



Article

Assessing the Relational Abilities of Large Language Models and Large Reasoning Models

Matthias Raemaekers *, Martin Finn and Jan De Houwer

Department of Experimental Clinical and Health Psychology, Ghent University, 9000 Ghent, Belgium; martin.finn@ugent.be (M.F.); jan.dehouwer@ugent.be (J.D.H.)

* Correspondence: matthias.raemaekers@ugent.be

Abstract

We assessed the relational abilities of two state-of-the-art large language models (LLMs) and two large reasoning models (LRMs) using a new battery of several thousand syllogistic problems, similar to those used in behavior-analytic tasks for relational abilities. To probe the models' general (as opposed to task- or domain-specific) abilities, the problems involved multiple relations (sameness, difference, comparison, hierarchy, analogy, temporal and deictic), specified between randomly selected nonwords and varied in terms of complexity (number of premises, inclusion of irrelevant premises) and format (valid or invalid conclusion prompted). We also tested transformations of stimulus function. Our results show that the models generally performed well in this new task battery. The models did show some variability across different relations and were to a limited extent affected by task variations. Model performance was, however, robust against the randomization of premise order in a replication study. Our research provides a new framework for testing a core aspect of intellectual (i.e., relational) abilities in artificial systems; we discuss the implications of this and future research directions.

Keywords: large language models; reasoning models; relational reasoning; relational abilities index; transformation of function

1. Introduction

Large language models (LLMs, e.g., GPT4, Open AI et al., 2023; LLaMa 3, Dubey et al., 2024) and large reasoning models (LRMs, e.g., Open AI o1, Open AI et al., 2024; DeepSeek R1, Guo et al., 2025) have taken the world by storm. Their ability to produce human-like text, answer questions and solve problems (and much more) has inspired much debate about how intelligent such systems are, and to what extent this aligns with human intelligence. While many have taken their performance as an indication that superhuman (or even general) intelligence is within reach (e.g., Bengio et al., 2024), such conclusions are nuanced by reports of those same systems failing on tasks that are trivial for the average human (e.g., Borji, 2023; McCoy et al., 2024; Berglund et al., 2023). The debate about these systems' level of intelligence is complicated by a range of issues. First, there is little consensus about a definition of human intelligence (e.g., Sternberg & Detterman, 1987; Legg & Hutter, 2007), let alone artificial or general intelligence (e.g., Wang, 2019; Chollet, 2019; Morris et al., 2023), which leads to inconsistencies in how different perspectives interpret the same findings. Second, the fact that these language models are so fluent at producing human-like text, and that they are often tested on tasks

Academic Editor: John Parkinson

Received: 30 October 2025

Revised: 12 December 2025

Accepted: 18 December 2025

Published: 25 December 2025

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

designed for humans (e.g., the bar exam, intelligence tests) brings the risk of anthropomorphizing these systems (Crockett & Messeri, 2023; Shiffrin & Mitchell, 2023), attributing human-like understanding, intelligence and even goals, purpose and sentience to them (often referred to as the ELIZA-effect). However, the alignment between human and artificial system's responses does not a priori reflect alignment of abilities or the mechanism underlying them (McCoy et al., 2024; Lin, 2025). It is important to see these models for what they are—large neural network models that are trained to predict the next token in a text—and evaluate them as such (McCoy et al., 2024). Finally, because these systems are trained on massive datasets, it is not uncommon that some of the test items used to evaluate them were present in the training data and that performance measures are biased as a result. This further highlights the need for systematic and rigorous testing of state-of-the-art LLMs and LRMs (Bowman & Dahl, 2021; Frank, 2023a; Srivastava et al., 2023; Zador et al., 2023; Lewis & Mitchell, 2024; Z. Wu et al., 2024). In this research, we take an approach towards evaluating LLM and LRM capacities inspired by behavior-analytic research on relational abilities. Doing so allows us to avoid getting caught up in philosophical debates about the nature of general intelligence, yet still obtain a measure closely related to it (i.e., do these systems demonstrate a behavior that is believed to be a hallmark of intelligence; Colbert et al., 2018).

Researchers from different fields in psychological research have emphasized the importance of relational abilities as a cornerstone of human cognition (e.g., Hayes et al., 2001; Penn et al., 2008; Halford et al., 2010; Gentner & Smith, 2013; Hughes & Barnes-Holmes, 2015) and intelligence more generally (e.g., Cassidy et al., 2011; Colbert et al., 2018; McLoughlin et al., 2020). Tests of relational abilities are part of many modern intelligence tests (i.e., tasks like matrix reasoning, verbal analogies, etc.; e.g., Raven & Raven, 2003; Kaufman & Lichtenberger, 2006; Wechsler, 2008), and some have been proposed to be a proxy-measure of intelligence (e.g., Colbert et al., 2017; May et al., 2022). Interestingly, different types (or levels of complexity) of relational abilities can be assessed and compared in different types of systems, be that of nonhuman animals, humans or artificially intelligent systems. To see how, we first need to describe the behavior-analytic perspective on relational abilities in more detail.

Relational frame theory (RFT, Hayes et al., 2001) provides a behavior-analytic definition of relational responding—the ability to respond to one event or stimulus in terms of its relation to another. This definition encompasses both responding to formal relations between events (e.g., choose the larger of two stimuli; referred to as non-arbitrarily applicable responding or NAARR) and responding to symbolic relations (referred to as arbitrarily applicable relational responding or AARR). In the latter case, relations can be arbitrarily applied to stimuli (i.e., regardless of formal properties, e.g., selecting a smaller, more valuable €1 coin over the larger, less valuable €0.50 coin) if particular contextual cues are present. Relational contextual cues signal the relationship between the stimuli (sameness, difference, hierarchy, and so on), whereas functional contextual cues indicate what stimulus function is related (e.g., color, value, threat). AARR has three core characteristics: (1) mutual entailment refers to the bidirectionality of relations (if A is related to B, B is also related to A); (2) combinatorial entailment or transitivity (if A is related to B and B is related to C, one can derive the relation between A and C); and (3) the transformation of function (i.e., if a neutral stimulus A is related to a stimulus B with a reinforcement value, the reinforcement value of A will be transformed according to the specified relation and one will respond to A as if it has that function). The definition of relational responding is in line with related constructs studied in cognitive psychology, such as relational reasoning (Alexander et al., 2016) or analogical reasoning (Holyoak, 2012; Gentner & Smith, 2013). We would like to emphasize the fact that the definition encompasses both NAARR (i.e., responding to formal stimulus relations) and

AARR (responding to contextually defined stimulus relations, regardless of formal stimulus relations). Its broad scope allows us to compare different systems (nonhuman animals or other organisms, humans, machines, ...) on a vast array of behaviors, from animals responding to simple, formal sameness and difference relations (e.g., a relational matching-to-sample task or transitive inference task; Penn et al., 2008) to humans or machines solving complex analogies or metaphors (e.g., Webb et al., 2023; Sourati et al., 2024).

Many tests have been developed to measure relational abilities (e.g., Premack, 1983; Goel & Dolan, 2001; Birney et al., 2006; Alexander et al., 2016; Colbert et al., 2017). In this research, we will use a measure developed within the behavior-analytic literature, referred to as the relational abilities index (RAI, Colbert et al., 2017). The procedure is similar to syllogistic reasoning tasks in that it presents the participant (or language model) with a number of relational premises involving nonwords (e.g., “COR is the same as WUG and WUG is different from LOM”) and a (or in some versions, multiple) conclusion(s) to respond to (e.g., “Is COR the same as LOM?”). Syllogistic reasoning has long been considered a hallmark of human intelligence (e.g., by Aristotle) and is studied in various research areas such as behavioral psychology (e.g., McHugh et al., 2004; Colbert et al., 2017); cognitive psychology (e.g., Goodwin & Johnson-Laird, 2005; Todd et al., 2019) and neurology (e.g., Goel & Dolan, 2001), as well as computer science (e.g., Hummel & Holyoak, 2001; Ando et al., 2023; Eisape et al., 2023; Ozeki et al., 2024). Traditionally, syllogistic reasoning tasks make use of realistic stimuli (e.g., names of people, animals, etc.) and involve either comparative relations (e.g., “Ben is smaller than Tom, Tom is bigger than Lisa”, see Goel & Dolan, 2001, for an example) or (logical) reasoning about object categories and features (e.g., category-based induction problems, e.g., “All birds have wings; Penguin is a bird; Therefore penguins have wings”, see Eisape et al., 2023, for an example). The RAI, on the other hand, involves multiple relations (sameness, difference, opposition, more or less than, hierarchical relations, temporal relations, deictic relations and analogies; see Cummins, 2023, for a detailed discussion) and uses arbitrarily chosen stimuli (i.e., nonwords). As such, it aims to capture the generalized (abstract) ability to respond relationally in an arbitrarily applied fashion (i.e., regardless of formal stimulus properties) and may provide a broader survey of relational reasoning abilities than similar measures do. The procedure is well-suited for testing language models’ relational abilities also. The fact that they are natural language reasoning problems makes them readily applicable in LLMs and LRMs and the use of nonwords reduces the likelihood that the models encountered identical problems in training. While the models may still benefit from structurally similar problems that are almost certainly present in training data, syllogistic reasoning problems can be manipulated in many ways to create different variations in the same problems (e.g., increasing the number of premises, including irrelevant premises, manipulating the order of premises, changing the response format to multiple choice) so as to further reduce the likelihood of contaminated results.

For this study, we created a large battery of several thousand syllogistic reasoning problems inspired by the RAI. We assessed a multitude of relations (sameness, difference, opposition, more than, less than, hierarchical relations, temporal relations, deictic relations and analogies). The task was created making use of relational derivation tables (essentially look-up tables that store the reversal and transitive combinations of different combinations of relations allowing for automated derivation of relations; Raemaekers, in preparation¹). It represents a significant increase in size relative to similar tasks, and to our knowledge, this is the largest and most varied RAI-inspired task to this date (previous versions ranged from 55, Colbert et al., 2017, to 67 trials, Colbert et al., 2020, Cummins, 2023). In addition to relations typically tested in the RAI, we also included test trials that specifically assessed transformation of function. RFT considers this to be an important

characteristic of relational responding, but it is currently not included in the RAI. We manipulated problem complexity by varying the number of premises (i.e., one to five relevant premises, with an optional irrelevant premise added to the problem). We also created further problem variations by manipulating whether a valid or invalid conclusion was prompted and by generating multiple instances of each unique problem using different nonwords in the premises, to get a reliable estimate of model performance. We describe the task in more detail below.

2. Materials and Methods

2.1. Syllogistic Reasoning Problems

We constructed a large battery of syllogistic relational reasoning problems ($N = 1364$ unique problems) that included a wide variety of relations and varied in structure and complexity. Similarly to the RAI, for descriptive purposes, we group the trials into blocks based on the relations they involve, the complexity of the relational arrangement, and the requirement to transform stimulus functions: (a) sameness and difference ($N = 152$), (b) sameness and opposition ($N = 488$), (c) more than and less than ($N = 72$), (d) before and after ($N = 72$), (e) hierarchy (i.e., ‘contains’ and ‘is part of’; $N = 72$), (f) deictic (I-you, here-there, now-then; $N = 48$), (g) analogy ($N = 52$) and (h) transformation of function ($N = 408$). The different trial numbers for the different relations were the result of the functioning of the generator function. For a number of specified relations, this function finds all valid combinations (i.e., the combination allows for valid derivation of a new relation) of the relations for a given number of premises (e.g., for two premises with sameness or difference relations, it could be sameness twice, or a one of two combinations of sameness and difference). For those combinations of relations, it then creates relational syllogistic problems with one relation per premise. Relations vary in terms of how they can be combined with other relations to allow valid derivations. For instance, two (or more) opposition relations can be specified between three (or more) relata, allowing one to derive novel transitive relations (e.g., A opposite to B and B opposite to C, so A is the same as C), whereas combinations of two or more difference relations does lead to valid derivations (if A is different from B and B is different from C, one cannot derive the relation between A and C with certainty). As a result, for some combinations of relations (e.g., same and different), fewer unique trials could be generated, leading to different numbers of trials between the relations. Each problem consisted of between one and six relational premises (e.g., ‘A is the same as B. B is the same as C’) and then prompted a conclusion, to which a binary ‘yes’ or ‘no’ response was required. We systematically manipulated problem complexity (varying number of premises and inclusion of an irrelevant premise) and the correct response (prompted syllogism conclusion was valid on half of trials, invalid on the other half). For deictic relations, we also manipulated whether the deictic dimension was to be reversed or not (see McHugh et al., 2004). Ten variants (each with different non-words as the relata) of each unique problem were created, adding up to a total of 13,640 problems (full set available on the Open Science Framework²) on which the LLMs and LRMs were assessed. To reduce the likelihood of prior experience with them in this context, the non-words used as relata were randomly selected from a pool of three-letter syllables, chosen to not be included in openly available versions of the RAI and to as much as possible not be parts of common words. Randomization of the non-words in different problems was constant across the different models. Below, we describe the main variables (beyond the relation involved) that we manipulated in the problems. Example problems are illustrated in Figure 1.

