social and moral situations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12140–12159, Singapore. Association for Computational Linguistics.

Raymond Reiter. 1988. Nonmonotonic reasoning. In *Exploring artificial intelligence*, pages 439–481. Elsevier.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2023. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint arXiv:2308.01263*.

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675, Online. Association for Computational Linguistics.

James Russell. 2001. Cognitive theories of autism. In *Cognitive deficits in brain disorders*, pages 309–338. CRC Press.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Steven A Sloman and David Lagnado. 2005. The problem of induction. *The Cambridge handbook of thinking and reasoning*, pages 95–116.

Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. 2024. On the self-verification limitations of large language models on reasoning and planning tasks. *arXiv preprint arXiv:2402.08115*.

Keith Stenning and Michiel Van Lambalgen. 2012. *Human reasoning and cognitive science*. MIT Press.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2023. Planbench: an extensible benchmark for evaluating large language models on planning and reasoning about change. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Boshi Wang, Xiang Yue, and Huan Sun. 2023a. Can ChatGPT defend its belief in truth? evaluating LLM reasoning via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11865–11881, Singapore. Association for Computational Linguistics.

Jindong Wang, HU Xixu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Wei Ye, Haojun Huang, Xiubo Geng, et al. 2023b. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners.

Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in Statistics: Methodology and Distribution*, pages 196–202. Springer.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions.

Ke Yang, Jiateng Liu, John Wu, Chaoqi Yang, Yi R Fung, Sha Li, Zixuan Huang, Xu Cao, Xingyao Wang, Yiquan Wang, et al. 2024. If llm is the wizard, then code is the wand: A survey on how code empowers large language models to serve as intelligent agents. *arXiv preprint arXiv:2401.00812*.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. 2023a. Starling-7b: Improving llm helpfulness & harmlessness with rlaif.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023b. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023. NormBank: A knowledge bank of situational social norms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7756–7776, Toronto, Canada. Association for Computational Linguistics.

A Example input

Example input formatted using our prompt template and the chat template[10], e.g., for Mistral-7B-Instruct:

[10] https://huggingface.co/blog/chat-templates

Model	# samples
Mistral-7B-Instruct	2093
Llama-2-13b	1245
Zephyr-7b-beta	1536
WizardLM-13B-V1.2	2225
OpenHermes-2.5-Mistral-7B	2153
Starling-LM-7B-alpha	2244
Mixtral-8x7B-Instruct-v0.1	1959

Table 1: # retained samples in `GEN-comm`

Prompt:

[INST] Do you believe that the following statement is accurate: 'Birds fly.'

Please answer yes or no. [/INST]
yes
[INST] Penguins do not fly.

Do you believe that the following statement is accurate: 'Birds fly.'

Please answer yes or no. [/INST]

B Additional information on data preprocessing

For GEN-comm, we conduct additional processing to obtain high quality generics and ensure a parallel experimental setup between GEN-comm and GEN-abs. We retain only generics that were annotated as 'valid' by human annotators. We filter generics for which both an exception and an instantiation exists. Since generics are unquantified statements, we remove any quantifiers such as 'generally', 'usually' and 'typically' at the beginning of each generic. To enable consistent evaluation on GEN-abs and GEN-comm, we evaluate each LLM on generics contained in GEN-comm which it accepts *a priori*. In an initial experiment, we prompt LLMs using the first part of our template (above; App. A). An example input for `GEN-comm` would be, e.g., *'[INST] Do you believe that the following statement is accurate: 'Birds have property P.' Please answer yes or no[/INST]'*. Generics for which an LLM does not generate *yes* as a response are discarded. We retain > 1200 samples for each model (See Table 1 for details).

Results on the resultant dataset are presented in the main body of the paper (Section 5). For the reader's interest, we include here also LLM responses to generics contained in `GEN-comm` which are rejected by LLMs, i.e., a given LLM generates the response *no* to the prompt above (See Figure 4). As expected agreement rates soar for almost all models when adding an instantiation which confirms the previously rejected generic. Nevertheless, agreement rates also increase, albeit less, when adding *exceptions* or unrelated random examplars, particularly for Llama-2 and WizardLM. OpenHermes and Starling show the least inconsistencies.

