Evaluating the Deductive Competence of Large Language Models

Spencer M. Seals[1,2,3] and **Valerie L. Shalin**[1,4]
[1]Wright State University [2]Oak Ridge Institute for Science and Education
[3]Air Force Research Laboratory [4]AI Institute - University of South Carolina
s.m.seals@outlook.com

Abstract

The development of highly fluent large language models (LLMs) has prompted increased interest in assessing their reasoning and problem-solving capabilities. We investigate whether several LLMs can solve a classic type of deductive reasoning problem from the cognitive science literature. The tested LLMs have limited abilities to solve these problems in their conventional form. We performed follow up experiments to investigate if changes to the presentation format and content improve model performance. We do find performance differences between conditions; however, they do not improve overall performance. Moreover, we find that performance interacts with presentation format and content in unexpected ways that differ from human performance. Overall, our results suggest that LLMs have unique reasoning biases that are only partially predicted from human reasoning performance and the human-generated language corpora that informs them.

1 Introduction

[1] [2] The development and availability of highly fluent large language models (LLMs) (i.e., (Brown et al., 2020; Devlin et al., 2019; Ouyang et al., 2022; Zhang et al., 2022)) has increased interest in assessing their reasoning and problem solving abilities (Bhargava and Ng, 2022; Geva et al., 2020; Jumelet et al., 2019; Mitchell, 2021; Trinh and Le, 2019; Webb et al., 2022). Despite considerable performance improvements on benchmark tasks, LLMs exhibit mixed results on reasoning tasks. Some research has suggested that LLMs may have emergent reasoning abilities that enable better performance than those of human subjects (Webb et al., 2022). Other research has suggested that LLM reasoning performance is inconsistent and task dependent. Such research has suggested that some tasks, such as four term analogy problems (Mikolov et al., 2013) and different natural language inference tasks (Williams et al., 2018), are simply easier to solve. Other types of reasoning tasks such as analogy generation (Bhavya et al., 2022) and deductive competence (Dasgupta et al., 2022) are more challenging.

(Dasgupta et al., 2022) has investigated deductive competence in LLMs with characteristically mixed results. They demonstrated that one LLM, Chinchilla (Hoffmann et al., 2022), showed content effects on reasoning similar to human behavior documented in the cognitive science literature. For zero-shot performance, they found 50% accuracy for what they call realistic problems but chance accuracy for unrealistic problems. A 5-shot presentation resulted in some performance improvement for realistic problems, but performance on unrealistic problems remained low.

In this paper, we extend the previous research in several ways. First, we investigate the extent to which limited performance may be due to how the task was formatted. Prior research has demonstrated that overall performance can vary according to how a particular task is formatted (Gao et al., 2021; Jiang et al., 2021; Li and Liang, 2021; Shin et al., 2020). Research on analogy generation (Bhavya et al., 2022) demonstrates that performance depends on the specific prompt given to the models. Thus the inability of a model to solve one particular format of a task only provides a lower limit for assessing whether a model can successfully solve that task (Jiang et al., 2021).

Second, limited performance on deductive reasoning may be due to how the researchers employed content familiarity, of direct relevance to the distribution of content in the training corpus. For familiar content, they tested several different types of problems including social rules (i.e. *If peo-*

[1]https://github.com/spencer-michael-s/deductive-competence

[2]The views expressed are those of the author and do not necessarily reflect the official policy or position of the Department of the Air Force, the Department of Defense, or the U.S. government.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 8611–8627

Inference	Definition
Modus ponens	$P \implies Q, P \therefore Q$
Deny the antecedent	$P \implies Q, \neg P \therefore \neg Q$
Affirm the consequent	$P \implies Q, Q \therefore P$
Modus tollens	$P \implies Q, \neg Q \therefore \neg P$

Table 1: Four conditional inferences in the Wason Task

ple are driving, they must have a license) and other relationships (i.e., *If the plants are flowers, they must be fertilized*). However, content familiarity does not fully capture the content benefit seen in human subjects (Griggs and Cox, 1982; Manktelow and Evans, 1979). Instead, previous research indicates that people perform substantially better on problems that involve social rules than those that do not (Cosmides, 1989), even when the problems contain other types of familiar content (Griggs and Cox, 1982).

We expand the research base on the reasoning capabilities of LLMs by: 1) examining the role of specifically social-rules in reasoning about realistic content, 2) investigating the role of alternative presentation formats in deductive reasoning performance, and 3) expanding the set of candidate LLMs to evaluate potential algorithmic effects. Our results show that social content does benefit LLM performance, but not to the extent that might be expected based on a human sourced training corpus. While presentation formats do influence performance, they interact with content in a surprising (non-human) fashion. These findings are independent of the LLM examined.

2 Evaluating Deductive Competence

The Wason selection task is a reasoning task from the cognitive science literature that evaluates deductive competence (Wason, 1968). Participants are presented with a rule of the form *If p, then q* and four cards with p status on one side and q status on the other that correspond to the options $P, \neg P, Q$, and $\neg Q$. Participants are asked to determine which card or cards must be flipped over to validate whether the rule holds for this set of cards.

In the traditional *abstract* version of the task, participants are given rules about letters and numbers (i.e., *if there is an odd number on one side of the card, there is a vowel on the other side of the card*). The correct response requires the identification of *two* cards. Typical human accuracy for these problems is 10-20% with common errors consistent with confirmation bias. In contrast, problems that deal with a social rule (i.e., *if a person is drinking beer, they must be at least 21 years old*) are easy to solve- most participants (70%+) correctly select both cards (Griggs and Cox, 1982).

There are four potential conditional inferences in task: *modus ponens*, denial of the antecedent, affirmation of the consequent, and *modus tollens* (Evans, 2013). Of these four inferences, only *modus ponens* and *modus tollens* are logically valid. These inferences and their logical forms are illustrated in Table 1.

The Wason task makes a good candidate task for evaluating the reasoning performance of LLMs for several reasons. First, the task is relatively close to certain language modeling objectives. While the task involves a reasoning component, it can be formatted as a completion task, where the objective is to predict the answer given the problem text. This suggests that prior training, particularly for LLMs with high numbers of parameters that have been trained on large text corpora, should provide sufficient information for performing the Wason task. Moreover, the construction of the task minimizes the potential for confounds that may artificially inflate performance (Hovy and Yang, 2021; Mitchell and Krakauer, 2023; Rudinger et al., 2017). Previous work has demonstrated that high performance on some natural language inference tasks (Bowman et al., 2015; Williams et al., 2018) can be due to exploitable properties of the training data (Gururangan et al., 2018). The standardized format of the Wason task allows for the creation of a large number of carefully constructed examples without the risk that some answers may be easily determined from the original prompt alone.

Second, the problem examines both straightforward and challenging aspects of deductive competence. As noted above, a correct answer to a Wason problem involves two logical processes: *modus ponens* and *modus tollens*. Results from the cognitive science literature indicate that these rules are not equally difficult- applying *modus ponens* is considerably easier than applying *modus tollens*. The vast majority of people correctly apply *modus ponens*, even for difficult problems (i.e., (Wason, 1968; Griggs and Cox, 1982; Manktelow and Evans, 1979). In contrast, people fail to apply *modus tollens*, unless the problem has a particular form of semantic content associated with it. Similarly, we might expect presentation format to assist

LLMs.