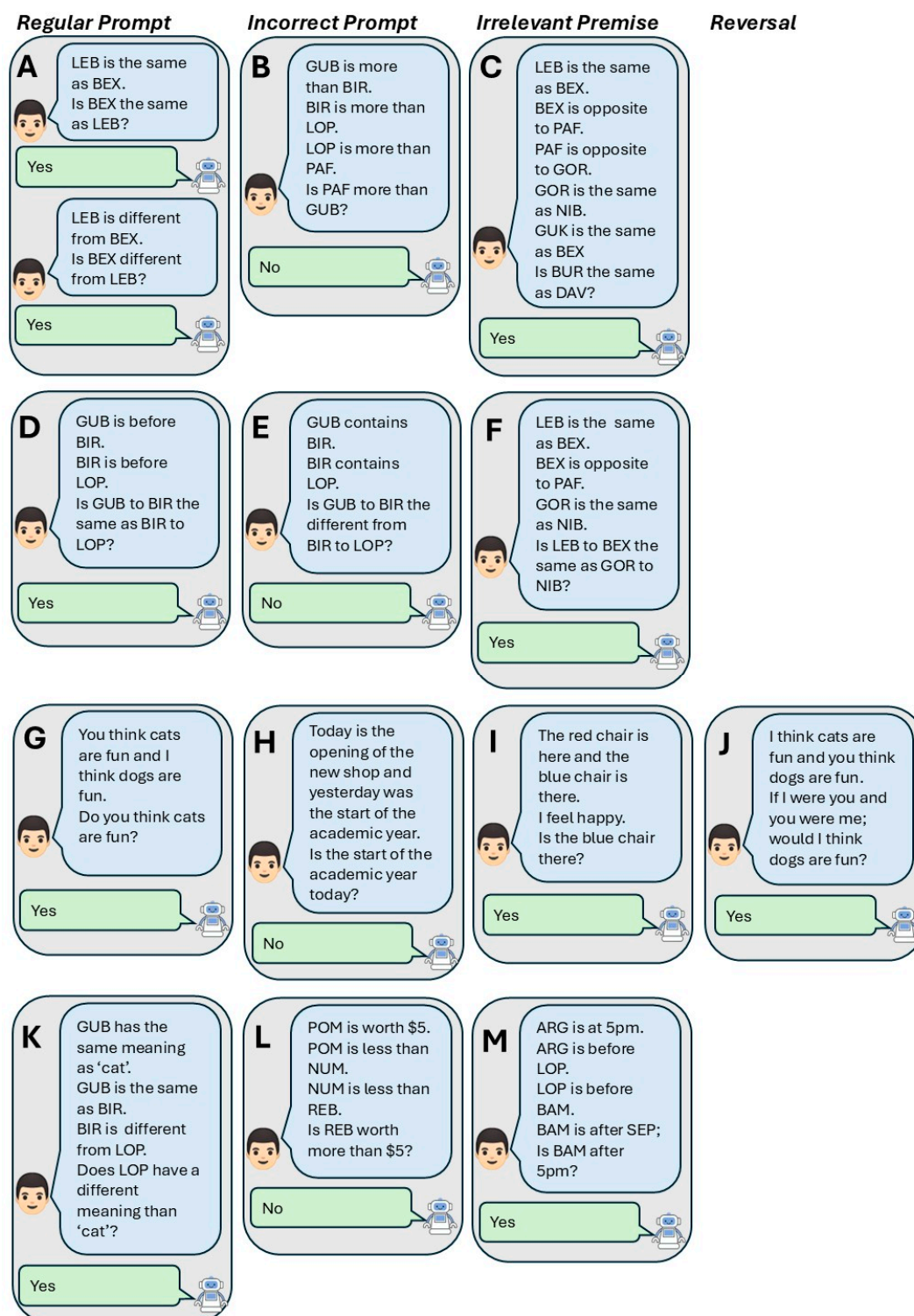


Figure 1. Example trials from the full task set. (A): Example of two one-premise problems with the correct reversal prompted. (B): A two-premise problem involving more than relations, with the incorrect conclusion prompted. (C): A four-premise problem with an irrelevant premise, all involving same–opposite relations. (D): An example of an analogy prompt involving temporal relations. (E): An analogy prompt with the incorrect conclusion prompted. (F): An analogy with an irrelevant premise included. (G–I): Deictic relations problems with the correct relation prompted, the incorrect relation prompted, and an irrelevant premise included, respectively. (J): A deictic relation problem with a reversal of the deictic dimension. (K–M): Examples of problems involving transformations of function, with the correct conclusion prompted, the incorrect conclusion prompted, and an irrelevant premise included, respectively.

2.1.1. Number of Premises and Irrelevant Premises

For all relations except deictic and analogy relations (which always had two premises with an optional irrelevant premise), we manipulated the number of premises in the problems from one to five premises. Problems with only one premise can be thought of as assessing mutual entailment or reversal of a relation (e.g., “A is the same as B, is B the same as A?”, see Figure 1A), whereas problems with two or more premises assessed understanding of transitivity of relations (e.g., “A is opposite to B, B is opposite to C, and C is the same as D. Is D the same as A?”, Figure 1B). For each level of complexity, we also included a variant where an additional, irrelevant premise was added before prompting the conclusion (e.g., Figure 1C,F,I,M). That is, this premise referred to one of the related nonwords, but was not relevant for the prompted derivation (e.g., “A is the same as B, B is opposite of C and C is the same as D. Is A the opposite of C?”).

2.1.2. Conclusion Validity

We also manipulated the validity of the prompted conclusion. That is, half of the problems had an invalid conclusion prompted (e.g., “A is the same as B. Is B different from A?”, e.g., Figure 1B), while the other half had a valid conclusion prompted (e.g., “A is the opposite of B and B is the opposite of C. Is C the same as A?”; e.g., Figure 1D). This manipulation allowed us to investigate whether the models show signs of affirmation bias (i.e., better performance when answering yes). Such biases could, for example, arise due to problems with correct conclusions being more prevalent in training than those with invalid conclusions, or because of the extensive finetuning process to tailor the models’ responses to human preferences.

2.1.3. Deictic Relations

The problems involving deictic relations had a slightly different structure to those involving other relations (see Figure 1G–J). RFT describes three types of deictic relations (McHugh et al., 2004): interpersonal relations (I–you), temporal relations (now–then) and spatial relations (here–there)). RFT-inspired tests of this ability (e.g., McHugh et al., 2004; Cummins et al., 2023) combine these relations (e.g., I am here now, you were there then. Am I here now?). For the current test battery, we separated the three relations. One set of trials assessed spatial relations (‘A is here. B is there. Is B here?’), another assessed purely temporal relations (e.g., ‘A is now, B is tomorrow. Is B tomorrow?’) and a final set assessed interpersonal relations (e.g., ‘I think A, you think B. Do I think A?’). Instead of different nonwords, different objects, events and thoughts were included in the various iterations of each problem. Similarly to the other relations, we created trial variations where an irrelevant premise was included (Figure 1I), and where the prompted conclusion was invalid (as well as a combination of both). We also included a variant that prompted a reversal of the deictic dimension (e.g., ‘If I were you and you were me, would I ...?’ or ‘If here were there and there were here, would A be here?’; Figure 1J).

2.1.4. Analogy

For all relations except deictic relations, we included problems that required analogical reasoning (see Figure 1D–F). These problems provided two premises (e.g., “A is opposite to B, B is opposite to C.”) and then prompted a comparison of the relation in the first premise to that in the second premise (Figure 1D). Here too, a version with an incorrect analogy prompt (e.g., “A same as B and B same as C. Is A to B different from B to C?”, Figure 1E), and a version with a third, irrelevant premise (i.e., not involved in the analogy prompt, Figure 1F) were included.

2.1.5. Transformation of Function

To assess whether LLMs and LRMs demonstrate the ability to transform psychological functions, we also created a set of problems that applied relations to semantic functions of nonwords (see Figure 1K–M). In these problems, a non-word is assigned a meaning, then related to other nonwords in the same relational premises, and then the meaning of one of the related nonwords was queried (e.g., “AGU means ‘yes’. AGU has the same meaning as BUR. BUR has the opposite meaning of LOP. Does LOP have the same meaning as ‘yes’?”, Figure 1K). For sameness, difference and opposition relations, the nonwords were related to existing words (e.g., “ARU has the same meaning as ‘cat’.”), whereas for more than, less than, before and after relations, nonwords were assigned a value (e.g., “ARU is worth €5”) or time (e.g., “ARU is at 4 p.m.”). Variations in these problems again included manipulating the validity of the presented conclusion (Figure 1L) and adding irrelevant premises to the problem (Figure 1M). While this is a limited form of transformation of function, we do consider it to be a transformation of semantic function via sameness, difference, opposition, comparative and temporal relations. In that sense, we go beyond what existing versions of the RAI measure and at least have some measure of a crucial aspect of relational responding abilities.

2.1.6. Replication with Randomized Order of Premises

As a final test of the generality of the models’ ability to solve relational syllogistic problems, we replicated the full task battery as described above (i.e., 13,640 problems) but randomized the order of premises in each problem. That is, for each problem consisting of two or more premises (including irrelevant premises and premises specifying a function), the premises were no longer presented in a linear fashion (i.e., “A related to B, B related to C, C related to ...”), but in a randomized order (e.g., A related to B, C related to D, B related to C). The (randomized) premises were still followed by a prompted conclusion requiring a yes or no response from the models. While randomizing the order of premises might also increase the complexity of the problems for humans, in principle, the relational complexity (i.e., the derivations to be made) remains the same. Therefore, reduced performance in this replication would hint at a lack of generalized ability for relational responding.

2.2. Models and Inference

We tested a sample of two state-of-the-art LLMs and two LRMs. This sample was chosen to include LLMs and LRMs varying in size, architecture and training. Although our sample was too limited to make general claims about LLMs, LRMs, or the difference in performance between the two, including a variety of models did allow us to assess relational behavior in both type of systems. Given the seminal nature of our study, we restricted the sample because of financial, environmental and practical concerns³. All models were accessed using the Together AI API service⁴. We describe them briefly below (summary in Table 1). For more technical descriptions, see the referenced publications.

We tested two models from the Llama 3 collection of LLMs developed by Meta AI (Dubey et al., 2024). Llama 3.1 405B, which has 405 billion parameters and a decoder-only transformer architecture, and Llama 3.3 70B, which has 70B parameters and has received further training (reinforcement learning, RL, from human feedback) to fine-tune it to human preferences. We also tested two recent models of the GPT-model family (Open AI, 2025): GPT OSS 20B and GPT OSS 120B. These are so-called open-weight models that are trained using reinforcement learning from human feedback and other techniques derived from past Open AI models, on only text data (Open AI, 2025). Both have a mixture-of-experts architecture (MoE, Jiang et al., 2024). GPT OSS 20B has 21 billion parameters with 3.6 billion active parameters, GPT OSS 120B has 117 billion parameters with 5.1 billion

active parameters. Based on the results of a pilot optimization study (see Appendix A), we chose to test the models in a zero-shot setting (i.e., no examples, information or additional prompts provided in the context; a system prompt was provided: “Answer with yes or no only, no punctuation is needed. Don’t put a space before your response.”), and with the temperature parameter set to 0.75. For all other parameters, the default settings were used (see Together AI documentation).

Table 1. Overview of the tested models (and group), their training and parameter size.

Model (Group)	Training	Parameters
LLaMa 3.1 405B IT (LLM)	Pre: 15T+ multilingual, open-source text tokens. Post: Supervised Fine-Tuning, Rejection Sampling, and Direct Preference Optimization ¹ .	405B
LLaMa 3.3 70B IT (LLM)	Pre: 15T+ multilingual, open-source text tokens. Post: Supervised fine-tuning and reinforcement learning with human feedback ¹ .	70B
GPT OSS 20B (LRM)	Pre: trillions of text tokens, focus on STEM, coding, and general knowledge. Post: supervised fine-tuning and high-compute RL ² .	21B, 3.6 active
GPT OSS 120B (LRM)	Pre: trillions of text tokens, focus on STEM, coding, and general knowledge. Post: supervised fine-tuning and high-compute RL ² .	117B, 5.1 active

Note: B = billion, T = trillion. RL = Reinforcement Learning. Model names are those used in the referenced publications. ¹ Meta AI (2024). ² Open AI (2025).

3. Results

3.1. Overall Task Performance

Figure 2 (and Table 2) shows the models’ overall performance by blocks (i.e., relations, including analogy and transformation of function separately). As per the preregistered analytic strategy, we limit ourselves to descriptive statistics here to keep the results digestible. Binomial Generalized Linear Model analyses are summarized in Appendix C. Our results demonstrated that the LRMs (GPT OSS 20B and GPT OSS 120B) mostly outperform the LLMs (LlaMa 3.1 405B and Llama 3.3 70B; except for deictic relations where LlaMa 3.1 405B outperformed the other models slightly). The larger models (more parameters) also tended to outperform their smaller counterparts. While the models generally reached a high level of performance (GPT models correctly solved 83% or more of the problems and reached maximum or near-maximum scores in most blocks), there was quite some variability in performance across the different blocks (more so in LlaMa models). Performance dropped notably in the analogy block, even for the GPT models (GPT OSS 20B: 83.27%; GPT OSS 120B: 83.85%), but more so for the LLMs (LlaMa 3.1 405B: 55.19%; LlaMa 3.3 70B: 55.31%). For the LLMs, performance also dropped significantly in the same–opposite block (LlaMa 3.1 405B: 61.25%; LlaMa 3.3 70B: 55.18%) and the transformation of function block (LlaMa 3.1 405B: 82.06%; LlaMa 3.3 70B: 69.04%), and to a lesser extent in the same–different block (LlaMa 3.1 405B: 83.16%; LlaMa 3.3 70B: 70.59%).

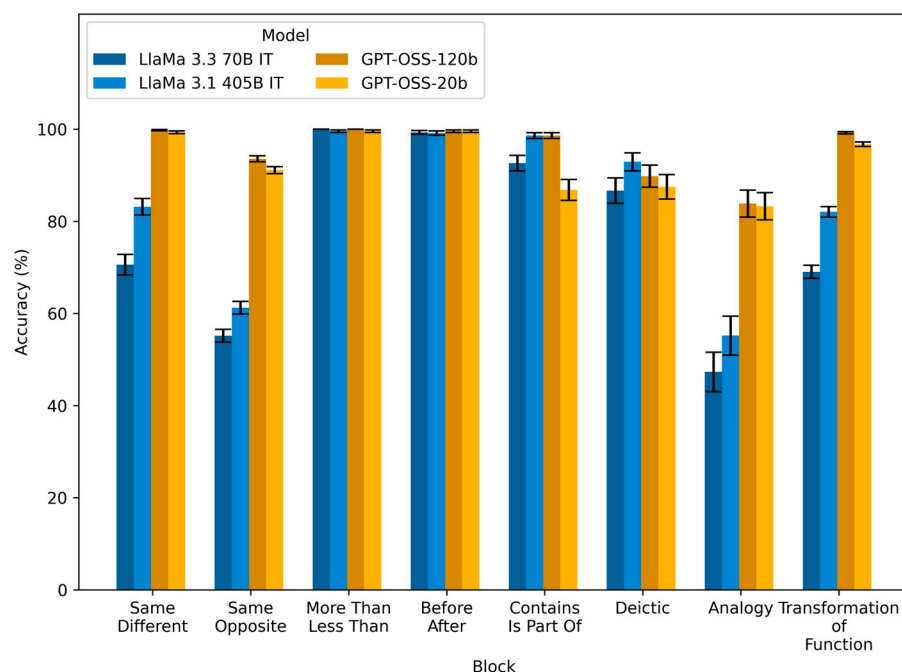


Figure 2. Model performance (percent accurate) grouped by the relations involved in the problems. Error bars are 95% Wilson binomial confidence intervals across items within each block.

Table 2. Model Accuracy (%) Across Blocks.

Block/Model	LLaMa 3.3 70B	LLaMa 3.3 405B	GPT OSS 20B	GPT OSS 120B
Same and Different	70.59	83.16	99.32	99.80
Same and Opposite	55.18	61.25	91.11	93.56
More Than and Less Than	100	99.58	99.58	100
Before and After	99.31	99.17	99.58	99.58
Contains and Is Part Of	92.64	98.61	86.81	98.61
Deictic	86.67	92.17	87.50	89.79
Analogy	47.31	55.19	83.27	83.85
Transformation Function	69.04	82.06	96.74	99.24

3.2. Effect of Problem Complexity

For all blocks except the deictic and analogy blocks, we varied the number of relational premises in each problem from one to five. For comparative (more than, less than) and temporal relations (before, after), all models were remarkably robust to this manipulation and performed at near-maximum level for all levels of complexity. For the other relations, however, we did observe that performance dropped increasingly with more premises. In the same–different and same–opposite blocks, this drop in performance was more pronounced for the LLaMa models (which dropped from maximum down to chance-level) than for the GPT models (which dropped by about ten percentage points). For hierarchical relations, we observed a reversed pattern, where the models’ performance dropped by up to 20% for problems with three premises, but then recovered slightly for more complex problems. These results are illustrated in Appendix B.1 Figure A2 and described in Table A5.