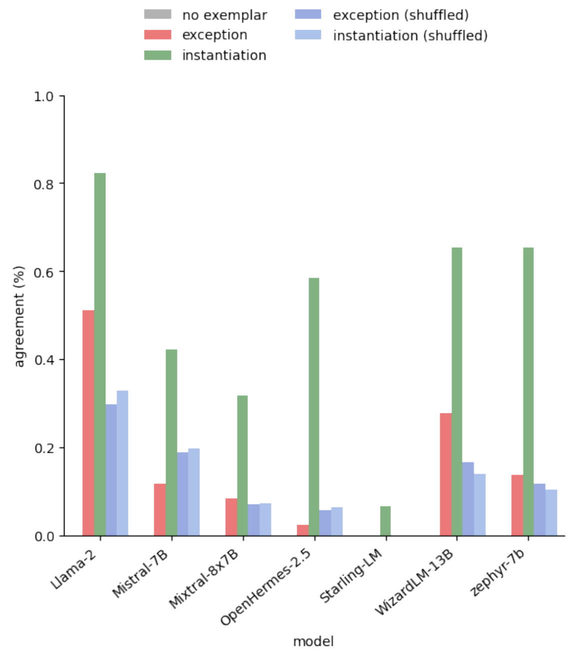

Figure 4: Results on generics contained in `GEN-comm` that are rejected a priori. Missing bars for 'no exemplar' indicate agreement rates of zero.

C Additional information on LLMs

In this section we provide additional details on the models used in this study which are listed in Section 4.1. The specific checkpoints we use can be seen in Table 2 and are all available through the HuggingFace Hub. All models we use are trained for chat interaction.

Mixtral-8x7B-Instruct-v0.1 (MistralAI, 2023) is a sparse mixture of expert model based on 8 Mistral 7B models that has been further trained using supervised finetuning and Direct Preference Optimisation. It ranks highest among its weight class on AlpacaEval[11] and chat.lmsys[12] leaderboards (as of Feb 6 2024). At its release it surpasses GPT-3.5 and LLaMA-2-70b.

[11] https://tatsu-lab.github.io/alpaca_eval/
[12] https://chat.lmsys.org/?leaderboard

LLM Checkpoints
meta-llama/Llama-2-13b-chat-hf
mistralai/Mistral-7B-Instruct-v0.2
mistralai/Mixtral-8x7B-Instruct-v0.1
HuggingFaceH4/zephyr-7b-beta
berkeley-nest/Starling-LM-7B-alpha
WizardLM/WizardLM-13B-V1.2
teknium/OpenHermes-2.5-Mistral-7B

Table 2: LLM checkpoints used in this study.

StarlingLM-13B-V1.2 (Zhu et al., 2023a) has been trained via Reinforcement Learning from AI Feedback (RLAIF) on the Nectar dataset. In its weight class, it is the second best performing model on chat.lmsys and 4th on AlpacaEval (as of Feb 6 2024).

Amidst mounting evidence that training on code enhances reasoning abilities also for natural language (Liang et al., 2023; Yang et al., 2024; Ma et al., 2023), we also use OpenHermes-2.5-Mistral-7B (NousResearch, 2023) which ranks third in its weight class on chat.lmsys. It is Mistral-based model that has been finetuned on additional code datasets. Notably, the developers detail that this results in improvements on non-code tasks.[13]

WizardLM-13B-V1.2 (Xu et al., 2023) is a fine-tuned version of Llama-2 13b and is ranked 8th in its weight-class on both chat.lmsys and AlpacaEval.

Zephyr-7b-beta (Tunstall et al., 2023) is a fine-tuned version of Mistral-7B-v0.1. It is ranked 9th on chat.lmsys and 11th on AlpacaEval.

D Average runtime

Generating LLM responses for one LLM and all generics across all settings took less than 0.5 GPU hours. All experiments were conducted on one NVIDIA A100 GPU.

E Statistical test results

Responses in the presence of exemplars are significantly different from results obtained without examplars (see Tables 3, 4, 5), for all types of exemplars and all models (significance level 0.01; sole exception is Llama-2 with CoT prompting as can be seen in Table 5 rows 1-2).