Third, the way the task is constructed allows for careful examination of how problem content influences reasoning performance. Because LLMs are built on co-occurring content in human-sourced corpora, they should benefit from problems that contain familiar relationships. This should be especially true for problems where the relationship between the antecedent and the consequent is highly familiar. For example, in the rule *If a person is driving, they must have a driver's license*, the antecedent *driving a car*, and the consequent *driver's license* have a familiar (and commonly occurring) relationship. In comparison, the antecedent and the consequent in a rule such as *If a person is driving, they must have a book bag* do not have the same familiar relationship. While it is likely that some problems may be more difficult than others (i.e., because some completions are more probable) we control for this experimentally. We create sets of problems where both arguments contain realistic content, but the relationship between them is unfamiliar (see Appendix A).

Lastly, there is a large body of human performance literature on this task. This literature provides a comparison point for evaluating the performance of LLMs.

3 Experiments

In this section, we discuss the conditions, task format, models, and evaluation metrics associated with our experiments.

3.1 Task Conditions

We evaluate a total of 350 problems, 325 of which we created for this project. The remaining 25 were drawn from recent work (Dasgupta et al., 2022) and sorted into our problem categories. To facilitate comparison between content conditions, our problems are created as *matched sets*. For each condition (except the arbitrary condition), we create problems that are nearly identical in content except for the feature at issue and minor grammatical corrections. We evaluate three different types of problem content: **realistic**, **shuffled**, and **arbitrary**. A complete diagram of the different problems we evaluate is in Figure 7. Example problems for each condition are in Appendix A.

For the **realistic** condition, we evaluate a total 140 problems. Of these 140, 70 take the form of **social rules** and 70 take the form of **non-social rules**. Of the social rule problems, 35 problems take the form of **familiar social rules**. These problems are designed to take the form of social rules governing human behavior and are designed to be *familiar* such that they reflect social rules that are consistent with the real world. The other 35 social rule problems take the form of **unfamiliar social rules**. These problems have the form of social rules but do not contain familiar relationships.

Of the 70 non-social rule problems, 35 are designed to be **familiar non-social rules** and 35 are designed to be **unfamiliar non-social rules**. For the **familiar non-social rule** condition, we evaluate 35 problems that are not social rules and are familiar such that the antecedent and the consequent have a relationship that is consistent with real-world expectations. For the **unfamiliar non-social rule** condition, we evaluate 35 problems that do not take the form of social rules and are designed such that the antecedent and the consequent do not have a familiar real-world relationship.

The **realistic** grouping is intended to capture the same types of realistic problems that have been used in previous work (Dasgupta et al., 2022). For some of our analyses, we compare these problems as a group.

For the **shuffled** condition, we evaluate problems where the antecedent and the consequent are switched. We create shuffled versions of each of the different categories of realistic problems. The purpose of the shuffled condition is twofold. First, the creation of shuffled non-social rules allows for the ability to stress the semantics of plausibility beyond mere co-occurrence. Shuffled rules allows us to directly evaluate whether models are sensitive to the words in a problem or the intended semantic meaning. Shuffled problems contain the same words, but convey different conditional logic relationships. Second, the creation of shuffled social rules allows for evaluating sensitivity to the cost-benefit structure of social rules. Standard social rules typically have an if-then format such that if a person receives a benefit, then they must pay the (metaphorical) cost for that benefit, per (Cosmides, 1989). In comparison, switched social rules occur in past tense- if a person has paid the cost, then they may receive the benefit. We create shuffled prompts for both the familiar non-social rule and the familiar social rule conditions. We make syntactic corrections to make these problems grammatically correct.

For the **arbitrary** condition, we evaluate 70 problems where there is no particular relationship between the antecedent and the consequent. For example, in the problem *The rule is that if the cards have a type of food then they must have an outdoor activity*, there is no particular pre-supposed relationship between types of food and outdoor activities.

3.2 Task Format

A complete prompt for each problem consists of the instruction sentence, a context sentence, the rule, and a list of cards formatted as a multiple choice question. The instruction sentence was the same for all questions. The context sentences were consistent with the content type. For the arbitrary problems, a neutral context sentence was used to prevent a potential length confounds. The instruction sentence and a sample context sentence are in Appendix A.

3.3 Models

For our main set of analyses, we evaluate four recently released large language models with approximately 7 billion parameters: Guanaco, MPT, BLOOM, and Falcon. Guanaco is a family of LLMs fine-tuned with QLoRa, a fine tuning approach designed to reduce memory demands while preserving model performance (Dettmers et al., 2023). We use the 16-bit version. MPT is an open-source family of LLMs released by Mosaic that are designed to support fast inference (Team, 2023; Dao et al., 2022). BLOOM was trained on a large multilingual corpus (Laurençon et al., 2022) and has a decoder-only transformer architecture (Workshop et al., 2023). Falcon is an open source LLM designed to support fast inference (Almazrouei et al., 2023; Dao et al., 2022; Shazeer, 2019). We also run several additional LLMs of varying sizes on this task. Results for these models are in Appendix C. These models include: the 7B and 13B versions of llama2 (Touvron et al., 2023), the 7B, 13B, and 30B versions of Wizard (Xu et al., 2023), the 40B version of Falcon (Almazrouei et al., 2023), and the 13B and 33B versions of Guanaco (Dettmers et al., 2023). See the cited papers for updated details about model licenses.

3.4 Implementation Details

We run our experiments on a A100 GPU with 12 vCPUs and 85GB of RAM, running Debian 10. Experiments were conducted in Python 3.10. A complete list of libraries is available in the supplementary materials. Total run time for three main experiments was approximately 3 hours.

3.5 Evaluation Metrics

Previous work has proposed various methods to correct for interactions between the specific form of a prompt and the answer generated by a language model (Brown et al., 2020; Holtzman et al., 2021; Zhou et al., 2019). We use one of these metrics, Domain Conditional PMI, as our scoring metric (Holtzman et al., 2021). DCPMI measures how much information a particular instruction domain provides about a particular answer. Formally, a correct answer is equivalent to

$$argmax \frac{P(y_i|x)}{P(y_i|baseline)}$$

where y_i is the ith answer choice, x is the input prompt, and $baseline$ is the probability associated with a task-specific premise. Candidate answers are evaluated independently. Chance performance is $1/6$. Tables with both traditional accuracy metrics and DCMPI scores for all models can be found in Appendix C.

To facilitate comparison between content conditions, our problems are created as *matched sets*. We model this shared variance statistically via random effects terms for sets of stimuli. We use a mixed-effects approach, which allows for modeling the hierarchical structure of the data (Gelman, 2006). Mixed-effects models are commonly used to analyze linguistic data (i.e., (Matuschek et al., 2017; Baayen et al., 2008)) and permit the generalization of performance beyond a specific set of problems (Clark, 1973). We perform follow-up tests by calculating estimated marginal means derived from the entire statistical model for each corresponding analysis. Because estimated marginal means account for other variables in the statistical model, interaction terms may have slightly different coefficients in different analyses. See (Lenth, 2016) and (Searle et al., 1980) for additional information.