3.3. Effect of Prompt Validity

We also manipulated whether a valid or invalid conclusion was prompted and whether an irrelevant premise was included, to test the models’ sensitivity to variations

in problem presentation. While the effect of these variations was more difficult to interpret, we can say that GPT models were less affected by them than were the LLaMa models (see Appendix B.2 Figure A3 and Table A6 for detailed results). For the same–different, more than–less than and before–after blocks, all models performed consistently (at a high level) across the problem variations. For same–opposite relations, the GPT models were relatively unperturbed by the problem variations and were even more accurate when an incorrect conclusion was prompted (10–15% more accurate). The LLaMa models showed the reversed patterns, being significantly less accurate when an irrelevant premise was included in the problem and with performance dropping even more when an invalid conclusion was prompted. Finally, for hierarchical relations, both the GPT and LLaMa models' performance dropped slightly (between ten and thirty percent) when an invalid conclusion was prompted, but the addition of an irrelevant premise did not noticeably affect performance. We discuss the effect of problem variations for deictic relations, analogy and transformation of function separately below.

3.4. Deictic Relations Performance

We describe the results for deictic relations separately, as these problems were slightly different from the other relations. Trials involved three types of deictic relations: interpersonal, temporal and spatial. All models were slightly less accurate on problems involving interpersonal relations than on trials involving spatial and temporal relations, on which they reached near-maximum performance (see Appendix B.2 Figure A3 and Table A7). In addition to manipulating the inclusion of an irrelevant premise and the validity of the prompted conclusion, we also included trials where the deictic dimension was reversed. While the models were only modestly affected by prompting an incorrect conclusion, performance dropped when a reversal was included, and more so when this was combined with an invalid conclusion or irrelevant premise. Most models were affected by trial variations in similar ways, but it is noteworthy that LLaMa 405B appeared to be significantly more robust to these manipulations, with its performance remaining largely stable or even increasing (relative to the regular prompt), and likely as a result outperforming the other models on deictic responding with 92.17% accuracy).

3.5. Analogy Performance

For all relations except deictic relations, we probed analogical reasoning abilities by providing two relational premises and asking whether the relations in the two premises were the same or not. Here too, we included variants with an invalid conclusion and with irrelevant premises. Across problem variants, models were more accurate solving analogies involving sameness, difference and opposition relations than those involving comparative (more than, less than), temporal (before, after) or hierarchical relations (see Appendix B.3 Figure A4 and Table A8). Still, in the same–different and same–opposite blocks, LLaMa 3.1 405B was significantly less accurate on problems where an incorrect conclusion was prompted (performance dropped from around eighty to below fifty percent). The smaller LLaMa 3.3 70B model only showed this sensitivity in the same–opposite block, while the GPT-models were relatively unaffected by problem variations in these blocks. In the blocks involving comparative, temporal and hierarchical relations, performance was markedly lower. The LLaMa models scored below 25% accuracy, while the GPT models reached about 50% accuracy. Performance on problems involving temporal and hierarchical relations was slightly higher, with LLaMa models reaching about 30% accuracy and GPT models reaching up to 75% accuracy. Interestingly, in the latter three blocks, performance appeared to increase when an invalid conclusion was prompted, or when an irrelevant premise was included.

3.6. Transformation of Function Performance

For sameness, difference, opposition, comparison and temporal relations, we also included trials that probed the models' ability to transform stimulus functions (nonword meaning). These problems again included variants with invalid conclusions and irrelevant premises. The GPT models were both highly accurate for all relations (near-maximum performance, GPT OSS 20B 96.74% on average, GPT OSS 120B 99.24% on average). The LLaMa models were slightly less accurate (LLaMa 3.3 70B was 69.04% accurate on average, LLaMa 3.1 405B was 82.06% accurate on average), but this drop in performance was limited to same–different and same–opposite relations. For those relations, both LLaMa models' performance decreased with an increasing number of premises and with invalid conclusions prompted (see Appendix B.4 Figure A5 and Tables A9 and A10).

3.7. Replication with Randomized Order of Relational Premises

As a final test of the robustness of our results, we conducted a replication study in which we recreated the full task battery, but with the order of relational premises randomized for each problem. The results of this replication are illustrated in Figure 3 and summarized in Table 3. As can be seen from comparing Figures 1 and 3, the models generally performed at about the same level in the replication study as in the original task. For some models, in particular blocks, performance dropped significantly, while for others, it even improved slightly. The most notable drop in performance was for GPT OSS 120B. While still performing at a high level, it lost around five percentage points in the same–different (91.84%), same–opposite (88.81%), comparison (94.03%), temporal (94.31%), hierarchy (92.92%) and transformation of function blocks (92.48%), in which it performed at near-maximum level in the original task (see Table 2). The smaller GPT OSS 20B model was more robust to the randomization of premise order. Also notable, the LLaMa 3.1 405B model dropped about ten percentage points in the transformation of function block (72.70%), but gained about seven percentage points in the same–different block (89.34%). Similarly, the smaller LLaMa 3.3 70B model also dropped about five percentage points in the transformation of function block (65.96%) and gained about eight percentage points in the same–different block (78.42%). The effect of increasing the number of relational premises in a problem, including irrelevant premises and manipulating the validity of the prompted conclusions were largely similar to the original results. The effect of increasing number of premises was slightly more pronounced in the replication study for GPT models (Figure A7 and Table A27). Effects of problem variations in deictic (Figure A8 and Table A28), analogy (Figure A9 and Table A29) and transformation of function blocks (Figure A10, Tables A30 and A31) were largely the same as in the original task.

Table 3. Model Accuracy (%) Across Blocks in the Replication with Randomized Premise Order.

Block/Model	LLaMa 3.3 70B	LLaMa 3.3 405B	GPT OSS 20B	GPT OSS 120B
Same and Different	78.42	89.34	99.41	91.84
Same and Opposite	55.37	58.14	89.45	88.81
More Than and Less Than	99.17	99.72	99.17	94.03
Before and After	98.06	99.03	99.58	94.31
Contains and Is Part Of	91.39	96.53	86.81	92.92
Deictic	84.58	94.17	89.58	89.58
Analogy	48.27	55.39	77.50	79.81
Transformation Function	65.96	72.70	92.99	92.48

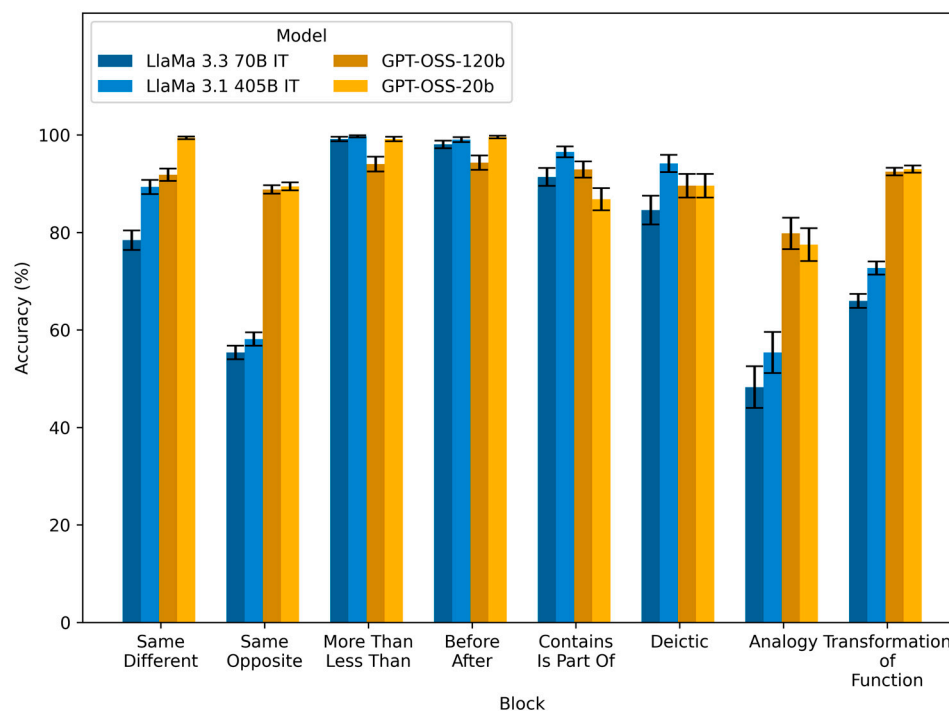


Figure 3. Model performance (percent accurate) in the replication study with randomized premise order, grouped by the relations involved in the problems. Error bars are 95% Wilson binomial confidence intervals across items within each block.

4. Discussion

Inspired by behavior-analytic research on the human ability for relational responding, we conducted survey of a small sample of LLMs and LRMs on a large battery of relational syllogistic reasoning problems. Our results demonstrated that both the LLMs (i.e., LLaMa 3.1 405B and 3.3 70B) and LRMs (i.e., GPT OSS 20B and 120B) in our sample generally perform well in these tasks. The LRMs reach at least eighty percent accuracy on all types of relating (including analogy and transformation of function) and reach near-maximum performance on same–different, comparison and temporal relations. The LLMs showed more variability in their performance, and while they performed at a similar level to the GPT models in the comparison, temporal, hierarchy and deictic blocks, their performance dropped markedly in the same–different, same–opposite, analogy and transformation of function blocks. This finding appears to show that the extensive post-training that LRMs go through does benefit them when it comes to solving syllogistic reasoning problems, potentially because this training involves finetuning of step-by-step reasoning (e.g., Rafailov et al., 2023; Li et al., 2023; Ouyang et al., 2022; Guo et al., 2025). Also in line with past research on LLMs and LRMs (e.g., Kaplan et al., 2020; Brown et al., 2020; Chowdhery et al., 2023; Guo et al., 2025), we observed that the larger models (i.e., LLaMa 3.1 405B and GPT OSS 120B) generally outperformed their smaller counterparts (i.e., LLaMa 3.3 70B and GPT OSS 20B, respectively).

These results appear to show that state-of-the-art systems have acquired an ability for relational responding, but they do show sensitivities to the type(s) of relations involved in the problem premises, and to variations in the problem presentation (i.e., varying the number of premises, including irrelevant premises and prompting invalid conclusions). While the effects of task variations differed between the models and between types of relations, they may reflect artifacts from the models' training. The models' decreased performance when invalid conclusions were prompted (except for GPT models, which were more accurate for invalid conclusions prompted in the same–opposite block)

may reflect a bias for affirmative responding. Similarly, decreased performance in problems with more premises and for problems with an irrelevant premise could also reflect a lack of generalization from training. While our use of non-words likely reduces the probability that the models have seen these exact problems in training, they could still take advantage of structural overlap between these problems and those seen in training, and from the presence of cues indicating that overlap (i.e., performance is a function of overlap in problem topography, rather than generalized relational understanding). We were not able to control the training data of the models, but future research could address this. The decrease in performance on these problem variants, however, could also simply reflect the increased difficulty of these problems (i.e., problems with more relational premises have higher relational complexity, Halford et al., 2010), differences in prompt sensitivity or tokenization idiosyncrasies, which could also be elucidated in future research. Results of our replication study showed that the models' performance was relatively robust against perturbations of the order of relational premises (which affected performance in prior studies, e.g., W. Wu & Deng, 2025), suggesting the ability to solve relational syllogisms is relatively generalized. Future research could investigate the models' sensitivity to problem variations by other aspects of the task (e.g., multiple choice format, testing other relations) and by assessing human's sensitivity to those same task variations.

The LLMs' relatively low performance on analogy problems (LLaMa models performed around chance-level, while GPT models reached about 83% accuracy) somewhat nuances claims of emergent analogical reasoning in LLMs e.g., Webb et al., 2023; but see Lewis & Mitchell, 2024, for a counterexample). Taken together with the fact that we probed analogical reasoning in a somewhat unusual way (e.g., "A is the same as B. B is the same as C. Is A to B the same as B to C?"), one could argue that the models' performance does not reflect truly generalized, abstract relational understanding. However, we must also note that the decrease in performance for analogy problems was most pronounced for comparative and temporal relations, where technically speaking, the model can be considered correct when it determines two identical relations (e.g., "A more than B" and "B more than C") as not being identical, because they might not be the exact same instance of the relation (e.g., A may be five more than B, while B could be twenty more than C). Further research investigating the models' reasoning process is needed to address this question.

A final aspect of our test of relational abilities that deserves highlighting is the transformation of function. The GPT models appeared to have little difficulty with these problems and performed at near-maximum level. LLMs' lower performance on transformation of function trials was limited to same-different and same-opposite relations. Performance on those problems decreased with an increasing number of premises, analogous to regular same-different and same-opposite trials (no transformation of function) and may therefore reflect the effect of increasing the number of same-different and same-opposite premises, rather than a difficulty with the transformation of stimulus function itself. Both LLMs and LRMs thus appear to be capable of transforming semantic functions in a relational network, which is an important aspect of relational responding (Hayes et al., 2001).

Further research is needed to answer the question how the models' performance relates to that of humans. Given the limited number of studies using the RAI or its derivatives (and the different versions used therein), normative (block-level) data for human performance is unavailable. Colbert et al. (2020) validated a version of the RAI (69 trials involving sameness, difference, opposition, comparative, temporal and analogy) and reported that adult participants scored between 80% and 90% for same-different, same-opposite, more-less and before-after trials, and that performance dropped to

around 65% for analogy trials. Future research needs to collect human data for this test battery to allow for a more direct comparison of human and artificial performance, as well as the effect of problem variations on human performance.

Readers should be careful interpreting these findings, however. As we mentioned in the introduction, there are many pitfalls, such as the risk of anthropomorphizing the models and attributing human abilities and mechanisms (e.g., understanding, reasoning) to them (Lin, 2025). It is important to emphasize that output alignment (the fact that artificial systems produce seemingly human-like responses) does not a priori imply alignment of the underlying mechanisms or processes. Indeed, these models may solve these tasks in entirely different ways than humans do. The current study only investigated overall performance (whether models can solve relational syllogistic problems), but to further investigate similarities between the models' and human reasoning, future research should study the models' 'reasoning' process in more detail, and compare it with that of humans (e.g., Sourati et al., 2024; Ozeki et al., 2024; Eisape et al., 2023; Bertolazzi et al., 2024). Another limitation of the current work concerns our use of non-words in our test. We chose to do so in line with previous work in behavior-analytic literature (e.g., Colbert et al., 2020), and argue that the use of non-words allows us to assess whether the system (human or artificial) has acquired a general ability for relational responding (i.e., if it can act as if randomly selected non-words are related, it can learn to do so with any stimulus). A downside of using non-words is that it reduces the ecological validity of our task. While we do not wish to downplay this concern, it is reassuring to know that past research has shown that performance in procedures like these (i.e., syllogistic reasoning problems that specify relations between non-words) can be considered a proxy-measure of intellectual abilities (Colbert et al., 2017) and that they can be used to improve relational and intellectual abilities (e.g., Cassidy et al., 2011; Dixon et al., 2022). Furthermore, we would argue that our use of non-words is intended to both (i) reduce training data contamination, and (ii) increase reliance on relational contextual cues (i.e., the relations specified in the relational premises), because the non-words are unlikely to possess functions that might control the relational responding. There are therefore theoretical reasons (see Hayes et al., 2001) to suspect that the use of non-words may not impact ecological validity when the same cues are present in the tested syllogisms and the "real world". Nevertheless, future research could address the question of ecological validity of these procedures in more detail.