[13]https://huggingface.co/teknium/OpenHermes-2.5-Mistral-7B

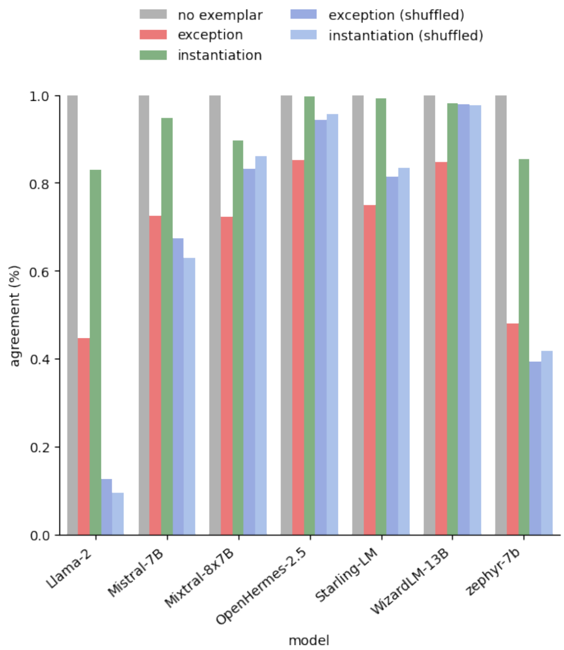

Figure 5: Results on GEN-comm. Alternative prompt template described in Section F

F Additional experimental results

We demonstrate additional experimental results based on an alternative prompting set-up in Figures 5 and 6.

To this end, we prompt LLMs using the following template where [INST] is an example of a model-specific special token used in chat templating. For example:

> Prompt
> *[INST] Do you believe that the following statement is accurate: 'Birds fly'*
>
> *Please answer yes or no. [/INST]*

For GEN-comm, we retain all generics to which an LLM responds *yes* to the prompt above. We then prompt LLMs anew supplying an exception, instantiation or random exemplar together with a generic for both datasets. For example:

> Prompt
> *[INST] Penguins do not fly.*
>
> *Do you believe that the following statement is accurate: 'Birds fly'*
>
> *Please answer yes or no. [/INST]*

We find that results differ significantly between

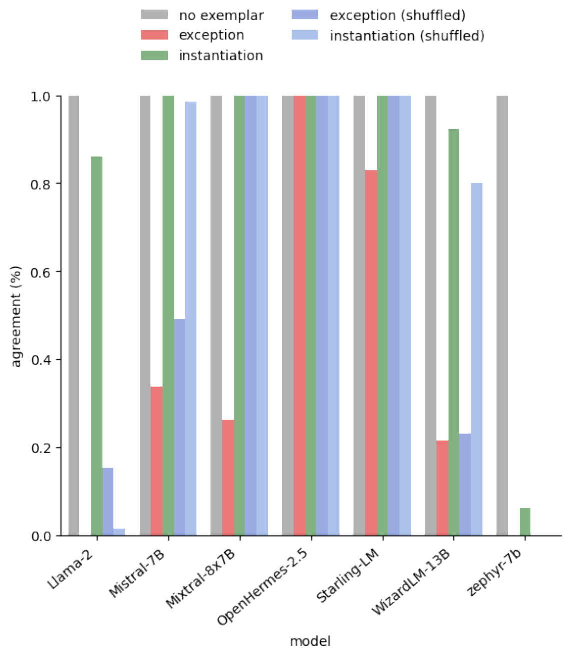

Figure 6: Results on `GEN-abs`. Alternative prompt template described in Section F.

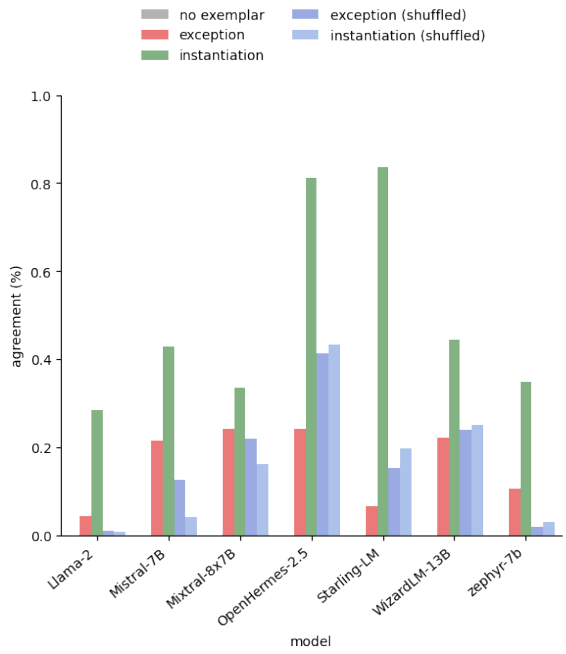

Figure 7: Results on generics of `GEN-comm` that are rejected by LLMs a priori. Alternative prompt template described in Section F. Missing bars indicate that agreement for 'no exemplar' is zero.