4 Results

4.1 Analysis 1 Results

For our first analysis, we evaluate two different crossed factors: social rule status and content type, using DCPMI as our scoring metric. For content type, we evaluate **arbitrary**, **shuffled**, and **realistic** rules. The realistic group contains social rules

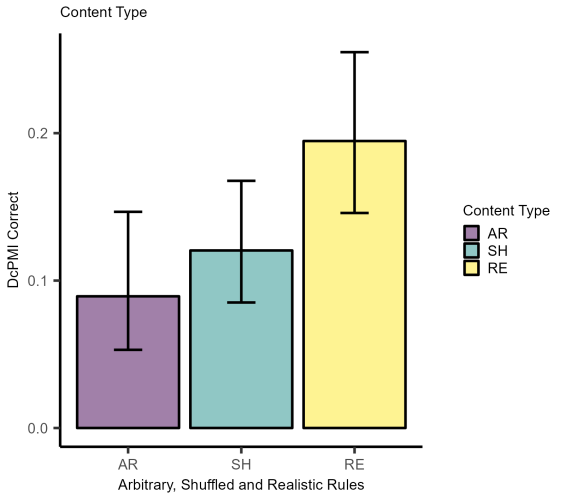

Figure 1: Model performance by content type for **arbitrary** (AR), **shuffled** (SH), and **realistic** (RE) rules. RE contains both social and non-social rules. Error bars represent 95 % confidence intervals. We do not find effects for LLM or familiarity, thus performance is collapsed. Relative to arbitrary content, most models result in a benefit for realistic rules, with mixed influences of shuffling.

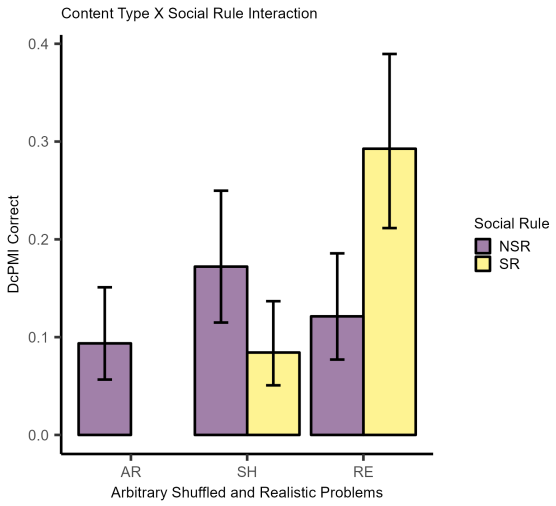

Figure 2: Interaction between content type and social rule status for Analysis 1. Content type: **arbitrary** (AR), **shuffled** (SH), or **realistic** (RE) rules. The realistic category contains social rules and non-social rules. Social rule status: **social rule** (SR) or **non- social rule** (NSR) problems. We do not find effects for LLM or familiarity, thus performance is collapsed.

Effect	**OR, CI,** Z
RE v SH	1.30 [1.05 - 1.62] 2.29*
SR Status x Content	1.59 [1.26 - 2.02] 3.97**
RE NSR v RE SR	0.33 [0.22 - 0.44] -3.34**
SH NSR v SH SR	2.25 [1.47 - 3.03] 2.35*
SH SR v RE SR	0.22 [0.15 - 0.29] -4.38**

Table 2: Statistical results for analysis 1. AR=arbitrary, SH=shuffled, RE=realistic, SR=social rule, NSR=non-social rule, OR=odds ratio, CI=confidence interval, * = $p < 0.05$, ** = $p < 0.01$. Interactions in bottom half.

and non-social rules. We find a significant beneficial main effect for realistic rules compared to shuffled rules and a significant interaction between social rule status and content type (Lines 1 and 2 in Table 2 respectively). Factors for LLM and familiarity do not improve model fit, suggesting that the overall pattern of results does not significantly differ between LLMs or between familiar and unfamiliar problems. See Appendix B for follow-up interaction tests. Overall performance is illustrated in Figure 1; see Figure 2 for the interaction. Performance remains rather low overall.

4.2 Analysis 1 Discussion

Initial results from analysis 1 do seem to replicate the general *pattern* of results demonstrated for human subjects: performance is better for social rule problems than non-social rule problems. We also find an effect for switched social rules such that switched social rules have considerably lower performance than standard social rule problems. This effect is similar to one reported for human subjects (Cosmides, 1989).

However, we do not replicate the *magnitude* of the content effect. While humans find abstract problems quite difficult, they successfully solve social rule problems ~70% of the time (Griggs and Cox, 1982). In comparison, the LLMs solve social rule problems approximately 30% of the time when using DCPMI scoring. Comparison with traditional highest probability scoring indicates that this pattern of results is not due to domain conditional scoring; highest probability scoring produces worse overall performance and the overall pattern of results is similar.

Additionally, we find an interaction between social rule status and content type. This interaction demonstrates that LLMs are sensitive to some aspects of the structure of social rules. This is somewhat consistent with results from human subjects which predict that performance should be lower for shuffled social rules compared to standard social rules (Cosmides, 1989). However, human subject responses do not predict the observed differences between shuffled social rules and shuffled non-social rules. Performance for LLMs is influ-

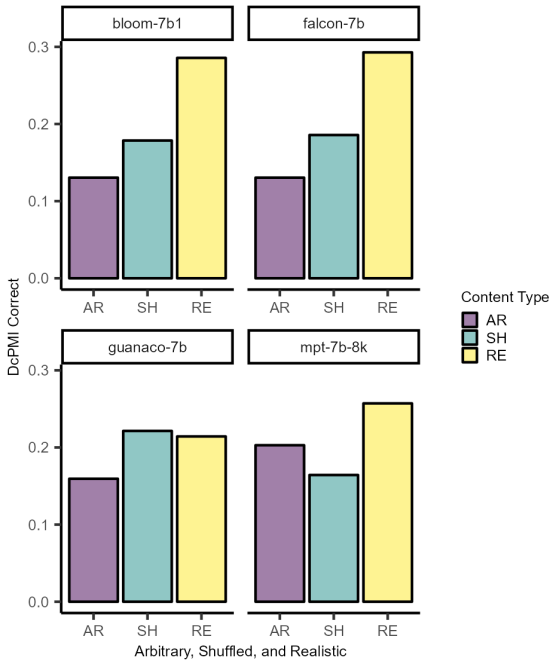

Figure 3: Performance across all models for **arbitrary** (AR), **shuffled** (SH), and **realistic** (RE) rules. The realistic category contains both social rules and non-social rules. We collapse across LLM and familiarity.

enced by problem content, but in a manner that is not parallel to human behavior.

4.3 Analysis 2 Results

A follow-up experiment examines whether alternative presentation formats may improve performance, given some previous results that suggest LLMs may have better reasoning performance with more explicit representations (e.g., Saparov et al., 2023). In addition to the standard presentation format (**classic**), we investigate three additional formats. In the **front** condition, the problems include descriptions of the front of each card. In the **back** condition, the problems include descriptions of the hypothetical category of the item on the back of the card. In the **both** condition, the problems include descriptions of both the front and the hypothetical category on the back of the card. We created alternative formats for all of the content types: **realistic social rules**, **realistic non-social rules**, **shuffled social rules**, and **shuffled non-social rules**. We use DCPMI as our scoring metric.