Beyond merely assessing the relational abilities of LLMs and LRMs to get another measure of their competence, we mainly conducted this study to illustrate a broader point about using (a behavior-analytic perspective on) relational responding as a framework to evaluate artificial systems and compare them to other systems (humans or other animals). As we mentioned in the introduction, researchers in different fields of psychological research and computer science agree that relational abilities are a cornerstone of human cognition (McLoughlin et al., 2020) and may even be considered a proxy-measure of general intelligence (Colbert et al., 2017; Raven & Raven, 2003). By studying relational abilities, we can evaluate different systems without getting caught up in philosophical debates about the nature of general intelligence, yet still obtain a measure related to it. Our results showed that state-of-the-art LLMs and LRMs display the ability for relational reasoning, which is a core aspect of intelligence. Furthermore, taking RFT's definition of relational responding (Hayes et al., 2001), encompassing both relatively simple responding to formal stimulus relations and more complex AARR (symbolic behavior), we can assess different types or levels of relational responding across a wide range of systems (nonhuman animals, humans, artificial systems). Many animals appear to be capable of responding to formal stimulus relations (e.g., Penn et al., 2008), while humans were until recently assumed to be unique in their ability for AARR (e.g., Hayes et al., 2001;

Lionello-DeNolf, 2009; Lionello-DeNolf, 2021). Our results show that language models have developed the ability to AARR, at least as far as our test is concerned. Further research in this direction would allow us to make fair comparisons of these systems' relational abilities and possibly provide clues about the bigger questions surrounding the level of intelligence of current state-of-the-art artificial intelligence.

Finally, RFT also emphasizes transformation of psychological function as a crucial aspect of AARR. We are not aware of any research that has directly assessed whether LLMs and LRMs have this ability. Our results provide evidence that they do. While our current procedure only probed a limited form of function transformation (i.e., transforming the meaning of nonwords), it does go beyond existing measures (e.g., Colbert et al., 2020; Cummins et al., 2023) that currently do not test it at all. Future research could investigate other types of transformations of function (e.g., transforming response functions: "When I say GUK, you respond with 'yes'. GUK is opposite to TOK. I say TOK! ['no']") or probe it in different ways (e.g., procedures like the recently developed function transformation tasks: Finn & De Houwer, 2021; Finn et al., 2023). However, there is also the broader, more philosophical question of whether artificial systems like LLMs and LRMs can develop this ability in the first place. According to RFT, the ability to AARR (and thus, the ability for transformation of function) develops through a long history of reinforcement in a verbal environment and involves abstraction of generalized patterns of relating away from formal stimulus properties involved in multiple exemplars of relations. Humans experience this learning history grounded in a multimodal sensory environment. It has been argued that experiencing many different examples of relational responding in a long and structured learning history is required for the ability to AARR (and thus, transform of function) to arise. LLMs strictly speaking only experience our world in a linguistic sense (except for more recent multimodal LLMs and LRMs, e.g., GPT o1, Open AI et al., 2024; Gemma 3, Gemma Team et al., 2025) and require orders of magnitude more training than the average human which looks much different from that of humans (e.g., Frank, 2023b). One could therefore argue that they cannot develop the ability for generalized relational understanding altogether (e.g., Bender et al., 2021). We hope that the framework we propose here can foster progress in these debates.

5. Conclusions

We conducted a large-scale assessment of the relational abilities of state-of-the-art LLMs and LRMs. Our results demonstrated that these systems generally perform well in our relational syllogistic reasoning task and appear to have developed the ability for relational responding, a core characteristic of intelligence. We did observe significant differences between the models and differences in performance for different relations. Furthermore, results suggest that model size and training significantly affect performance (bigger, more intensively post-trained models perform better). Finally, sensitivities to task variations hint at the possibility that performance does not fully reflect generalized understanding, so caution is warranted when interpreting these findings.

Author Contributions: Conceptualization, M.R., M.F. and J.D.H.; methodology, M.R., M.F. and J.D.H.; software, M.R.; validation, M.R.; formal analysis, M.R.; investigation, M.R.; resources, M.R.; data curation, M.R.; writing—original draft preparation, M.R.; writing—review and editing, M.R., M.F. and J.D.H.; visualization, M.R.; supervision, M.F. and J.D.H.; project administration, M.R.; funding acquisition, M.R. and J.D.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Flemish Research Foundation grant number 11M0325N (M.R.), Flemish Research Foundation grant number 12A2B26N (M.F.) and Special Research Fund grant BOF22/MET_V/002-01M00209 (J.D.H.).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available on the OSF at https://osf.io/78u36/overview?view_only=8f2df70d8ff845e9ad393d407f4c27c1 (accessed on 29 October 2025).

Acknowledgments: During the preparation of this study, the authors used ChatGPT (version 5, Thinking) for the purposes of code revision. The authors have reviewed and edited the output and take full responsibility for the content of this publication. GenAI was not used in the writing process.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LLM	Large Language Model
LRM	Large Reasoning Model
RFT	Relational Frame Theory
RAI	Relational Abilities Index
OSF	Open Science Framework

Appendix A

Pilot Optimization Study

The default hyperparameter settings were used for all models, except for the temperature. To determine the optimal temperature settings, we piloted a subset of the test battery to two of the models, GPT OSS 120B and LLaMa 405B IT. We tested four temperature parameter settings, zero, 0.25, 0.5 and 0.75, and chose the one setting that produced the highest performance, or any comparison deemed interesting. Model performance will be compared in a zero-shot reasoning, few-shot (in-context) reasoning, and a ‘chain-of-thought’ reasoning setting. In the zero-shot reasoning setting, the model only receives a system prompt about the task and response format (i.e., “Answer with yes or no only, no capitalization or punctuation needed.”) and is then prompted with the syllogism and a conclusion. For the chain-of-thought (CoT) prompt variant, the model is prompted to reason step by step after being presented with the syllogism (i.e., “Try and think this through step-by-step.”). In the few-shot learning setting, the model is first presented with two example syllogisms with the correct answer (one single premise reversal with an incorrect conclusion and one three-premise problem with a correct conclusion), before being prompted with the syllogism and possible conclusion(s) (again, with the additional prompt to reason step-by-step in the chain-of-thought reasoning setting).

The pilot study results showed overall high performance for both models, with GPT OSS 120B reliably outperforming LLaMa 3.1 405B IT by about 3 percentage points. With respect to the temperature parameter, our results indicated no significant differences between the four settings we tested. With respect to prompting techniques, our results indicated that few-shot prompting (i.e., providing a few example problems with correct responses before prompting the test problem) caused performance to increase by about 4.5 percentage points for both models. Based on these results, we decided to set the temperature parameter for testing models on the full battery to 0.75 (as Together AI

recommends). We decided to only test the models in a zero-shot setting (i.e., no examples or additional prompt), seen as performance in that condition was already high, and few-shot prompting only improved it by a couple of percentage points.

Table A1. Pilot Study Results. Model Accuracy (%) at Different Temperature Parameter Settings.

Model\Temperature	0	0.25	0.5	0.75
GPT OSS 120B	95.1	95.9	96.2	97.0
LlaMa 3.1 405B IT	92.3	92.3	92.9	92.6

Table A2. Binomial GLM Results: Model $\times$ Temperature.

	Coeff.	Std. Err.	z	p
Intercept	2.49	0.20	12.634	0
Model (GPT)	0.47	0.25	1.88	0.06
Temperature 0.25	0	0.06	0	1
Temperature 0.5	0.08	0.08	1.00	0.32
Temperature 0.75	0.04	0.07	0.58	0.57
Model-GPT $\times$ Temperature—0.25	0.19	0.26	0.75	0.45
Model-GPT $\times$ Temperature—0.5	0.18	0.27	0.69	0.49
Model-GPT $\times$ Temperature—0.75	0.47	0.28	1.69	0.09

Table A3. Pilot Study Results. Model Accuracy (%) for Different Prompting Techniques.

Model\Prompt	Zero-Shot	Few-Shot	Chain-of-Thought
GPT OSS 120B	93.9	98.6	95.7
LlaMa 3.1 405B IT	90.8	95.3	91.6

Table A4. Binomial GLM Results: Model $\times$ Prompt Technique.

	Coeff.	Std. Err.	z	p
Intercept	2.39	0.16	14.61	0
Model (GPT)	0.71	0.23	3.05	0.002
Prompt-Few-shot	0.62	0.22	2.82	0.005
Prompt-CoT	−0.10	0.10	−1	0.318
Model-GPT $\times$ Prompt-Few-shot	0.51	0.55	0.93	0.350
Model-GPT $\times$ Prompt-CoT	−0.27	0.23	−1.20	0.229

Appendix B
Appendix B.1. Effect of Problem Complexity and Variations

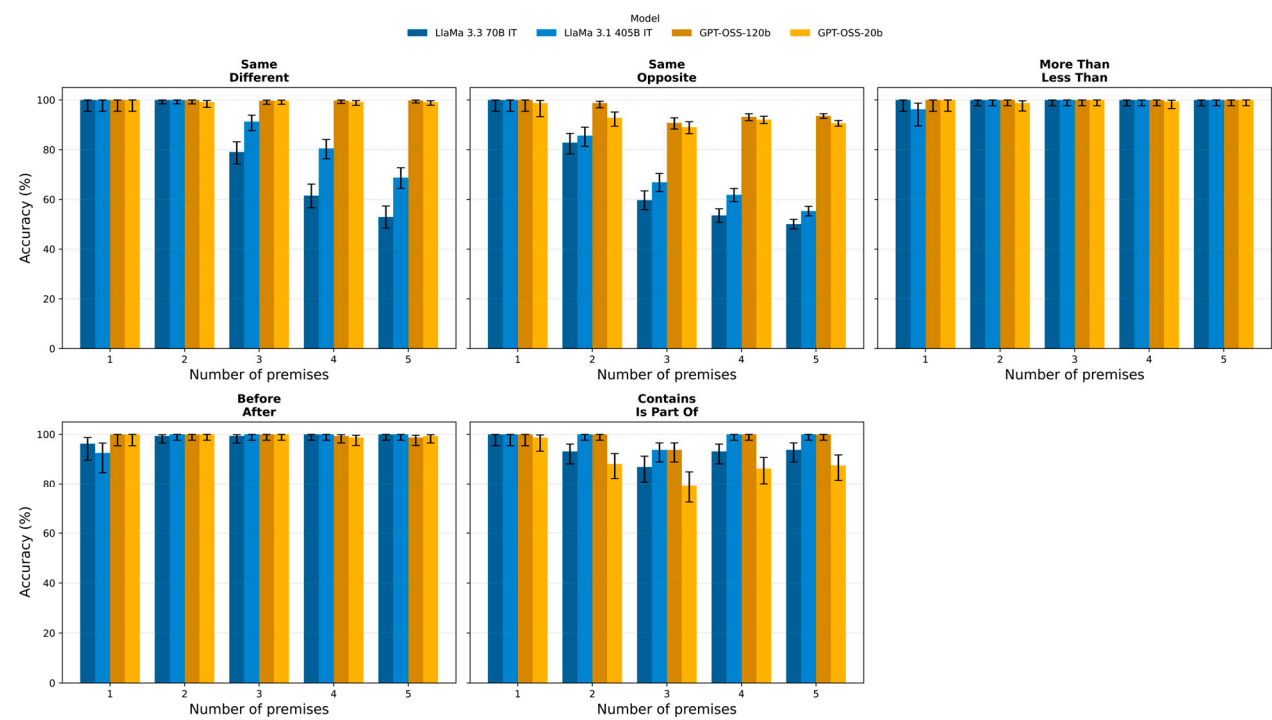


Figure A1. Performance for increasing number of relational premises grouped by relations involved. Error bars are 95% Wilson binomial confidence intervals.

Table A5. Model Accuracy (%) For Increasing Number of Premises Grouped by Relations Involved.

Block	Number of Premises	LLaMa 3.3 70B	LLaMa 3.3 405B	GPT OSS 20B	GPT OSS 120B
Same and Different	One	100.00	100.00	100.00	100.00
	Two	100.00	100.00	99.17	100.00
	Three	79.06	91.25	99.69	99.69
	Four	61.50	80.50	99.25	99.75
	Five	52.92	68.75	99.17	99.79
Same and Opposite	One	100.00	100.00	98.75	100.00
	Two	82.81	85.625	92.81	98.75
	Three	59.69	66.88	89.06	90.78
	Four	53.52	61.80	92.11	93.20
	Five	50.04	55.31	99.66	93.59
More Than and Less Than	One	100.00	96.25	100.00	100.00
	Two	100.00	100.00	98.75	100.00
	Three	100.00	100.00	100.00	100.00
	Four	100.00	100.00	99.38	100.00
	Five	100.00	100.00	100.00	100.00
Before and After	One	96.25	92.50	100.00	100.00
	Two	99.38	100.00	100.00	100.00
	Three	99.38	100.00	100.00	100.00
	Four	100.00	100.00	98.75	93.75
	Five	100.00	100.00	99.38	98.75
Contains and Is Part Of	One	100.00	100.00	98.75	100.00
	Two	93.13	100.00	88.13	100.00
	Three	86.38	93.75	79.38	93.75
	Four	93.13	100.00	86.25	100.00
	Five	93.75	100.00	87.50	100.00

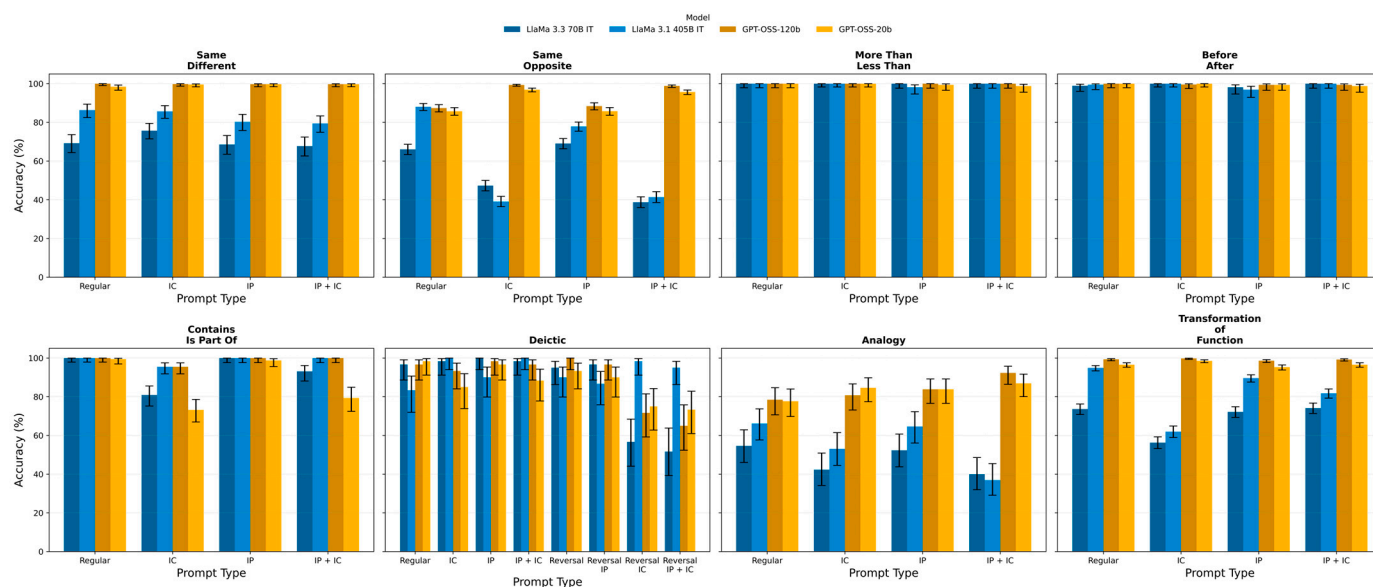


Figure A2. Performance on Different Trial Variations, grouped by Relations Involved. Error bars are 95% Wilson binomial confidence intervals. IC = incorrect Conclusion, IP = Irrelevant Premise.