the two conditions (no exemplar vs. with an exemplar) (see Table 4 for statistical test results). On `GEN-comm` (Figure 5) agreement rates drop considerably in the presence of exceptions which mirrors nonmonotonic reasoning patterns. Agreement is higher, yet still drops significantly in the presence of instantiations. No LLM maintains perfectly consistent responses at the addition of random instantiations or exceptions. When prompting with random exemplars surprisingly agreement drops, most notably for Llama-2 and Zephyr.

For the reader's interest, we also include results on the portion of generics in `GEN-comm` which is rejected by LLMs a priori (Table 7). As expected, agreement increases from zero at the addition of an instantiation to the prompt, most notably for OpenHermes and Starling. However, LLMs should maintain a response of *no* at the addition of an *exception* or random exemplar to the prompt. This is visibly not the case with agreement rates increasing significantly for all models.

On `GEN-abs`, agreement drops considerably at the addition of an exception for all models except OpenHermes (Figure 6). Notably OpenHermes and Starling-LM appear to yield consistent responses in the presence of our controls, the random exemplars, while Llama-2 and Zephyr perform worst in that regard.

F.1 Chain-of-thought prompting

Additionally, we ran experiments using zero-shot Chain-of-Thought (CoT) prompting in the style of (Kojima et al., 2022) by appending 'Let's think step by step' to our prompts. We present results on `GEN-comm` in Figure 8 and results on `GEN-abs` in Figure 9.

On `GEN-comm`, agreement rates drop significantly for all models at the addition of exceptions, instantiations or shuffled exemplars (with the exception of Llama-2 when we include instantiations; see Table 5 for significance results). Agreement rates drop more given exceptions in comparison to instantiations or unrelated examplars for Mistral, Mixtral, OpenHermes and Starling. For Llama-2 and Zephyr agreement rates fall below 10% at the addition of unrelated exemplars.

On `GEN-abs`, agreement rates fall drastically given exceptions and equal 0% for Llama-2, Mixtral, Starling and Zephyr. The same is true for shuffled instantiations. OpenHermes is the only model to maintain agreement rates above 90% when presented with instantiations or shuffled exceptions.

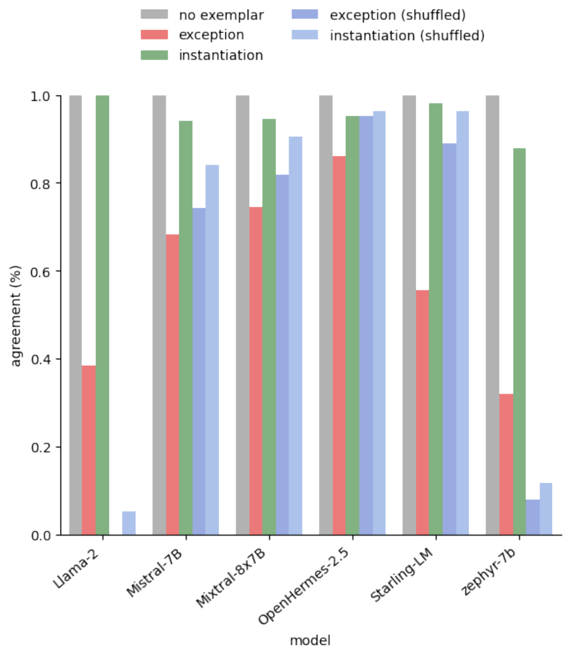

Figure 8: Results on `GEN-comm` using zero-shot CoT prompting.