The best fitting statistical model includes an interaction between presentation format, social rule status, and content type plus a random effect for item instances. Adding factors for LLM and problem familiarity did not improve overall model fit,

Main Effect	**OR, CI,** Z
Classic v Front	1.19 [1.06 - 1.34] 2.79**
Front v Back	1.28 [1.09 - 1.50] 3.17**
Back v Both	1.15 [1.00 - 1.31] 1.98*
SR v NSR x F v B	1.22 [1.06 - 1.40] 2.6*
SR v NSR x B v Both	1.25 [1.09 - 1.44] 3.30**
SH v RE x C v F	1.22 [1.08 - 1.37] 2.96**
SH v RE x F v B	1.17 [1.02 - 1.34] 2.04*
SR v NSR x SH v RE	1.32 [1.17 - 1.48] 4.66**

Table 3: Statistical results for analysis 2. AR=arbitrary, SH=shuffled, RE=realistic, SR=social rule, NSR=non-social rule, OR=odds ratio, CI=confidence interval, * = $p < 0.05$, ** = $p < 0.01$. Interactions in bottom half.

suggesting that performance does not vary substantially by model or by familiarity. The main effect for presentation format was significant. Comparisons between classic versus front, front versus back, and back versus both presentation formats were all significant (lines 1, 2, and 3 respectively in the top half of Table 3). The main effects for social rule status and content type were not significant.

For the two-way interaction between social rule status and presentation format, comparisons between social rule status and front versus back presentation formats and social rule status and back versus both presentation formats were significant (lines 1 and 2 respectively in the bottom half of Table 3).

For the two-way interaction between content type and presentation format, the comparison between shuffled versus realistic content types and classic and front presentation formats and the comparison between shuffled versus realistic content types and front versus back presentation formats were significant (lines 3 and 4 respectively in the bottom half of Table 3).

The two-way interaction between social rule status and content type (shuffled or realistic) was significant (lines 5 in the bottom half of Table 3).

No three way interactions were significant (z-values $= (1.92, 1.88, 1.50)$, all $p > 0.05$). See Figures 4 and 5 for plots of the interactions. Follow up tests for individual level comparisons within each two-way interaction are located in Appendix B.2.

Problem familiarity did not improve model fit, suggesting that performance does not vary according to the familiarity of problem content.

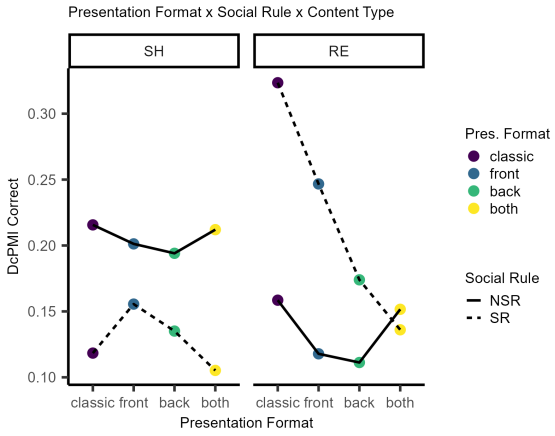

Figure 4: Interaction between presentation format (**classic**, **front**, **back**, or **both**), content type (**shuffled** (SH) or **realistic** (RE)), and social rule status (**social rule** or **non social rule**) broken out by presentation format.

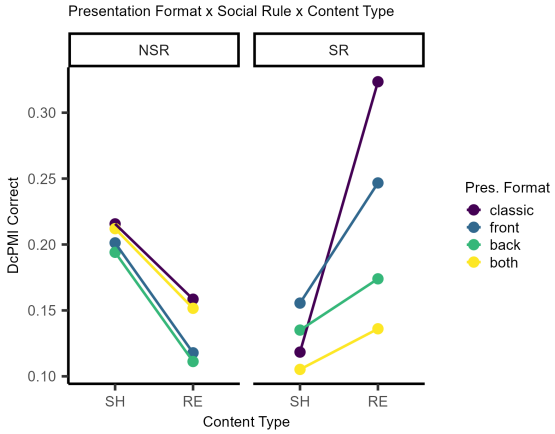

Figure 5: Interaction between presentation format (**classic**, **front**, **back**, or **both**), content type (**shuffled** (SH) or **realistic** (RE)), and social rule status (**social rule** or **non social rule**) broken out by content type.

4.4 Analysis 2 Discussion

In general, we find that models are sensitive to different presentation formats. However, we do not find performance improvements for different presentation formats. Results suggest that there are some interactions between presentation format and problem content; however, most of our follow up tests were not significant. A follow up analysis of treatment effects broken out by content type suggests that presentation formats have more of an effect on realistic rules than shuffled rules. Such interactions are not anticipated in human performance data (e.g., Wason and Green, 1984; Manktelow and Evans, 1979).

Effect	**OR, CI,** Z
C v F	1.34 [1.15 - 1.57] 3.76**
F v B	1.29 [1.10 - 1.51] 3.10**
B v Both	1.23 [1.07 - 1.41] 3.06*
AR v NAR	1.53 [1.23 - 1.90] 3.74**
SH v RE	1.30 [1.14 - 1.50] 3.59**

Table 4: Statistical results for antecedent selection in analysis 3. AR=arbitrary, NAR=non-arbitrary, SH=shuffled, RE=realistic, SR=social rule, NSR=non-social rule, C=classic, F=front, B=back, OR=odds ratio, CI=confidence interval, * = $p < 0.05$, ** = $p < 0.01$

4.5 Analysis 3 Results

For analysis 3, we examine antecedent selection, also scored with DCPMI, across all conditions. Specifically, we evaluate whether content type, presentation format, or social rule status influences whether the models select answers that contain any antecedent card. Complete statistical results are in Table 4.

The best-fitting statistical model contains a main effect for social rule status and an interaction between presentation format and content type, plus a random effect for overall item. A term for LLM did not improve model fit. For the presentation formats main effect, we find significant main effects for classic versus front presentation formats, front versus back presentation formats, and back versus both presentation formats (lines 1, 2, and 3 respectively in Table 4). For the content type main effect, we find significant differences between arbitrary versus non-arbitrary problems and between shuffled and realistic problems (lines 4 and 5 respectively in Table 4). The main effect for social rule status was not significant.

For the interaction between presentation format and content type, we find significant differences for contrasts between arbitrary and non-arbitrary problems and classic versus front presentation formats, shuffled versus realistic problems and classic versus front presentation formats, shuffled versus realistic problems and front versus back formats, shuffled versus realistic problems and back versus both formats (Lines 1-4 respectively in Table 5). Interactions are displayed in Figure 6. Follow-up tests for each of the interaction effects can be found in Table 7 in Appendix B.

4.6 Analysis 3 Discussion

Results from analysis 3 are particularly interesting because there is limited variance in human per-

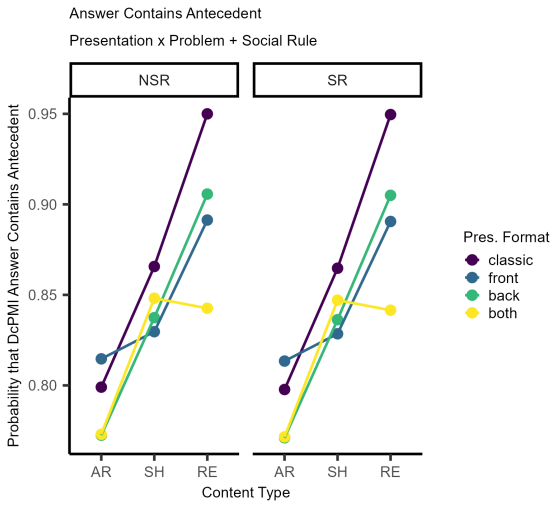

Figure 6: Evaluation of whether the LLMs select an antecedent card. Content type: **arbitrary** (AR), **shuffled** (SH), and **familiar** (FM). Presentation formats: **classic**, **front**, **back**, and **both**. Social rule status: **non-social rule**, **social rule**. Collapsed over LLM.

formance according to this metric. Regardless of content type, human subjects select at least one antecedent card (Johnson-Laird et al., 1972; Griggs and Cox, 1982). Even for conditions that influence antecedent card selection, the overwhelming tendency is for participants to select the alternative antecedent card (Cosmides, 1989). In contrast, our results suggest that whether LLMs select antecedent cards varies significantly according to content type and presentation format.