Table A6. Model Accuracy (%) For Different Problem Variants Grouped by Relations Involved.

Block	Problem Variant	LLaMa 3.3 70B	LLaMa 3.3 405B	GPT OSS 20B	GPT OSS 120B
Same and Different	Regular	69.21	86.32	98.42	100.00
	Incorrect Conclusion	78.68	85.68	99.55	99.77
	Irrelevant Premise	68.57	80.29	99.71	99.71
	Irrelevant Incorrect	67.71	79.43	99.71	99.71
Same and Opposite	Regular	66.07	88.03	85.74	87.38
	Incorrect Conclusion	47.31	39.08	96.85	99.31
	Irrelevant Premise	69.07	77.88	85.76	88.39
	Irrelevant Incorrect	38.73	41.36	95.68	98.81
More Than and Less Than	Regular	100.00	100.00	100.00	100.00
	Incorrect Conclusion	100.00	100.00	100.00	100.00
	Irrelevant Premise	100.00	98.13	99.38	100.00
	Irrelevant Incorrect	100.00	100.00	98.75	100.00
Before and After	Regular	98.89	99.44	100.00	100.00
	Incorrect Conclusion	100.00	100.00	100.00	99.46
	Irrelevant Premise	98.13	96.88	99.38	99.38
	Irrelevant Incorrect	100.00	100.00	98.75	99.38
Contains and Is Part Of	Regular	100.00	100.00	99.44	100.00
	Incorrect Conclusion	80.91	95.46	73.18	95.46
	Irrelevant Premise	100.00	100.00	98.75	100.00
	Irrelevant Incorrect	93.13	100.00	79.38	100.00

Appendix B.2. Deictic Responding

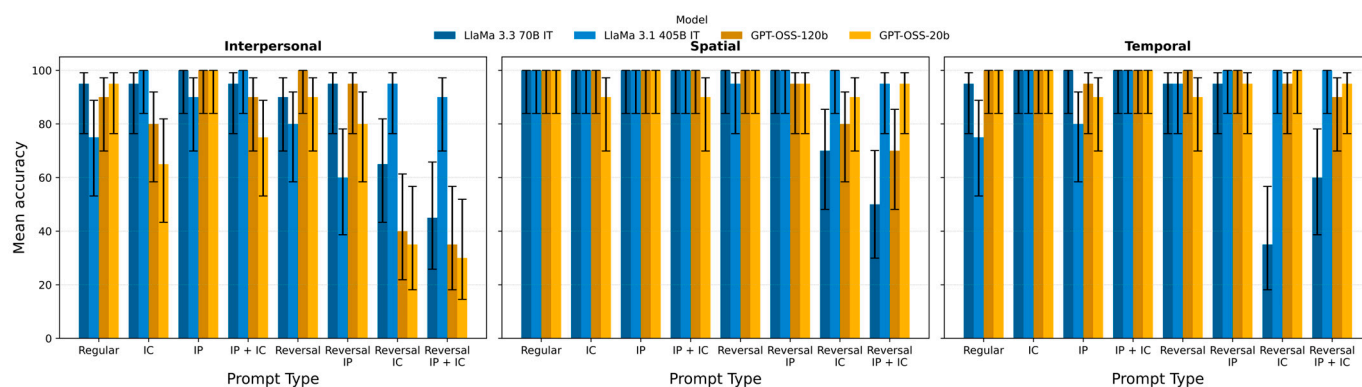


Figure A3. Performance on different variations in deictic relations trials, grouped by relation involved. Error bars are 95% Wilson binomial confidence intervals. IC = incorrect Conclusion, IP = Irrelevant Premise.

Table A7. Model Accuracy (%) For Different Deictic Problem Variants Grouped by Relations Involved.

Block	Problem Variant	LlaMa 3.3 70B	LlaMa 3.3 405B	GPT OSS 20B	GPT OSS 120B
Interpersonal	Regular	95.00	75.00	95.00	90.00
	Incorrect Conclusion	95.00	100.00	65.00	80.00
	Irrelevant Premise	100.00	90.00	100.00	100.00
	Irrelevant Incorrect	95.00	100.00	75.00	90.00
	Reversal	90.00	80.00	90.00	100.00
	Reversal Incorrect	65.00	95.00	35.00	40.00
	Reversal + Irrelevant	95.00	60.00	80.00	95.00
	Reversal Inc. + Irrelevant	45.00	90.00	30.00	35.00
Temporal	Regular	95.00	75.00	100.00	100.00
	Incorrect Conclusion	100.00	100.00	100.00	100.00
	Irrelevant Premise	100.00	80.00	90.00	95.00
	Irrelevant Incorrect	100.00	100.00	100.00	100.00
	Reversal	95.00	95.00	90.00	100.00
	Reversal Incorrect	35.00	100.00	100.00	95.00
	Reversal + Irrelevant	95.00	100.00	95.00	100.00
	Reversal Inc. + Irrelevant	60.00	100.00	95.00	90.00
Spatial	Regular	100.00	100.00	100.00	100.00
	Incorrect Conclusion	100.00	100.00	90.00	100.00
	Irrelevant Premise	100.00	100.00	100.00	100.00
	Irrelevant Incorrect	100.00	100.00	90.00	100.00
	Reversal	100.00	95.00	100.00	100.00
	Reversal Incorrect	70.00	100.00	90.00	80.00
	Reversal + Irrelevant	100.00	100.00	95.00	95.00
	Reversal Inc. + Irrelevant	50.00	95.00	95.00	70.00

Appendix B.3. Analogy

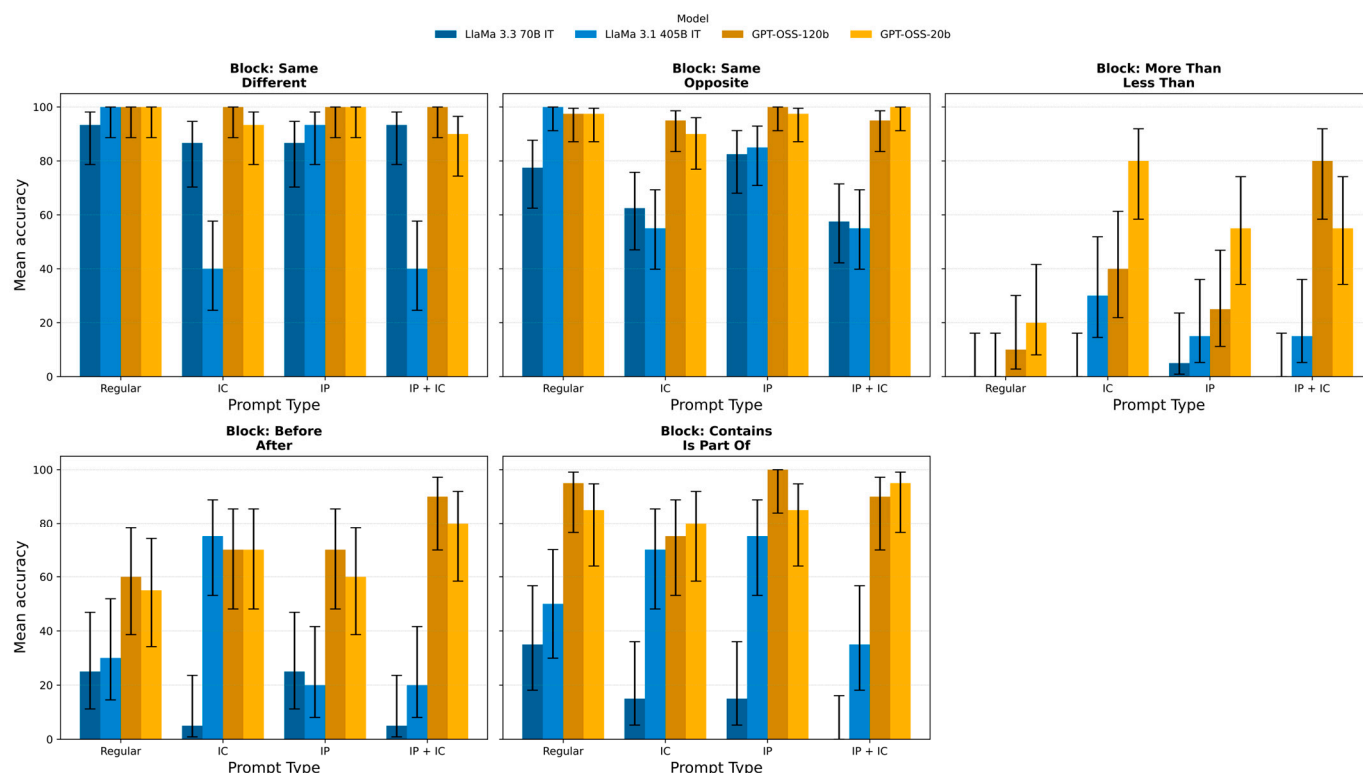


Figure A4. Performance on different variations in analogy problems, grouped by relations involved. Error bars are 95% Wilson binomial confidence intervals. IC = incorrect Conclusion, IP = Irrelevant Premise.

Table A8. Model Accuracy (%) For Different Analogy Problem Variants Grouped by Relations Involved.

Block	Problem Variant	LlaMa 3.3 70B	LlaMa 3.3 405B	GPT OSS 20B	GPT OSS 120B
Same and Different	Regular	93.33	100.00	100.00	100.00
	Incorrect Conclusion	86.67	40.00	93.33	100.00
	Irrelevant Premise	86.67	93.33	100.00	100.00
	Irrelevant Incorrect	93.33	40.00	90.00	100.00
Same and Opposite	Regular	75.5	100.00	97.50	97.50
	Incorrect Conclusion	62.50	55.00	90.00	95.00
	Irrelevant Premise	82.50	85.00	97.50	100.00
	Irrelevant Incorrect	57.50	55.00	100.00	95.00
More Than and Less Than	Regular	0.00	0.00	20.00	10.00
	Incorrect Conclusion	0.00	30.00	80.00	40.00
	Irrelevant Premise	5.00	15.00	55.00	25.00
	Irrelevant Incorrect	0.00	15.00	55.00	80.00
Before and After	Regular	25.00	30.00	55.00	60.00
	Incorrect Conclusion	5.00	75.00	70.00	70.00
	Irrelevant Premise	25.00	20.00	60.00	70.00
	Irrelevant Incorrect	5.00	20.00	80.00	90.00
Contains and Is Part Of	Regular	35.00	50.00	85.00	95.00
	Incorrect Conclusion	15.00	70.00	80.00	75.00
	Irrelevant Premise	15.00	75.00	85.00	100.00
	Irrelevant Incorrect	0.00	35.00	95.00	90.00

Appendix B.4. Transformation of Function

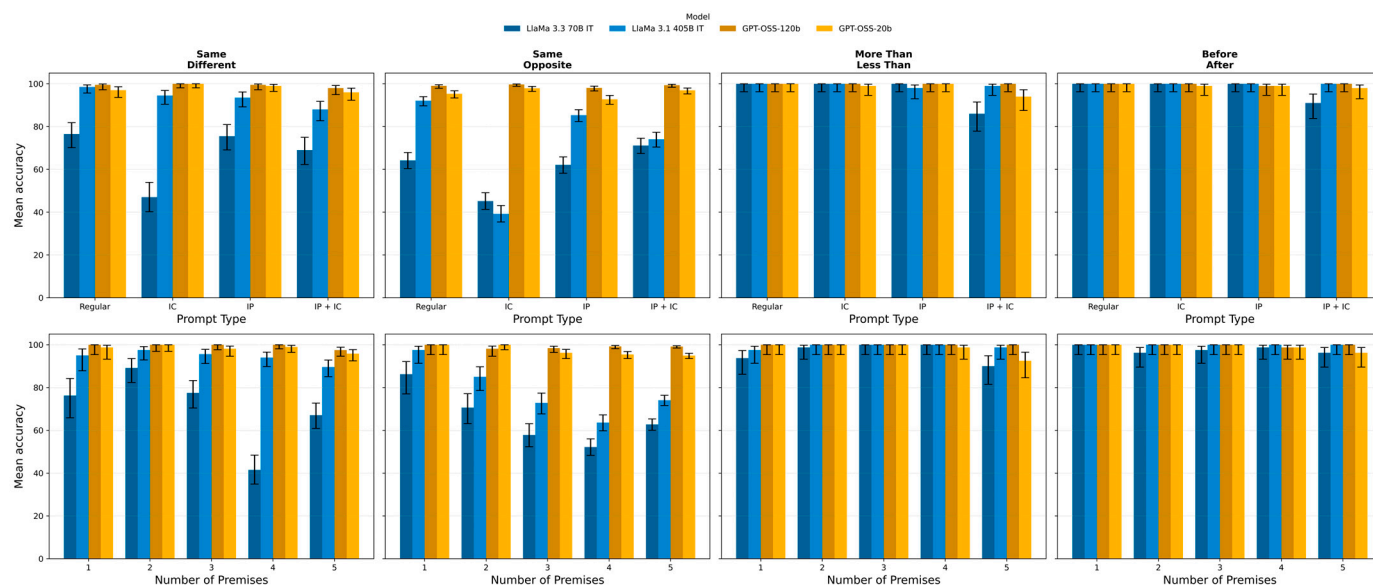


Figure A5. Performance on transformation of function trials for different problem variation and number of premises, grouped by relations involved. Error bars are 95% Wilson binomial confidence intervals. IC = incorrect Conclusion, IP = Irrelevant Premise.

Table A9. Model Accuracy (%) For Different Transformation of Function Problem Variants Grouped by Relations Involved.