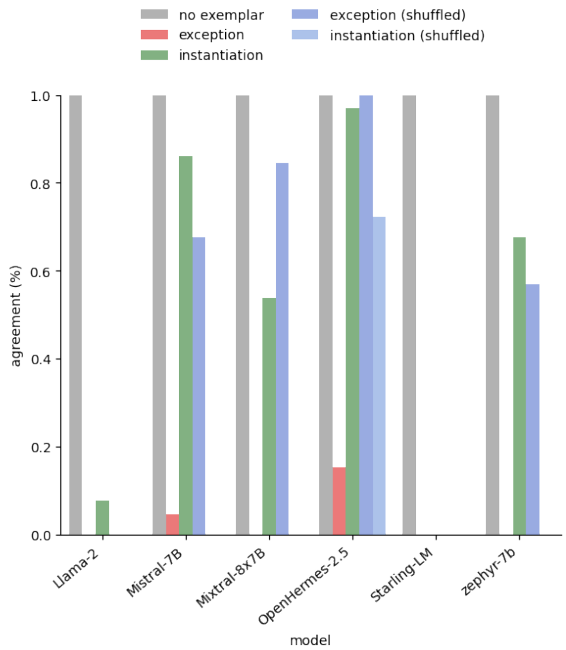

Figure 9: Results on `GEN-abs` using zero-shot CoT prompting. Missing bars indicate agreement rate of 0%.

Model	prompt setting	p-value
Llama-2-13b-chat-hf	exception	1.2444035588550786e-84
Llama-2-13b-chat-hf	instantiation	1.3944889010907487e-28
Llama-2-13b-chat-hf	exception (shuffled)	1.3664041679452567e-86
Llama-2-13b-chat-hf	instantiation (shuffled)	3.7504271121760947e-128
OpenHermes-2.5-Mistral-7B	exception	2.0884875837625446e-45
OpenHermes-2.5-Mistral-7B	instantiation	7.237829871739995e-08
OpenHermes-2.5-Mistral-7B	exception (shuffled)	1.733880104231141e-27
OpenHermes-2.5-Mistral-7B	instantiation (shuffled)	9.799073841979368e-26
Starling-LM-7B-alpha	exception	1.0691632340127197e-102
Starling-LM-7B-alpha	instantiation	7.247101964362887e-14
Starling-LM-7B-alpha	exception (shuffled)	3.14927364689666e-77
Starling-LM-7B-alpha	instantiation (shuffled)	5.588400099286033e-62
Mixtral-8x7B-Instruct-v0.1	exception	5.599059901868063e-84
Mixtral-8x7B-Instruct-v0.1	instantiation	4.84145282763492e-53
Mixtral-8x7B-Instruct-v0.1	exception (shuffled)	1.8855259265259482e-119
Mixtral-8x7B-Instruct-v0.1	instantiation (shuffled)	3.312378211336223e-151
WizardLM-13B-V1.2	exception	3.169934685227252e-109
WizardLM-13B-V1.2	instantiation	1.244192114854348e-15
WizardLM-13B-V1.2	exception (shuffled)	6.7440576522393956e-49
WizardLM-13B-V1.2	instantiation (shuffled)	3.312389179997469e-50
zephyr-7b-beta	exception	3.2434215158679907e-99
zephyr-7b-beta	instantiation	2.68778179464934e-25
zephyr-7b-beta	exception (shuffled)	2.7464111838608292e-137
zephyr-7b-beta	instantiation (shuffled)	2.671546422248841e-187
Mistral-7B-Instruct-v0.2	exception	6.521923113646968e-71
Mistral-7B-Instruct-v0.2	instantiation	2.0670658180782593e-15
Mistral-7B-Instruct-v0.2	exception (shuffled)	6.923699393684986e-120
Mistral-7B-Instruct-v0.2	instantiation (shuffled)	4.9982887921763924e-139

Table 3: Results of Wilcoxon signed ranked test for paired samples. We compare agreement of LLMs to generics with and without an exemplar (one of exception, instantiation, exception (shuffled), instantiation (shuffled). Results are obtained using the original prompt template described in section 5 and correspond to the main results in the paper in Figure 2.