Despite differences in training datasets and tasks and model architecture, we do not find any effects for LLM. Additionally, the interaction between content type and presentation format suggests that different presentation formats have differential effects on antecedent selection for different content types. LLMs may benefit from different types of formatting depending on the content of the reasoning task.

5 Discussion

We set out to expand the research base on evaluating the reasoning capabilities of LLMs with a classic experiment contrasting *a priori* conditions. We do replicate some effects found in the literature for human subjects. Performance is higher for familiar problems than arbitrary ones and social rule have higher performance than non-social rules. However, we do not replicate the magnitude of the effects; social content does not benefit LLM performance as might be expected based on training corpus content.

We do find effects for different presentation formats; however, they do not improve performance. For shuffled problems, the specific presentation format does not make a difference. For realistic problems, we find presentation format does make a difference-the classic presentation format has the highest performance.

However, our systematic, content and format controlled experimentation and performance measurement has also revealed a number of inexplicable interactions that *appear to be consistent across different LLMs* and are therefore independent of architecture. Given the literature on human performance, an interaction between social rule status and content type is expected. However, many of the other interactions are not expected and are inconsistent with human performance (e.g., Wason and Green, 1984; Manktelow and Evans, 1979). We find some evidence that LLMs benefit from different types of presentation formats, (depending on the specific content of the problem), as might be expected from popular compensatory prompt engineering efforts. However, it is not immediately clear what types of information facilitate overall reasoning performance. This limits the ability to make general predictions about the conditions under which LLM reasoning is accurate.

In addition to content and presentation interactions, we find that LLMs do not pick antecedent cards at the same rate that human subjects do. Moreover, this behavior is influenced by the task condition- LLMs are less likely to select antecedent cards for arbitrary and realistic problems than for social rules.

Overall low performance is particularly surprising for realistic social and non-social rules as the relationships for solving these problems are plausibly available in the training data of LLMs. Yet we find that all LLMs diverge from documented human performance.

Overall performance is also remarkably consistent across models despite different training data, objectives, and model architectures. Moreover, we find that the interaction results are also independent of the LLM examined. Some consistency between LLMs is to be expected, given that LLMs are trained on human-generated text corpora. However, the commonalities are not consistent with human behavior. This suggests a common yet surprising emergent reasoning bias without any apparent adap-

tive benefit.

Fine-tuning the models for this task would likely improve task performance. We did not fine-tune the models for several reasons. First, the Wason task is intended to be a general task for evaluating reasoning performance. These tap into general knowledge and a set of reasoning skills that transfers to new tasks. Thus, the position that networks specifically trained on large reasoning task corpora is the best way to evaluate the reasoning performance of models is questionable. This is particularly true given that models often have high performance on the specific training task and limited performance on related reasoning tasks (Mitchell, 2021).

Second, several researchers have proposed that the Wason task can be solved via linguistic and real world knowledge (Pollard, 1982; Tversky and Kahneman, 1973; Wason, 1983). Human participants achieve high performance on problems that deal with familiar social rules with no experience with the task, using prior knowledge. However, the knowledge that this task requires is plausibly available in the training data for LLMs. The words used in this task are all common English words. Moreover, many of the relationships between items are plausibly available in training text, particularly for problems that deal with familiar social rules. Yet, performance remained quite low.

Previous work has proposed that ideal tasks for evaluating the reasoning of computational algorithms are those that do not require task-specific training (Chollet, 2019; Mitchell, 2021). We concur with this position and suggest that the Wason task is an ideal task in this regard.

6 Conclusion

Despite substantial performance improvements on standard benchmark datasets, existing LLMs have considerable room for improvement with regards to many aspects of human intelligence (Lake et al., 2017; Mitchell, 2021). In these experiments, we specifically investigate two of these aspects: generalized performance on related tasks and generation of answers at the limits of available knowledge.

Overall, our results replicate some of the same patterns found in the cognitive science literature. However, performance remains poor with inexplicable interactions between problem content and our efforts to manipulate presentation format. LLMs are sensitive to different sets of task criteria than human subjects. These criteria are not predictable across conditions and suggest areas where the reasoning of LLMs is not consistent with that of human capability.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

R. H. Baayen, D. J. Davidson, and D. M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412. Publisher: Elsevier Inc.

Prajjwal Bhargava and Vincent Ng. 2022. Commonsense Knowledge Reasoning and Generation with Pre-trained Language Models: A Survey. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12317–12325. Number: 11.

Bhavya Bhavya, Jinjun Xiong, and Chengxiang Zhai. 2022. Analogy Generation by Prompting Large Language Models: A Case Study of InstructGPT. ArXiv:2210.04186 [cs].

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

François Chollet. 2019. On the Measure of Intelligence. ArXiv:1911.01547 [cs].

Herbert H Clark. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12(4):335–359.

Leda Cosmides. 1989. The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, 31(3):187–276.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. ArXiv:2205.14135 [cs].

Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning. ArXiv:2207.07051 [cs].

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. ArXiv:2305.14314 [cs].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jonathan St B.T. Evans. 2013. Reasoning. In *Oxford Handbook of Cognitive Psychology*, pages 635–649.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Andrew Gelman. 2006. Multilevel (hierarchical) modeling: What It can and cannot do. *Technometrics*, 48(3):432–435.

Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. Injecting numerical reasoning skills into language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 946–958, Online. Association for Computational Linguistics.

Richard A. Griggs and James R. Cox. 1982. The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology*, 73(3):407–420.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

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, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training Compute-Optimal Large Language Models. ArXiv:2203.15556 [cs].

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

P. N. Johnson-Laird, Paolo Legrenzi, and Maria Sonino Legrenzi. 1972. Reasoning and a sense of reality. *British journal of psychology*, 63(3):395–400. Publisher: Wiley-Blackwell.

Jaap Jumelet, Willem Zuidema, and Dieuwke Hupkes. 2019. Analysing neural language models: Contextual decomposition reveals default reasoning in number and gender assignment. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 1–11, Hong Kong, China. Association for Computational Linguistics.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret

Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. 2022. The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

Russell V. Lenth. 2016. Least-Squares Means: The *R* Package **lsmeans**. *Journal of Statistical Software*, 69(1).

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

K. I. Manktelow and J. St B. T. Evans. 1979. Facilitation of reasoning by realism: Effect or non-effect? *British Journal of Psychology*, 70(4):477–488.

Hannes Matuschek, Reinhold Kliegl, Shravan Vasishth, Harald Baayen, and Douglas Bates. 2017. Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, 94:305–315. ArXiv: 1511.01864 Publisher: The Authors.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Melanie Mitchell. 2021. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101.

Melanie Mitchell and David C. Krakauer. 2023. The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. ArXiv:2203.02155 [cs].