Block	Problem Variant	LlaMa 3.3 70B	LlaMa 3.3 405B	GPT OSS 20B	GPT OSS 120B
Same and Different	Regular	76.50	98.50	97.00	99.50
	Incorrect Conclusion	47.00	94.50	100.00	100.00
	Irrelevant Premise	75.50	93.50	99.00	99.50
	Irrelevant Incorrect	69.00	88.00	96.00	98.00
Same and Opposite	Regular	64.19	92.10	95.32	99.03
	Incorrect Conclusion	45.16	39.19	97.90	99.68
	Irrelevant Premise	62.10	85.32	92.74	98.07
	Irrelevant Incorrect	71.13	74.03	96.94	99.36
More Than and Less Than	Regular	100.00	100.00	100.00	100.00
	Incorrect Conclusion	100.00	100.00	99.00	100.00
	Irrelevant Premise	100.00	98.00	100.00	100.00
	Irrelevant Incorrect	86.00	99.00	94.00	100.00
Before and After	Regular	100.00	100.00	100.00	100.00
	Incorrect Conclusion	100.00	100.00	99.00	100.00
	Irrelevant Premise	100.00	100.00	99.00	99.00
	Irrelevant Incorrect	91.00	100.00	98.00	100.00

Table A10. Model Accuracy (%) for Transformation of Function Problems with Increasing Number of Premises Grouped by Relations Involved.

Block	Number of Premises	LlaMa 3.3 70B	LlaMa 3.3 405B	GPT OSS 20B	GPT OSS 120B
Same and Different	One	76.25	95.00	98.75	100.00
	Two	89.17	97.50	100.00	100.00
	Three	77.50	95.63	98.13	100.00
	Four	41.50	94.00	99.00	100.00

Same and Opposite	Five	67.08	89.58	95.83	97.50
	One	86.25	97.50	100.00	100.00
	Two	70.63	85.00	100.00	98.13
	Three	57.81	72.81	96.25	98.44
	Four	52.19	63.59	84.47	99.22
More Than and Less Than	Five	62.73	74.06	94.92	99.14
	One	93.75	97.50	100.00	100.00
	Two	98.75	100.00	100.00	100.00
	Three	100.00	100.00	100.00	100.00
	Four	100.00	100.00	98.75	100.00
Before and After	Five	100.00	98.75	92.50	100.00
	One	100.00	100.00	100.00	100.00
	Two	96.25	100.00	100.00	100.00
	Three	97.50	100.00	100.00	100.00
	Four	98.75	100.00	98.75	98.75
	Five	98.75	100.00	96.25	100.00

Appendix C

Appendix C.1. Effect of Model and Block (Relations) on Performance

Table A11. Binomial GLM Results: Model $\times$ Block.

	Coeff.	Std. Err.	z	p
Intercept	−0.88	0.34	−2.57	0.010
Model (LlaMa 3.1 405B IT)	−0.72	0.16	−4.41	0.000
Model (GPT OSS 120B)	−5.35	0.59	−9.01	0.000
Model (GPT OSS 20B)	−4.14	0.39	−10.52	0.000
Block (Same–Opposite)	0.67	0.36	1.90	0.062
Block (More–Less)	−26.69	119.33	−7.04	0.823
Block (Before–After)	−4.09	0.58	−3.81	0.000
Block (Hierarchy)	−1.66	0.44	−2.76	0.000
Block (Deictic)	−1.00	0.36	1.86	0.006
Block (Analogy)	0.98	0.53	0.21	0.062
Block (Transformation)	0.07	0.34	2.50	0.831
Model (LlaMa 3.1 405B IT) $\times$ Block (Same–Opposite)	0.47	0.19	2.50	0.012
Model (GPT OSS 120B) $\times$ Block (Same–Opposite)	2.88	0.62	4.65	0.000
Model (GPT OSS 20B) $\times$ Block (Same–Opposite)	2.02	0.42	4.87	0.000
Model (LlaMa 3.1 405B IT) $\times$ Block (More–Less)	22.81	114.67	0.20	0.842
Model (GPT OSS 120B) $\times$ Block (More–Less)	5.35	119.99	0.05	0.964
Model (GPT OSS 20B) $\times$ Block (More–Less)	26.23	120.64	0.22	0.828
Model (LlaMa 3.1 405B IT) $\times$ Block (Before–After)	0.90	0.72	1.26	0.209
Model (GPT OSS 120B) $\times$ Block (Before–After)	4.84	0.99	4.90	0.000
Model (GPT OSS 20B) $\times$ Block (Before–After)	3.63	1.00	3.62	0.000
Model (LlaMa 3.1 405B IT) $\times$ Block (Hierarchy)	−1.01	0.89	−1.14	0.256
Model (GPT OSS 120B) $\times$ Block (Hierarchy)	3.62	1.06	3.43	0.001
Model (GPT OSS 20B) $\times$ Block (Hierarchy)	4.79	0.43	11.05	0.000
Model (LlaMa 3.1 405B IT) $\times$ Block (Deictic)	0.02	0.40	0.05	0.962
Model (GPT OSS 120B) $\times$ Block (Deictic)	5.05	0.77	6.59	0.000
Model (GPT OSS 20B) $\times$ Block (Deictic)	4.07	0.69	5.91	0.000
Model (LlaMa 3.1 405B IT) $\times$ Block (Analogy)	0.41	0.33	1.24	0.215
Model (GPT OSS 120B) $\times$ Block (Analogy)	3.60	0.69	5.25	0.000
Model (GPT OSS 20B) $\times$ Block (Analogy)	2.43	0.44	5.73	0.000

Model (LlaMa 3.1 405B IT) × Block (Transformation)	0.00	0.19	0.02	0.987
Model (GPT OSS 120B) × Block (Transformation)	1.28	0.63	2.01	0.044
Model (GPT OSS 20B) × Block (Transformation)	1.55	0.43	3.65	0.000

Appendix C.2. Effect of Model Problem Complexity on Performance

Table A12. Binomial GLM Results: Model × Complexity (same–different block).

	Coeff.	Std. Err.	z	p
Intercept	7.92	0.11	70.77	0.000
Model (LlaMa 3.1 405B IT)	1.12	0.16	7.11	0.000
Model (GPT OSS 120B)	4.66	0.16	29.48	0.000
Model (GPT OSS 20B)	2.45	0.16	45.46	0.000
Complexity (2 Premises)	2.07	0.13	16.00	0.000
Complexity (3 Premises)	−6.59	0.18	−37.17	0.000
Complexity (4 Premises)	−7.45	0.15	−49.03	0.000
Complexity (5 Premises)	−7.80	0.14	−53.981	0.000
Model (LlaMa 3.1 405B IT) × Complexity (2 Premises)	0.13	0.18	0.74	0.463
Model (GPT OSS 120B) × Complexity (2 Premises)	0.00	0.18	0.00	1.00
Model (GPT OSS 20B) × Complexity (2 Premises)	−7.65	0.73	−10.45	0.000
Model (LlaMa 3.1 405B IT) × Complexity (3 Premises)	−0.11	0.29	−0.38	0.705
Model (GPT OSS 120B) × Complexity (3 Premises)	−0.22	1.03	−0.21	0.834
Model (GPT OSS 20B) × Complexity (3 Premises)	1.98	1.02	1.95	0.051
Model (LlaMa 3.1 405B IT) × Complexity (4 Premises)	−0.18	0.23	−0.77	0.440
Model (GPT OSS 120B) × Complexity (4 Premises)	0.85	1.01	0.84	0.400
Model (GPT OSS 20B) × Complexity (4 Premises)	1.97	0.61	3.32	0.001
Model (LlaMa 3.1 405B IT) × Complexity (5 Premises)	−0.45	0.21	−2.18	0.029
Model (GPT OSS 120B) × Complexity (5 Premises)	1.39	1.02	1.37	0.171
Model (GPT OSS 20B) × Complexity (5 Premises)	2.21	0.53	4.16	0.000

Table A13. Binomial GLM Results: Model × Complexity (same–opposite block).

	Coeff.	Std. Err.	z	p
Intercept	6.40	0.11	57–159	0.000
Model (LlaMa 3.1 405B IT)	2.39	0.16	15.13	0.000
Model (GPT OSS 120B)	3.05	0.16	19.30	0.000
Model (GPT OSS 20B)	−2.00	1.05	−1.91	0.056
Complexity (2 Premises)	−4.83	0.19	−26.00	0.000
Complexity (3 Premises)	−6.01	0.39	−43.55	0.000
Complexity (4 Premises)	−6.26	0.13	−49.99	0.000
Complexity (5 Premises)	−6.40	0.12	−53.89	0.000
Model (LlaMa 3.1 405B IT) × Complexity (2 Premises)	−2.18	0.70	−8.11	0.000
Model (GPT OSS 120B) × Complexity (2 Premises)	−0.26	0.55	−0.47	0.639
Model (GPT OSS 20B) × Complexity (2 Premises)	2.98	1.08	2.77	0.006
Model (LlaMa 3.1 405B IT) × Complexity (3 Premises)	−2.08	0.20	−10.61	0.000
Model (GPT OSS 120B) × Complexity (3 Premises)	−1.16	0.22	−5.18	0.000
Model (GPT OSS 20B) × Complexity (3 Premises)	3.70	1.06	3.51	0.000
Model (LlaMa 3.1 405B IT) × Complexity (4 Premises)	−2.05	0.17	−11.58	0.000
Model (GPT OSS 120B) × Complexity (4 Premises)	−0.58	0.20	−2.86	0.004
Model (GPT OSS 20B) × Complexity (4 Premises)	4.32	1.05	4.10	0.000
Model (LlaMa 3.1 405B IT) × Complexity (5 Premises)	−2.18	0.19	−13.00	0.000
Model (GPT OSS 120B) × Complexity (5 Premises)	−0.8	0.18	−2.05	0.040
Model (GPT OSS 20B) × Complexity (5 Premises)	4.27	1.05	4.08	0.000

Table A14. Binomial GLM Results: Model $\times$ Complexity (more–less block).

	Coeff.	Std. Err.	z	p
Intercept	8.72	0.11	77.96	0.000
Model (LlaMa 3.1 405B IT)	−5.46	0.61	−8.98	0.000
Model (GPT OSS 120B)	1.84	0.16	11.61	0.000
Model (GPT OSS 20B)	−0.12	0.16	−0.78	0.437
Complexity (2 Premises)	1.33	0.14	9.73	0.000
Complexity (3 Premises)	4.98	0.14	36.37	0.000
Complexity (4 Premises)	4.52	0.14	11.10	0.000
Complexity (5 Premises)	4.98	0.14	36.37	0.000
Model (LlaMa 3.1 405B IT) $\times$ Complexity (2 Premises)	5.02	0.62	8.12	0.000
Model (GPT OSS 120B) $\times$ Complexity (2 Premises)	0.40	0.19	2.05	0.041
Model (GPT OSS 20B) $\times$ Complexity (2 Premises)	−5.55	0.74	−7.52	0.000
Model (LlaMa 3.1 405B IT) $\times$ Complexity (3 Premises)	1.47	0.62	2.38	0.017
Model (GPT OSS 120B) $\times$ Complexity (3 Premises)	0.00	0.19	0.00	1.00
Model (GPT OSS 20B) $\times$ Complexity (3 Premises)	0.00	0.19	0.00	1.00
Model (LlaMa 3.1 405B IT) $\times$ Complexity (4 Premises)	5.15	0.62	8.33	0.000
Model (GPT OSS 120B) $\times$ Complexity (4 Premises)	0.00	0.19	0.00	1.00
Model (GPT OSS 20B) $\times$ Complexity (4 Premises)	−5.04	1.03	−4.90	0.000
Model (LlaMa 3.1 405B IT) $\times$ Complexity (5 Premises)	1.47	0.62	2.38	0.017
Model (GPT OSS 120B) $\times$ Complexity (5 Premises)	0.00	0.19	0.00	1.00
Model (GPT OSS 20B) $\times$ Complexity (5 Premises)	0.00	0.19	0.00	1.00

Table A15. Binomial GLM Results: Model $\times$ Complexity (before–after block).

	Coeff.	Std. Err.	z	p
Intercept	3.26	0.60	5.48	0.000
Model (LlaMa 3.1 405B IT)	−0.74	0.73	−1.01	0.314
Model (GPT OSS 120B)	5.09	0.61	8.42	0.000
Model (GPT OSS 20B)	5.13	0.61	8.48	0.000
Complexity (2 Premises)	1.83	1.18	1.55	0.121
Complexity (3 Premises)	1.83	1.18	1.55	0.121
Complexity (4 Premises)	5.28	0.60	8.80	0.000
Complexity (5 Premises)	5.34	0.60	8.89	0.000
Model (LlaMa 3.1 405B IT) $\times$ Complexity (2 Premises)	5.15	1.26	4.10	0.000
Model (GPT OSS 120B) $\times$ Complexity (2 Premises)	0.00	1.19	0.00	1.00
Model (GPT OSS 20B) $\times$ Complexity (2 Premises)	0.00	1.19	0.00	1.00
Model (LlaMa 3.1 405B IT) $\times$ Complexity (3 Premises)	5.15	1.26	4.01	0.000
Model (GPT OSS 120B) $\times$ Complexity (3 Premises)	0.00	1.19	0.00	1.00
Model (GPT OSS 20B) $\times$ Complexity (3 Premises)	0.00	1.19	0.00	1.00
Model (LlaMa 3.1 405B IT) $\times$ Complexity (4 Premises)	3.19	0.74	4.31	0.000
Model (GPT OSS 120B) $\times$ Complexity (4 Premises)	−8.55	1.18	−7.22	0.000
Model (GPT OSS 20B) $\times$ Complexity (4 Premises)	−9.29	0.94	−9.88	0.000
Model (LlaMa 3.1 405B IT) $\times$ Complexity (5 Premises)	3.19	0.74	4.31	0.000
Model (GPT OSS 120B) $\times$ Complexity (5 Premises)	−9.31	0.94	−9.90	0.000
Model (GPT OSS 20B) $\times$ Complexity (5 Premises)	8.64	1.18	−7.31	0.000

Table A16. Binomial GLM Results: Model $\times$ Complexity (hierarchy block).

	Coeff.	Std. Err.	z	p
Intercept	6.80	0.11	60.74	0.000
Model (LlaMa 3.1 405B IT)	5.34	0.16	33.77	0.000
Model (GPT OSS 120B)	5.26	0.16	33.27	0.000

Model (GPT OSS 20B)	−2.38	1.06	−2.24	0.025
Complexity (2 Premises)	−4.19	0.33	−12.59	0.000
Complexity (3 Premises)	−4.91	0.26	−18.92	0.000
Complexity (4 Premises)	−4.19	0.33	−12.59	0.000
Complexity (5 Premises)	−4.09	0.35	−11.81	0.000
Model (LlaMa 3.1 405B IT) × Complexity (2 Premises)	1.97	0.36	5.47	0.000
Model (GPT OSS 120B) × Complexity (2 Premises)	1.92	0.36	5.33	0.000
Model (GPT OSS 20B) × Complexity (2 Premises)	1.78	1.13	1.57	0.117
Model (LlaMa 3.1 405B IT) × Complexity (3 Premises)	−4.52	0.43	−10.49	0.000
Model (GPT OSS 120B) × Complexity (3 Premises)	−4.44	0.43	−10.28	0.000
Model (GPT OSS 20B) × Complexity (3 Premises)	1.84	1.10	1.67	0.096
Model (LlaMa 3.1 405B IT) × Complexity (4 Premises)	1.97	0.36	5.47	0.000
Model (GPT OSS 120B) × Complexity (4 Premises)	1.92	0.36	5.33	0.000
Model (GPT OSS 20B) × Complexity (4 Premises)	1.61	1.13	1.42	0.155
Model (LlaMa 3.1 405B IT) × Complexity (5 Premises)	1.87	0.37	5.01	0.000
Model (GPT OSS 120B) × Complexity (5 Premises)	1.81	0.37	4.87	0.000
Model (GPT OSS 20B) × Complexity (5 Premises)	1.61	1.14	1.42	0.155

Appendix C.3. Effect of Model Problem Variations on Performance

Table A17. Binomial GLM Results: Model × Problem Variant (same–different block).