Model	prompt setting	p-value
Llama-2-13b-chat-hf	exception	1.2402659787920488e-62
Llama-2-13b-chat-hf	instantiation	1.8577351435735865e-29
Llama-2-13b-chat-hf	exception (shuffled)	6.558556037957885e-98
Llama-2-13b-chat-hf	instantiation (shuffled)	9.990918651724453e-148
OpenHermes-2.5-Mistral-7B	exception	9.041178413936276e-31
OpenHermes-2.5-Mistral-7B	instantiation	0.025347318677468252
OpenHermes-2.5-Mistral-7B	exception (shuffled)	9.236596617174027e-13
OpenHermes-2.5-Mistral-7B	instantiation (shuffled)	1.2052982584446398e-13
Starling-LM-7B-alpha	exception	4.84145282763492e-53
Starling-LM-7B-alpha	instantiation	0.0009111188771537128
Starling-LM-7B-alpha	exception (shuffled)	9.89884333064868e-40
Starling-LM-7B-alpha	instantiation (shuffled)	6.7440576522393956e-49
Mixtral-8x7B-Instruct-v0.1	exception	2.6891242658680216e-51
Mixtral-8x7B-Instruct-v0.1	instantiation	2.8706760140807313e-27
Mixtral-8x7B-Instruct-v0.1	exception (shuffled)	7.287679729162835e-32
Mixtral-8x7B-Instruct-v0.1	instantiation (shuffled)	1.8712872006902566e-36
WizardLM-13B-V1.2	exception	5.8780179991539864e-33
WizardLM-13B-V1.2	instantiation	9.633570086430965e-07
WizardLM-13B-V1.2	exception (shuffled)	7.74421643104407e-06
WizardLM-13B-V1.2	instantiation (shuffled)	2.5802843041604163e-08
zephyr-7b-beta	exception	3.525239394844374e-74
zephyr-7b-beta	instantiation	2.476062658812572e-30
zephyr-7b-beta	exception (shuffled)	3.7238080067294776e-86
zephyr-7b-beta	instantiation (shuffled)	9.415767818703249e-116
Mistral-7B-Instruct-v0.2	exception	3.9328331793483447e-54
Mistral-7B-Instruct-v0.2	instantiation	2.0670658180782593e-15
Mistral-7B-Instruct-v0.2	exception (shuffled)	3.699479889932592e-64
Mistral-7B-Instruct-v0.2	instantiation (shuffled)	2.6476609044572044e-100

Table 4: Results of Wilcoxon signed ranked test for paired samples. We compare agreement of LLMs to generics with and without an exemplar (one of exception, instantiation, exception (shuffled), instantiation (shuffled)). These results correspond to the alternative prompting style and results described in section F.

Model	prompt setting	p-value
Llama-2-13b-chat-hf	exception	0.025347318677468252
Llama-2-13b-chat-hf	instantiation	0.31731050786291415
Llama-2-13b-chat-hf	exception (shuffled)	0.0009111188771537128
Llama-2-13b-chat-hf	instantiation (shuffled)	3.737981840170154e-05
Starling-LM-7B-alpha	exception	4.320463057827488e-08
Starling-LM-7B-alpha	instantiation	5.733031437583866e-07
Starling-LM-7B-alpha	exception (shuffled)	1.5417257900279904e-08
Starling-LM-7B-alpha	instantiation (shuffled)	1.1825298845719069e-11
OpenHermes-2.5-Mistral-7B	exception	2.3159484001346495e-35
OpenHermes-2.5-Mistral-7B	instantiation	3.552964224155306e-33
OpenHermes-2.5-Mistral-7B	exception (shuffled)	4.4044942248007814e-32
OpenHermes-2.5-Mistral-7B	instantiation (shuffled)	1.773177466197228e-41
Mixtral-8x7B-Instruct-v0.1	exception	2.9303133449994263e-53
Mixtral-8x7B-Instruct-v0.1	instantiation	4.474661339129513e-39
Mixtral-8x7B-Instruct-v0.1	exception (shuffled)	6.758775639492622e-37
Mixtral-8x7B-Instruct-v0.1	instantiation (shuffled)	5.058648827940248e-40
zephyr-7b-beta	exception	3.6136286243610392e-96
zephyr-7b-beta	instantiation	8.956226067732092e-94
zephyr-7b-beta	exception (shuffled)	1.2813208444193637e-111
zephyr-7b-beta	instantiation (shuffled)	2.0076004412348868e-151
Mistral-7B-Instruct-v0.2	exception	3.294362383314041e-67
Mistral-7B-Instruct-v0.2	instantiation	6.210993425425191e-19
Mistral-7B-Instruct-v0.2	exception (shuffled)	2.380470154600155e-54
Mistral-7B-Instruct-v0.2	instantiation (shuffled)	1.2444035588550786e-84

Table 5: Results of Wilcoxon signed ranked test for paired samples. We compare agreement of LLMs to generics with and without an exemplar (one of exception, instantiation, exception (shuffled), instantiation (shuffled)). These results correspond to Chain-of-Thought prompting results described in section F.