P. Pollard. 1982. Human reasoning: Some possible effects of availability. *Cognition*, 12(1):65–96.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Seyed Mehran Kazemi, Najoung Kim, and He He. 2023. Testing the General Deductive Reasoning Capacity of Large Language Models Using OOD Examples. ArXiv:2305.15269 [cs].

S. R. Searle, F. M. Speed, and G. A. Milliken. 1980. Population Marginal Means in the Linear Model: An Alternative to Least Squares Means. *The American Statistician*, 34(4):216.

Noam Shazeer. 2019. Fast Transformer Decoding: One Write-Head is All You Need. ArXiv:1911.02150 [cs].

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

MosaicML NLP Team. 2023. Introducing MPT-7B: A new standard for open-source, commercially usable LLMs.

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

Trieu H. Trinh and Quoc V. Le. 2019. A Simple Method for Commonsense Reasoning. ArXiv:1806.02847 [cs].

Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2):207–232.

P. C. Wason. 1968. Reasoning about a Rule. *Quarterly Journal of Experimental Psychology*, 20(3):273–281. Publisher: SAGE Publications.

P. C. Wason and D. W. Green. 1984. Reasoning and Mental Representation. *The Quarterly Journal of Experimental Psychology Section A*, 36(4):597–610.

Peter C. Wason. 1983. Realism and rationality in the selection task. In *Thinking and Reasoning (Psychology Revivals)*. Psychology Press. Num Pages: 32.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. 2022. Emergent Analogical Reasoning in Large Language Models. ArXiv:2212.09196 [cs].

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. ArXiv:1704.05426 [cs].

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M. Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A. Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S. Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tan-

may Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. ArXiv:2211.05100 [cs].

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. ArXiv:2304.12244 [cs].

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pre-trained Transformer Language Models. ArXiv:2205.01068 [cs].

Kun Zhou, Kai Zhang, Yu Wu, Shujie Liu, and Jingsong Yu. 2019. Unsupervised context rewriting for open domain conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1834–1844, Hong Kong, China. Association for Computational Linguistics.

A Example Problems

In this section, we provide example problems. For details on the conditions and their associated rationale, see Section 1.

A.1 Definitions

Below are definitions and examples of the different problem components. A complete design matrix for the study is provided in 7.

Context Sentence: Used as context for for the problem. Example: *An attendant needs to make sure that customers are following the rules.*

Instruction sentence: Used at the end of all problems to prompt the model. Example: *Pick two cards that are required to determine if the rule is true.*

Familiar Social Rule: Problems that take the form of a social rule (i.e., *If a person pays the cost, they receive the benefit*) that deals with familiar relationships. Example: *The rule is that if the customer is over 25 they can drive a rental car.*

Unfamiliar Social Rule: Problems that take the form of a social rule (i.e., *If a person pays the cost, they receive the benefit*) and deals with unfamiliar relationships. Example: *The rule is that if the customer is over 25 they must be in elementary school.*

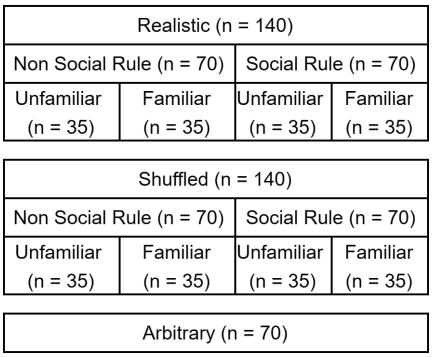

Figure 7: Breakdown of the different types of problems we examine.

Familiar Non-social Rule: Problems that do not take the form of a social rule and deal with familiar relationships. Example: *The rule is that if the equipment is a laptop then it must have a plastic keyboard.*

Unfamiliar Non-social Rule: Problems that do not take the form of a social rule and deal with unfamiliar relationships. Example: *The rule is that if the equipment is a laptop then it must have a grass keyboard.*

Shuffled Rules: Created from the rule types above. Shuffling allows us to evaluate the extent to which models are sensitive to the cost benefit structure of the rules. Example **shuffled familiar social rule**: *The rule is that if the customer can drive a rental car they must be over 25.* **Arbitrary**: Rules that contain arbitrary relationships. Example: *The rule is that if the cards have a type of food then they must have an outdoor activity.*

A.2 Complete Example Problems

Below are complete example problems for each of the presentation conditions. Note that the order of the answers was shuffled in the actual experiments, answers are presented in here the $P, \neg P, Q, \neg Q$ order for all problems for ease of interpretation.

Familiar Social Rule (Classic): *Context:* Pick two cards that are required to determine if the rule is true: *Problem:* An attendant needs to make sure that customers are following the rules. The rule is that if the customer is over 25 they can drive a rental car: A: 27 years old B: 15 years old C: driving rental car D: riding the bus

Familiar Social Rule (Front): *Context:* Pick two cards that are required to determine if the rule is true: *Problem:* An attendant needs to make sure

that customers are following the rules. The rule is that if the customer is over 25 they can drive a rental car. The first card has 15 years old on the front. The second card has driving rental car on the front. The third card has riding the bus on the front. The fourth card has 27 years old on the front: A: 27 years old B: 15 years old C: driving rental car D: riding the bus

Familiar Social Rule (Back): *Context:* Pick two cards that are required to determine if the rule is true: *Problem:* An attendant needs to make sure that customers are following the rules. The rule is that if the customer is over 25 they can drive a rental car. The first card has an unknown age on the back. The second card has an unknown activity on the back. The third card has an unknown age on the back. The fourth card has an unknown activity on the back: A: 27 years old B: 15 years old C: driving rental car D: riding the bus

Familiar Social Rule (Both): *Context:* Pick two cards that are required to determine if the rule is true: *Problem:* An attendant needs to make sure that customers are following the rules. The rule is that if the customer is over 25 they can drive a rental car. The first card has riding the bus on the front and an unknown age on the back. The second card has driving rental car on the front and an unknown age on the back. The third card has 15 years old on the front and an unknown activity on the back. The fourth card has 27 years old on the front and an unknown activity on the back: A: 27 years old B: 15 years old C: driving rental car D: riding the bus

B Follow-up Interaction Tests

B.1 Analysis 1 Follow-up Tests

Results from analysis 1 found support for one interaction between social rule status and content type.

B.1.1 Content Type and Social Rule Status

Three of the follow-up tests for the interaction between social rule status and content type were significant. The test comparing realistic non-social rules and realistic social rules test was significant (Bottom half of Table 2, line 1). The test comparing shuffled non-social rules and shuffled social rules was significant (Bottom half of Table 2, line 2). The test comparing shuffled social rules and realistic social rules was significant (Bottom half of Table 2, line 3).

Interaction Effect	**OR, CI,** Z
AR v NAR x C v F	1.28 [1.03 - 1.59] 2.17*
SH v RE x C v F	1.29 [1.08 - 1.54] 2.74**
SH v RE x F v B	1.28 [1.05 - 1.56] 2.54*
SH v RE x B v Both	1.33 [1.14-1.56] 3.61**

Table 5: Effects for interactions for Analysis 3. AR=arbitrary, NAR=non-arbitrary, SH=shuffled, RE=realistic, SR=social rule, NSR=non-social rule, C=classic, F=front, B=back, OR=odds ratio, CI=confidence interval, * = $p < 0.05$, ** = $p < 0.01$

B.2 Analysis 2 Follow-up Tests

Results from analysis 2 found support for three two-way interactions: one between social rule status and presentation format, one between content type and presentation format, and one between social rule status and content type.