	Coeff.	Std. Err.	z	p
Intercept	1.14	0.11	10.22	0.000
Model (LlaMa 3.1 405B IT)	0.65	0.18	3.72	0.000
Model (GPT OSS 120B)	4.96	1.02	4.88	0.000
Model (GPT OSS 20B)	4.26	0.72	5.89	0.000
Variant (Irrelevant)	−0.35	0.16	−2.22	0.027
Variant (Irrelevant Incorrect)	−0.39	0.16	−2.47	0.014
Variant (Regular)	−0.32	0.16	−2.07	0.039
Model (LlaMa 3.1 405B IT) × Variant (Irrelevant)	−0.03	0.25	−0.12	0.906
Model (GPT OSS 120B) × Variant (Irrelevant)	0.11	1.42	0.08	0.938
Model (GPT OSS 20B) × Variant (Irrelevant)	0.82	1.24	0.66	0.512
Model (LlaMa 3.1 405B IT) × Variant (Irr. Incorrect)	−0.04	0.25	−0.18	0.859
Model (GPT OSS 120B) × Variant (Irr. Incorrect)	0.14	1.42	0.10	0.923
Model (GPT OSS 20B) × Variant (Irr. Incorrect)	0.83	1.22	0.68	0.498
Model (LlaMa 3.1 405B IT) × Variant (Regular)	0.38	0.26	1.48	0.140
Model (GPT OSS 120B) × Variant (Regular)	5.18	1.02	5.07	0.000
Model (GPT OSS 20B) × Variant (Regular)	−0.94	0.84	−1.12	0.262

Table A18. Binomial GLM Results: Model × Problem Variant (same–opposite block).

	Coeff.	Std. Err.	z	p
Intercept	−0.11	0.06	−1.94	0.052
Model (LlaMa 3.1 405B IT)	−0.34	0.08	−4.23	0.000
Model (GPT OSS 120B)	5.07	0.34	14.96	0.000
Model (GPT OSS 20B)	3.53	0.17	21.01	0.000
Variant (Irrelevant)	0.91	0.08	10.85	0.000
Variant (Irrelevant Incorrect)	−0.35	0.08	−4.30	0.000
Variant (Regular)	0.77	0.08	9.43	0.000
Model (LlaMa 3.1 405B IT) × Variant (Irrelevant)	0.79	0.12	6.42	0.000
Model (GPT OSS 120B) × Variant (Irrelevant)	−3.85	0.38	−10.79	0.000
Model (GPT OSS 20B) × Variant (Irrelevant)	−2.54	0.20	−12.83	0.000
Model (LlaMa 3.1 405B IT) × Variant (Irr. Incorrect)	0.45	0.12	3.85	0.000

Model (GPT OSS 120B) × Variant (Irr. Incorrect)	−0.19	0.44	−0.44	0.659
Model (GPT OSS 20B) × Variant (Irr. Incorrect)	0.02	0.23	0.10	0.918
Model (LlaMa 3.1 405B IT) × Variant (Regular)	1.67	0.13	12.50	0.000
Model (GPT OSS 120B) × Variant (Regular)	−3.81	0.36	−10.72	0.000
Model (GPT OSS 20B) × Variant (Regular)	−2.41	0.20	−12.24	0.000

Table A19. Binomial GLM Results: Model × Problem Variant (more–less block).

	Coeff.	Std. Err.	z	p
Intercept	11.97	0.07	177.59	0.000
Model (LlaMa 3.1 405B IT)	0.00	0.10	0.00	1.00
Model (GPT OSS 120B)	0.79	0.10	8.29	0.000
Model (GPT OSS 20B)	−2.30	0.10	−24.12	0.000
Variant (Irrelevant)	−2.64	0.10	−25.45	0.000
Variant (Irrelevant Incorrect)	−0.23	0.10	−2.17	0.030
Variant (Regular)	0.63	0.10	6.23	0.000
Model (LlaMa 3.1 405B IT) × Variant (Irrelevant)	−5.37	0.60	−9.00	0.000
Model (GPT OSS 120B) × Variant (Irrelevant)	0.00	0.15	0.00	1.00
Model (GPT OSS 20B) × Variant (Irrelevant)	−1.94	1.03	−1.89	0.059
Model (LlaMa 3.1 405B IT) × Variant (Irr. Incorrect)	0.04	0.15	0.27	0.785
Model (GPT OSS 120B) × Variant (Irr. Incorrect)	0.00	0.15	0.00	1.00
Model (GPT OSS 20B) × Variant (Irr. Incorrect)	−5.07	0.73	−6.98	0.000
Model (LlaMa 3.1 405B IT) × Variant (Regular)	0.00	0.14	0.00	1.00
Model (GPT OSS 120B) × Variant (Regular)	0.00	0.14	0.00	1.00
Model (GPT OSS 20B) × Variant (Regular)	0.36	0.14	2.54	0.011

Table A20. Binomial GLM Results: Model × Problem Variant (before–after block).

	Coeff.	Std. Err.	z	p
Intercept	9.02	0.07	133.82	0.000
Model (LlaMa 3.1 405B IT)	0.70	0.10	7.33	0.000
Model (GPT OSS 120B)	−3.62	1.02	−3.56	0.000
Model (GPT OSS 20B)	1.24	0.10	13.05	0.000
Variant (Irrelevant)	−5.06	0.59	−8.60	0.000
Variant (Irrelevant Incorrect)	0.00	0.10	0.00	1.00
Variant (Regular)	−4.52	0.72	−6.27	0.000
Model (LlaMa 3.1 405B IT) × Variant (Irrelevant)	−1.22	0.75	−1.64	0.101
Model (GPT OSS 120B) × Variant (Irrelevant)	4.72	1.54	3.07	0.002
Model (GPT OSS 20B) × Variant (Irrelevant)	−0.13	1.17	−0.11	0.909
Model (LlaMa 3.1 405B IT) × Variant (Irr. Incorrect)	0.05	0.15	0.32	0.746
Model (GPT OSS 120B) × Variant (Irr. Incorrect)	−0.32	1.44	−0.22	0.827
Model (GPT OSS 20B) × Variant (Irr. Incorrect)	−5.89	0.73	−8.11	0.000
Model (LlaMa 3.1 405B IT) × Variant (Regular)	−0.00	1.24	0.00	0.998
Model (GPT OSS 120B) × Variant (Regular)	9.40	1.25	7.53	0.000
Model (GPT OSS 20B) × Variant (Regular)	4.10	0.73	5.63	0.000

Table A21. Binomial GLM Results: Model × Problem Variant (hierarchy block).

	Coeff.	Std. Err.	Z	p
Intercept	1.45	0.17	8.42	0.000
Model (LlaMa 3.1 405B IT)	1.60	0.37	4.37	0.000
Model (GPT OSS 120B)	1.60	0.37	4.37	0.000
Model (GPT OSS 20B)	−0.44	0.23	−1.93	0.054
Variant (Irrelevant)	8.04	0.19	42.57	0.000

Variant (Irrelevant Incorrect)	1.16	0.36	3.26	0.001
Variant (Regular)	7.57	0.29	40.43	0.000
Model (LlaMa 3.1 405B IT) × Variant (Irrelevant)	0.00	0.38	0.00	1.00
Model (GPT OSS 120B) × Variant (Irrelevant)	0.00	0.38	0.00	1.00
Model (GPT OSS 20B) × Variant (Irrelevant)	−4.67	0.76	−6.19	0.000
Model (LlaMa 3.1 405B IT) × Variant (Irr. Incorrect)	5.79	0.49	11.87	0.000
Model (GPT OSS 120B) × Variant (Irr. Incorrect)	5.79	0.49	11.87	0.000
Model (GPT OSS 20B) × Variant (Irr. Incorrect)	−0.82	0.43	−1.88	0.060
Model (LlaMa 3.1 405B IT) × Variant (Regular)	0.00	0.38	0.00	1.00
Model (GPT OSS 120B) × Variant (Regular)	0.00	0.38	0.00	1.00
Model (GPT OSS 20B) × Variant (Regular)	−3.37	1.05	−3.21	0.001

Appendix C.4. Effect of Model and Problem Variations on Performance in Analogy Blocks (By Relation)

Table A22. Binomial GLM Results: Model × Problem Variant (same–different analogies).

	Coeff.	Std. Err.	z	p
Intercept	2.64	0.73	3.61	0.000
Model (LlaMa 3.1 405B IT)	11.23	0.75	14.88	0.000
Model (GPT OSS 120B)	14.78	0.75	19.60	0.000
Model (GPT OSS 20B)	12.14	0.75	16.09	0.000
Variant (Irrelevant)	−0.77	0.91	−0.85	0.398
Variant (Incorrect)	−0.77	0.91	−0.85	0.397
Variant (Irrelevant Incorrect)	0.00	1.04	0.00	1.00
Model (LlaMa 3.1 405B IT) × Variant (Irrelevant)	−10.46	1.18	−8.87	0.000
Model (GPT OSS 120B) × Variant (Irrelevant)	2.59	0.94	2.74	0.006
Model (GPT OSS 20B) × Variant (Irrelevant)	8.94	0.94	9.47	0.000
Model (LlaMa 3.1 405B IT) × Variant (Incorrect)	−13.50	1.00	−13.53	0.000
Model (GPT OSS 120B) × Variant (Incorrect)	2.60	0.94	2.76	0.006
Model (GPT OSS 20B) × Variant (Incorrect)	−11.37	1.18	−9.64	0.000
Model (LlaMa 3.1 405B IT) × Variant (Irr. Incor.)	−14.27	1.12	−12.80	0.000
Model (GPT OSS 120B) × Variant (Irr. Incorrect)	1.22	1.07	1.14	0.253
Model (GPT OSS 20B) × Variant (Irr. Incorrect)	−12.58	1.22	−10.36	0.000

Table A23. Binomial GLM Results: Model × Problem Variant (same–opposite analogies).

	Coeff.	Std. Err.	z	p
Intercept	1.24	0.38	3.27	0.001
Model (LlaMa 3.1 405B IT)	11.98	0.41	29.20	0.000
Model (GPT OSS 120B)	2.43	1.08	2.24	0.025
Model (GPT OSS 20B)	2.43	1.08	2.25	0.025
Variant (Irrelevant)	0.32	0.56	0.56	0.574
Variant (Incorrect)	−0.72	0.50	−1.45	0.147
Variant (Irrelevant Incorrect)	−0.93	0.50	−1.89	0.059
Model (LlaMa 3.1 405B IT) × Variant (Irrelevant)	−11.79	0.73	−16.09	0.000
Model (GPT OSS 120B) × Variant (Irrelevant)	10.80	1.17	9.22	0.000
Model (GPT OSS 20B) × Variant (Irrelevant)	−0.32	1.54	−0.21	0.836
Model (LlaMa 3.1 405B IT) × Variant (Incorrect)	−12.29	0.61	−20.04	0.000
Model (GPT OSS 120B) × Variant (Incorrect)	0.00	1.34	0.00	1.00
Model (GPT OSS 20B) × Variant (Incorrect)	−0.74	1.24	−0.59	0.553
Model (LlaMa 3.1 405B IT) × Variant (Irr. Incor.)	−12.08	0.61	−19.81	0.000
Model (GPT OSS 120B) × Variant (Irr. Incorrect)	0.21	1.34	0.16	0.873
Model (GPT OSS 20B) × Variant (Irr. Incorrect)	13.53	1.14	11.91	0.000

Table A24. Binomial GLM Results: Model $\times$ Problem Variant (more–less analogies).

	Coeff.	Std. Err.	z	p
Intercept	−16.41	0.18	−90.55	0.000
Model (LlaMa 3.1 405B IT)	0.79	0.29	2.79	0.005
Model (GPT OSS 120B)	14.21	0.77	18.51	0.000
Model (GPT OSS 20B)	15.02	0.59	25.59	0.000
Variant (Irrelevant)	13.47	1.04	12.92	0.000
Variant (Incorrect)	−2.57	0.65	−3.95	0.000
Variant (Irrelevant Incorrect)	−2.53	0.00	0.00	1.00
Model (LlaMa 3.1 405B IT) $\times$ Variant (Irrelevant)	0.41	1.23	0.34	0.738
Model (GPT OSS 120B) $\times$ Variant (Irrelevant)	−12.37	1.38	−8.96	0.000
Model (GPT OSS 20B) $\times$ Variant (Irrelevant)	−11.88	1.26	−9.40	0.000
Model (LlaMa 3.1 405B IT) $\times$ Variant (Incorrect)	17.34	0.82	21.13	0.000
Model (GPT OSS 120B) $\times$ Variant (Incorrect)	4.36	1.06	4.12	0.000
Model (GPT OSS 20B) $\times$ Variant (Incorrect)	5.34	1.02	5.24	0.000
Model (LlaMa 3.1 405B IT) $\times$ Variant (Irr. Incor.)	16.41	0.62	26.47	0.000
Model (GPT OSS 120B) $\times$ Variant (Irr. Incorrect)	6.11	0.88	6.96	0.000
Model (GPT OSS 20B) $\times$ Variant (Irr. Incorrect)	4.12	0.69	5.98	0.000

Table A25. Binomial GLM Results: Model $\times$ Problem Variant (before–after analogies).

	Coeff.	Std. Err.	z	p
Intercept	−1.10	0.52	−2.13	0.033
Model (LlaMa 3.1 405B IT)	0.25	0.71	0.35	0.724
Model (GPT OSS 120B)	1.50	0.69	2.18	0.029
Model (GPT OSS 20B)	1.30	0.69	1.90	0.058
Variant (Irrelevant)	0.00	0.73	0.00	1.00
Variant (Incorrect)	−1.85	1.15	−1.61	0.108
Variant (Irrelevant Incorrect)	−1.85	1.15	−1.61	0.108
Model (LlaMa 3.1 405B IT) $\times$ Variant (Irrelevant)	−0.54	1.04	−0.52	0.605
Model (GPT OSS 120B) $\times$ Variant (Irrelevant)	0.44	0.99	0.45	0.655
Model (GPT OSS 20B) $\times$ Variant (Irrelevant)	0.20	0.97	0.21	0.833
Model (LlaMa 3.1 405B IT) $\times$ Variant (Incorrect)	3.79	1.35	2.81	0.005
Model (GPT OSS 120B) $\times$ Variant (Incorrect)	2.29	1.33	1.72	0.085
Model (GPT OSS 20B) $\times$ Variant (Incorrect)	2.49	1.33	1.88	0.060
Model (LlaMa 3.1 405B IT) $\times$ Variant (Irr. Incor.)	1.31	1.37	0.96	0.339
Model (GPT OSS 120B) $\times$ Variant (Irr. Incorrect)	3.64	1.44	2.52	0.012
Model (GPT OSS 20B) $\times$ Variant (Irr. Incorrect)	3.03	1.35	2.24	0.025

Table A26. Binomial GLM Results: Model $\times$ Problem Variant (hierarchy analogies).