B.2.1 Social Rule Status and Presentation Format

None of the follow-up tests were significant for the front versus back or the back versus both presentation format interactions with social rule status.

B.2.2 Content Type and Presentation Format

Follow-up tests for the shuffled versus realistic rules and classic versus front presentation format contrast found significant differences between shuffled and realistic rules for the classic problems (line 1 Table 6) and other tests were non-significant.

All follow-up tests for the shuffled versus realistic rules and the front versus back presentation format contrast were non-significant.

We conducted a follow-up set of comparisons to examine differences within the shuffled and realistic content types. We found significant differences between social rules within the both presentation format (line 4 Table 6). For realistic rules, all four comparisons were significant (lines 5-8 Table 6).

B.2.3 Social Rule Status and Content Type

For the interaction between social rule status and content type (shuffled versus realistic), follow-up tests between shuffled vs realistic non-social rules (line 2 Table 6) and realistic social rules vs shuffled social rules were significant (line 3 Table 6).

B.3 Analysis 3 Follow-up Tests

Analysis 3 found support for an interaction between content type (arbitrary, shuffled, or realistic) and presentation format (classic, front, back, or both).

Interaction Effect	OR, Z, CI
SH C v RE C	1.56, 2.5*, [1.29-1.83]
SH SR v RE NSR	1.68, 3.04**, [1.39 - 1.96]
RE SR v SH SR	1.84, 3.55**, [1.53 - 2.15]
SH SR both	0.60, -2.78*
RE NSR F	0.65, -2.40*
NSR B	0.61, -2.73*
SR C	2.35, 5.95**
SR F	1.61, 3.20**

Table 6: Follow-up interaction tests for Analysis 2. AR=arbitrary, NAR=non-arbitrary, SH=shuffled, RE=realistic, SR=social rule, NSR=non-social rule, C=classic, F=front, B=back, OR=odds ratio, CI=confidence interval, * = $p < 0.05$, ** = $p < 0.01$

Follow-up Test	OR, CI, Z
AR C v NAR C	2.73 [1.49-5.19] 4.09**
NAR C v NAR F	0.57 [0.39-0.83] -3.68**
SH C v RE C	0.33 [0.17-0.65] -4.091**
SH F v RE F	0.59 [0.34-1.0] -2.35*
SH F v RE C	0.25 [0.13-0.49] -5.23*
SH F v RE F	0.59 [0.34-1.0] -2.35*
SH B v RE B	0.53 [0.30-0.93] -2.77*
SH B v RE F	0.62 [0.36-1.0] -2.09*
SH F v RE B	0.50 [0.29-0.88] -3.03**
SH Back v RE Back	0.53 [0.30-0.93] -2.77*
SH Both v RE Back	0.58 [0.33-1.0] -2.41*

Table 7: Follow-up interaction tests for analysis 3. AR=arbitrary, NAR=non-arbitrary, SH=shuffled, RE=realistic, SR=social rule, NSR=non-social rule, C=classic, F=front, B=back, OR=odds ratio, CI=confidence interval, * = $p < 0.05$, ** = $p < 0.01$

B.3.1 Content Type and Presentation Format

Tests for this interaction indicated significant differences for four contrasts: one between arbitrary and non-arbitrary content types and classic versus front presentation formats (1), one between shuffled versus realistic content types and classic versus front presentation formats (2), one between shuffled versus realistic content types and front versus back presentation formats (3), and one between shuffled versus realistic content types and back versus both presentation formats (4).

Contrast 1: For the contrast between arbitrary and non-arbitrary content types and classic versus front presentation formats, two follow-up tests were significant. These include: 1) the test between arbitrary classic versus non-arbitrary classic (line 1, Table 7) and 2) the test between non-arbitrary classic and non-arbitrary front (line 2, Table 7).

Contrast 2: For the contrast between shuffled and realistic content types and classic versus front presentation formats, three of the follow-up tests were significant. These include: the test between shuffled classic and realistic classic (line 3, Table 7), shuffled front and realistic front (line 4, Table 7), and shuffled front and realistic classic (line 5, Table 7).

Contrast 3: For the contrast between shuffled and realistic content types and front versus back presentation formats, all four follow up tests were significant. These include: shuffled front versus realistic front (line 6, Table 7), shuffled back versus realistic back (line 7, Table 7), shuffled back versus realistic front (line 8, Table 7), and shuffled front versus realistic back (line 9, Table 7).

Contrast 4: For the contrast between shuffled versus realistic content types and back versus both presentation formats, two follow-up tests were significant. These include: the follow up test between shuffled back versus realistic back (line 10, Table 7) and shuffled both versus realistic back (line 11, Table 7). Interaction plots are in Figure 6.

C Additional Models

In this section, we report full accuracy (see Table 8) and domain-conditional PMI (see Table 9) scores for all the models in the main paper, as well as several additional models. While we do find that some of the larger models perform somewhat better, the overall pattern of results is similar.

Model	Cond	Classic	Front	Back	Both
guanaco-7b	AR	0.1 (0.3)	0.14 (0.35)	0.16 (0.37)	0.09 (0.28)
guanaco-7b	SH	0.1 (0.3)	0.16 (0.37)	0.14 (0.34)	0.14 (0.34)
guanaco-7b	SR	0.06 (0.25)	0.09 (0.29)	0.09 (0.29)	0.07 (0.26)
mpt-7b-8k	AR	0.06 (0.23)	0.21 (0.41)	0.09 (0.28)	0.09 (0.28)
mpt-7b-8k	SH	0.09 (0.28)	0.14 (0.34)	0.09 (0.29)	0.19 (0.39)
mpt-7b-8k	SR	0.14 (0.34)	0.15 (0.36)	0.15 (0.36)	0.11 (0.32)
bloom-7b1	AR	0.09 (0.28)	0.24 (0.43)	0.13 (0.34)	0.11 (0.32)
bloom-7b1	SH	0.13 (0.34)	0.13 (0.34)	0.07 (0.26)	0.15 (0.36)
bloom-7b1	SR	0.16 (0.37)	0.16 (0.37)	0.11 (0.32)	0.08 (0.27)
falcon-7b	AR	0.11 (0.32)	0.21 (0.41)	0.06 (0.23)	0.06 (0.23)
falcon-7b	SH	0.09 (0.28)	0.11 (0.32)	0.07 (0.26)	0.11 (0.32)
falcon-7b	SR	0.14 (0.34)	0.16 (0.37)	0.13 (0.34)	0.11 (0.32)
WizardLM-7B-V1.0	AR	0.17 (0.38)	0.26 (0.44)	0.29 (0.46)	0.26 (0.44)
WizardLM-7B-V1.0	SH	0.15 (0.36)	0.12 (0.33)	0.11 (0.31)	0.2 (0.4)
WizardLM-7B-V1.0	SR	0.16 (0.37)	0.15 (0.36)	0.16 (0.37)	0.25 (0.43)
Llama-2-7b-hf	AR	0.11 (0.32)	0.11 (0.32)	0.13 (0.34)	0.16 (0.37)
Llama-2-7b-hf	SH	0.13 (0.34)	0.18 (0.38)	0.11 (0.32)	0.17 (0.38)
Llama-2-7b-hf	SR	0.1 (0.3)	0.07 (0.26)	0.16 (0.37)	0.11 (0.31)
guanaco-13b	AR	0.11 (0.32)	0.11 (0.32)	0.11 (0.32)	0.16 (0.37)
guanaco-13b	SH	0.16 (0.37)	0.16 (0.37)	0.17 (0.38)	0.16 (0.37)
guanaco-13b	SR	0.14 (0.34)	0.11 (0.32)	0.14 (0.34)	0.14 (0.34)
guanaco-33b-merged	AR	0.13 (0.34)		0.19 (0.39)	
guanaco-33b-merged	SH	0.24 (0.43)			
guanaco-33b-merged	SR	0.16 (0.37)			
mpt-30b	AR	0.1 (0.3)	0.17 (0.38)	0.09 (0.28)	0.1 (0.3)
mpt-30b	SH	0.17 (0.38)	0.16 (0.37)	0.14 (0.34)	0.12 (0.33)
mpt-30b	SR	0.08 (0.27)	0.13 (0.34)	0.07 (0.26)	0.1 (0.3)
WizardLM-13B-V1.2	AR	0.13 (0.34)	0.24 (0.43)	0.17 (0.38)	0.13 (0.34)
WizardLM-13B-V1.2	SH	0.13 (0.34)	0.16 (0.37)	0.1 (0.3)	0.19 (0.39)
WizardLM-13B-V1.2	SR	0.15 (0.36)	0.18 (0.38)	0.14 (0.35)	0.2 (0.4)
WizardLM-30B-V1.0	AR	0.11 (0.32)	0.17 (0.38)	0.14 (0.35)	0.11 (0.32)
WizardLM-30B-V1.0	SR	0.13 (0.34)	0.14 (0.34)	0.18 (0.38)	0.13 (0.34)
falcon-40b	AR	0.1 (0.3)	0.21 (0.41)	0.09 (0.28)	
falcon-40b	SR	0.13 (0.34)	0.14 (0.34)	0.13 (0.34)	0.14 (0.34)
Llama-2-13b-hf	AR	0.07 (0.26)	0.14 (0.35)	0.06 (0.23)	0.09 (0.28)
Llama-2-13b-hf	SH	0.06 (0.25)	0.06 (0.25)	0.06 (0.23)	0.09 (0.29)
Llama-2-13b-hf	SR	0.1 (0.3)	0.11 (0.31)	0.12 (0.33)	0.12 (0.33)

Table 8: Accuracy metrics for all models tested. mean(sd)

Model	Cond	Classic	Front	Back	Both
bloom-7b1	AR	0.13 (0.34)	0.07 (0.26)	0.2 (0.4)	0.14 (0.35)
bloom-7b1	SH	0.18 (0.38)	0.21 (0.41)	0.16 (0.37)	0.19 (0.39)
bloom-7b1	SR	0.29 (0.45)	0.16 (0.37)	0.18 (0.38)	0.21 (0.41)
falcon-7b	AR	0.13 (0.34)	0.17 (0.38)	0.16 (0.37)	0.09 (0.28)
falcon-7b	SH	0.19 (0.39)	0.18 (0.38)	0.21 (0.41)	0.14 (0.35)
falcon-7b	SR	0.29 (0.46)	0.21 (0.41)	0.16 (0.37)	0.19 (0.4)
guanaco-7b	AR	0.16 (0.37)	0.14 (0.35)	0.14 (0.35)	0.13 (0.34)
guanaco-7b	SH	0.22 (0.42)	0.2 (0.4)	0.16 (0.37)	0.19 (0.4)
guanaco-7b	SR	0.21 (0.41)	0.24 (0.43)	0.15 (0.36)	0.14 (0.34)
mpt-7b-8k	AR	0.2 (0.4)	0.16 (0.37)	0.2 (0.4)	0.11 (0.32)
mpt-7b-8k	SH	0.16 (0.37)	0.21 (0.41)	0.21 (0.41)	0.19 (0.4)
mpt-7b-8k	SR	0.26 (0.44)	0.21 (0.41)	0.18 (0.38)	0.13 (0.34)
Llama-2-7b-hf	AR	0.2 (0.4)	0.13 (0.34)	0.19 (0.39)	0.19 (0.39)
Llama-2-7b-hf	SH	0.16 (0.37)	0.18 (0.38)	0.16 (0.37)	0.14 (0.34)
Llama-2-7b-hf	SR	0.16 (0.37)	0.15 (0.36)	0.16 (0.37)	0.14 (0.34)
WizardLM-7B-V1.0	AR	0.14 (0.35)	0.3 (0.46)	0.19 (0.39)	0.19 (0.39)
WizardLM-7B-V1.0	SH	0.09 (0.29)	0.15 (0.36)	0.11 (0.32)	0.19 (0.4)
WizardLM-7B-V1.0	SR	0.19 (0.39)	0.18 (0.38)	0.16 (0.37)	0.21 (0.41)
guanaco-13b	AR	0.23 (0.42)	0.2 (0.4)	0.23 (0.42)	0.17 (0.38)
guanaco-13b	SH	0.17 (0.38)	0.17 (0.38)	0.22 (0.42)	0.21 (0.41)
guanaco-13b	SR	0.19 (0.4)	0.14 (0.35)	0.14 (0.35)	0.19 (0.4)
guanaco-33b-merged	SH	0.24 (0.43)			
guanaco-33b-merged	SR	0.19 (0.4)			
mpt-30b	AR	0.2 (0.4)	0.19 (0.39)	0.19 (0.39)	0.16 (0.37)
mpt-30b	SH	0.17 (0.38)	0.16 (0.37)	0.14 (0.34)	0.12 (0.33)
mpt-30b	SR	0.17 (0.38)	0.16 (0.37)	0.14 (0.34)	0.2 (0.4)
falcon-40b	AR	0.17 (0.38)	0.19 (0.39)	0.23 (0.42)	
falcon-40b	SR	0.16 (0.37)	0.19 (0.39)	0.1 (0.3)	0.11 (0.32)
Llama-2-13b-hf	AR	0.24 (0.43)	0.11 (0.32)	0.16 (0.37)	0.16 (0.37)
Llama-2-13b-hf	SH	0.24 (0.43)	0.2 (0.4)	0.23 (0.42)	0.22 (0.42)
Llama-2-13b-hf	SR	0.21 (0.41)	0.14 (0.35)	0.19 (0.4)	0.16 (0.37)
WizardLM-13B-V1.2	AR	0.17 (0.38)	0.27 (0.45)	0.21 (0.41)	0.2 (0.4)
WizardLM-13B-V1.2	SH	0.12 (0.33)	0.15 (0.36)	0.12 (0.33)	0.23 (0.42)
WizardLM-13B-V1.2	SR	0.15 (0.36)	0.14 (0.34)	0.14 (0.34)	0.15 (0.36)
WizardLM-30B-V1.0	AR	0.11 (0.32)	0.17 (0.38)	0.14 (0.35)	0.11 (0.32)
WizardLM-30B-V1.0	SH	0.1 (0.3)	0.19 (0.4)	0.12 (0.33)	0.13 (0.34)
WizardLM-30B-V1.0	SR	0.14 (0.34)	0.19 (0.39)	0.11 (0.32)	0.14 (0.35)

Table 9: DCPMI metrics for all models tested. mean(sd)