	Coeff.	Std. Err.	z	p
Intercept	−0.62	0.47	−1.32	0.187
Model (LlaMa 3.1 405B IT)	0.62	0.65	0.96	0.339
Model (GPT OSS 120B)	3.57	1.13	3.15	0.002
Model (GPT OSS 20B)	2.35	0.78	3.01	0.003
Variant (Irrelevant)	−1.12	0.78	−1.43	0.154
Variant (Incorrect)	−1.11	0.78	−1.43	0.154
Variant (Irrelevant Incorrect)	−13.45	0.52	−25.89	0.000
Model (LlaMa 3.1 405B IT) $\times$ Variant (Irrelevant)	2.21	1.04	2.13	0.033
Model (GPT OSS 120B) $\times$ Variant (Irrelevant)	13.57	1.31	10.35	0.000
Model (GPT OSS 20B) $\times$ Variant (Irrelevant)	1.12	1.18	0.94	0.345
Model (LlaMa 3.1 405B IT) $\times$ Variant (Incorrect)	1.96	1.02	1.91	0.056

Model (GPT OSS 120B) × Variant (Incorrect)	−0.74	1.39	−0.53	0.597
Model (GPT OSS 20B) × Variant (Incorrect)	0.77	1.15	0.67	0.504
Model (LlaMa 3.1 405B IT) × Variant (Irr. Incor.)	12.83	0.83	15.45	0.000
Model (GPT OSS 120B) × Variant (Irr. Incorrect)	12.70	1.37	9.25	0.000
Model (GPT OSS 20B) × Variant (Irr. Incorrect)	14.66	1.31	11.20	0.000

Appendix D

Appendix D.1. Effect of Problem Complexity and Variations When Premise Order Is Randomized

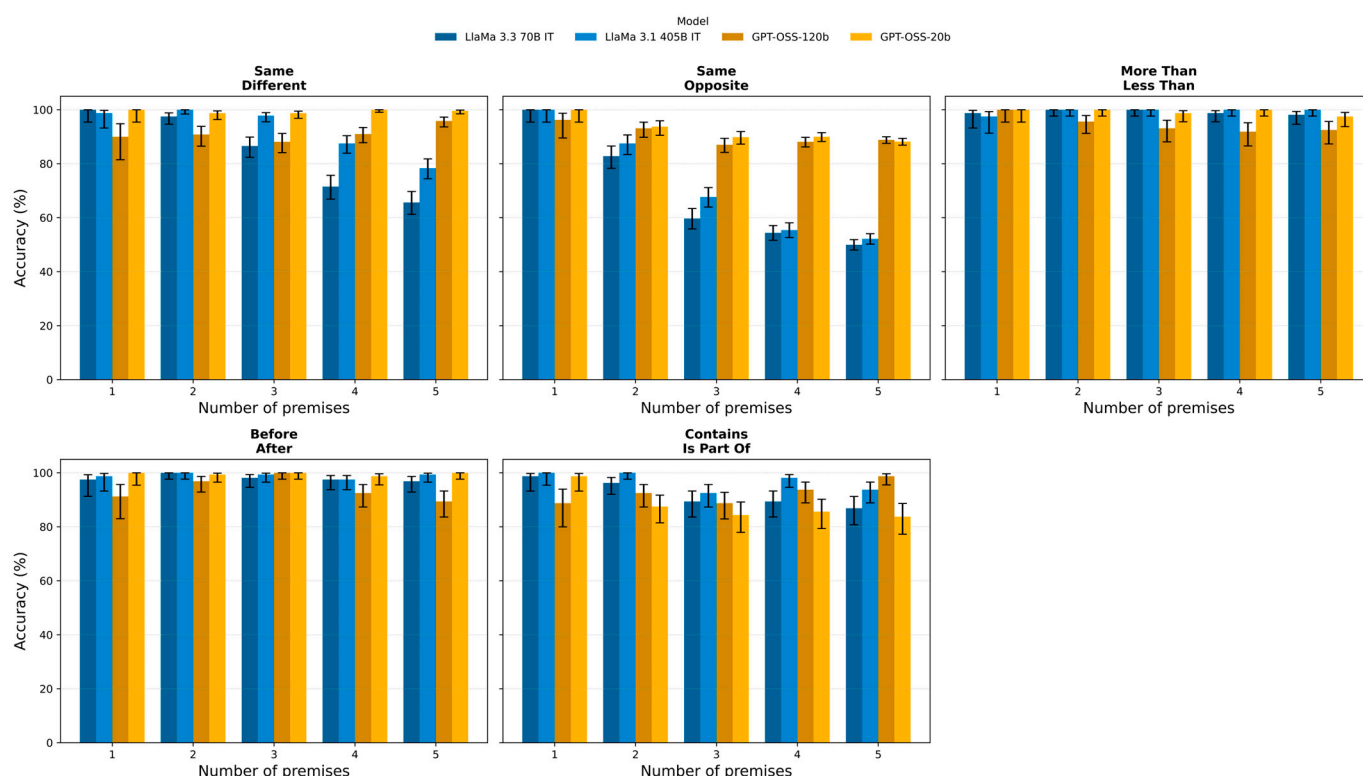


Figure A6. Replication study performance for increasing number of relational premises grouped by relations involved. Error bars are 95% Wilson confidence intervals.

Table A27. Model Accuracy (%) For Increasing Number of Premises in the Replication Study, Grouped by Relations Involved.

Block	Number of Premises	LLaMa 3.3 70B	LLaMa 3.3 405B	GPT OSS 20B	GPT OSS 120B
Same and Different	One	100.00	98.75	100.00	90.00
	Two	97.50	100.00	98.75	90.83
	Three	86.56	97.81	98.75	88.13
	Four	71.50	87.50	100.00	91.00
	Five	65.63	78.33	99.58	95.83
Same and Opposite	One	100.00	100.00	100.00	96.25
	Two	82.81	87.50	93.75	93.13
	Three	59.70	67.66	89.84	87.03
	Four	54.38	55.39	90.00	88.13
	Five	49.96	52.15	88.20	88.83
More Than and Less Than	One	98.75	97.50	100.00	100.00
	Two	100.00	100.00	100.00	95.63
	Three	100.00	100.00	98.75	93.13
	Four	98.75	100.00	100.00	91.88

Before and After	Five	98.13	100.00	97.50	92.50
	One	97.50	98.75	100.00	91.25
	Two	100.00	100.00	99.38	96.88
	Three	98.13	99.38	100.00	100.00
	Four	97.50	97.50	98.75	92.50
Contains and Is Part Of	Five	96.88	99.38	100.00	89.38
	One	98.75	100.00	98.75	88.75
	Two	96.25	100.00	87.50	92.50
	Three	89.38	92.50	84.38	88.75
	Four	89.38	98.13	85.63	93.75
	Five	86.88	93.75	83.75	98.75

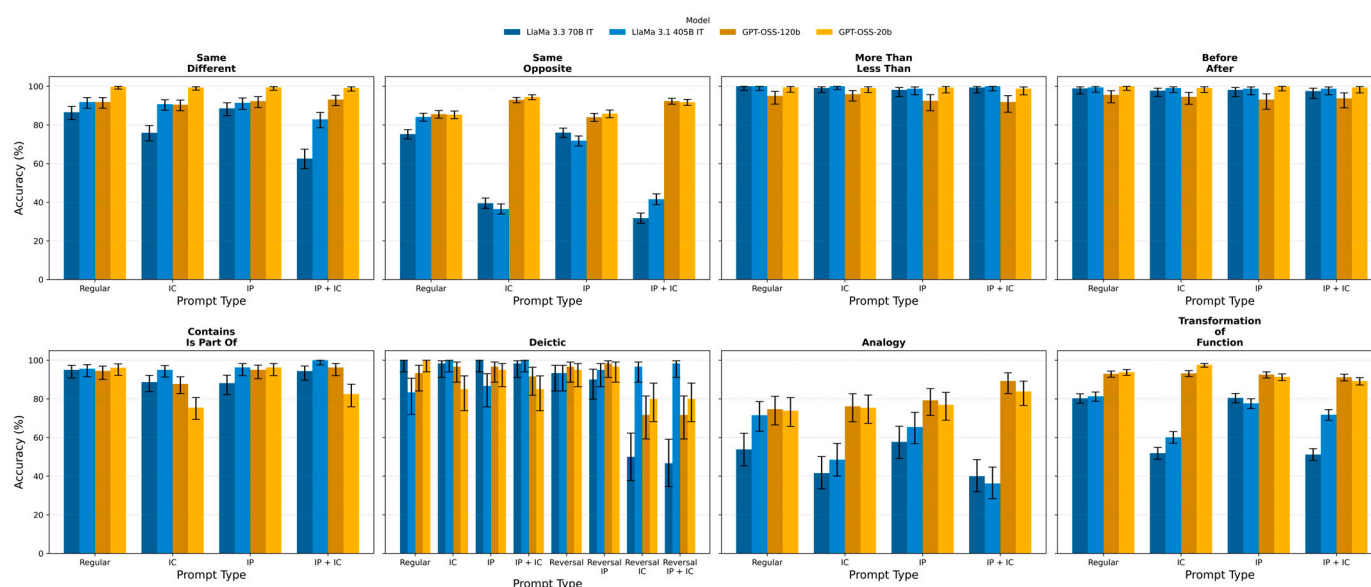


Figure A7. Replication study performance on Different Trial Variations, grouped by Relations Involved. Error bars are 95% Wilson confidence intervals. IC = incorrect Conclusion, IP = Irrelevant Premise.

Table A28. Model Accuracy (%) For Different Problem Variants in the Replication Study, Grouped by Relations Involved.

Block	Problem Variant	LLaMa 3.3 70B	LLaMa 3.3 405B	GPT OSS 20B	GPT OSS 120B
Same and Different	Regular	86.58	91.84	99.74	91.84
	Incorrect Conclusion	75.91	90.68	99.32	90.46
	Irrelevant Premise	88.57	91.43	99.43	92.29
	Irrelevant Incorrect	62.57	82.86	99.14	93.14
Same and Opposite	Regular	75.25	84.10	85.33	85.57
	Incorrect Conclusion	39.46	36.46	94.46	93.00
	Irrelevant Premise	76.02	71.78	85.85	83.98
	Irrelevant Incorrect	31.70	41.53	91.78	92.37
More Than and Less Than	Regular	100.00	100.00	99.44	95.00
	Incorrect Conclusion	99.09	100.00	99.09	95.91
	Irrelevant Premise	98.13	98.75	99.38	92.50
	Irrelevant Incorrect	99.38	100.00	98.75	91.88
Before and After	Regular	98.89	99.44	100.00	95.56
	Incorrect Conclusion	97.73	99.09	99.09	94.55
	Irrelevant Premise	98.13	98.75	100.00	93.13

Contains andIs Part Of	Irrelevant Incorrect	97.50	98.75	99.38	93.75
	Regular	95.00	95.56	96.11	94.44
	Incorrect Conclusion	88.64	95.00	75.46	87.73
	Irrelevant Premise	88.13	96.25	96.25	95.00
	Irrelevant Incorrect	94.38	100.00	82.50	96.25

Appendix D.2. Deictic Responding

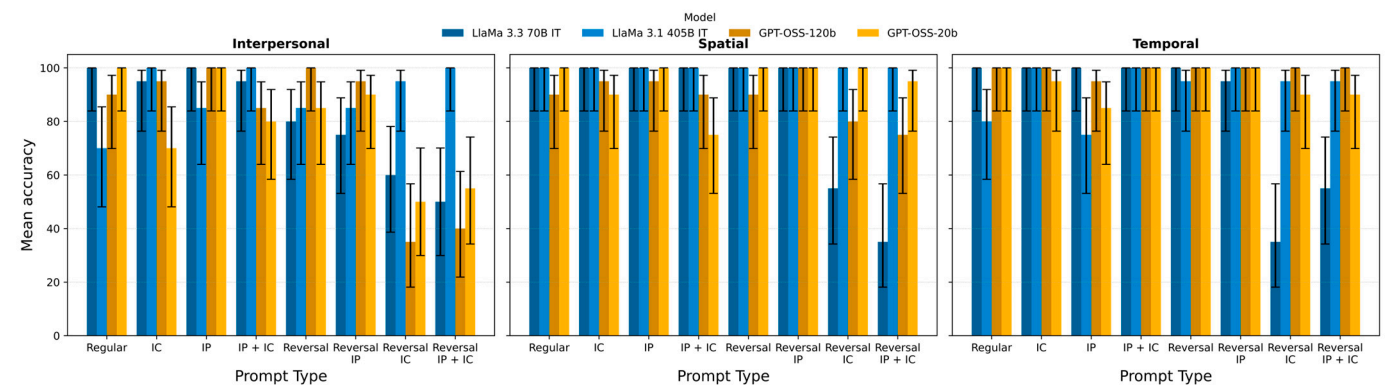


Figure A8. Replication study performance on different variations in deictic relations trials, grouped by relation involved. Error bars are 95% Wilson confidence intervals. IC = incorrect Conclusion, IP = Irrelevant Premise.

Table A29. Model Accuracy (%) For Different Deictic Problem Variants in the Replication Study, Grouped by Relations Involved.

Block	Problem Variant	LLaMa 3.3 70B	LLaMa 3.3 405B	GPT OSS 20B	GPT OSS 120B
Interpersonal	Regular	100.00	70.00	100.00	90.00
	Incorrect Conclusion	95.00	100.00	70.00	95.00
	Irrelevant Premise	100.00	85.00	100.00	100.00
	Irrelevant Incorrect	95.00	100.00	80.00	85.00
	Reversal	80.00	85.00	85.00	100.00
	Reversal Incorrect	60.00	95.00	50.00	35.00
	Reversal + Irrelevant	75.00	85.00	90.00	95.00
	Reversal Inc. + Irrelevant	50.00	100.00	55.00	40.00
Temporal	Regular	100.00	80.00	100.00	100.00
	Incorrect Conclusion	100.00	100.00	95.00	100.00
	Irrelevant Premise	100.00	75.00	85.00	95.00
	Irrelevant Incorrect	100.00	100.00	100.00	100.00
	Reversal	100.00	95.00	100.00	100.00
	Reversal Incorrect	35.00	95.00	90.00	100.00
	Reversal + Irrelevant	95.00	100.00	100.00	100.00
	Reversal Inc. + Irrelevant	55.00	95.00	90.00	100.00
Spatial	Regular	100.00	100.00	100.00	90.00
	Incorrect Conclusion	100.00	100.00	90.00	95.00
	Irrelevant Premise	100.00	100.00	100.00	95.00
	Irrelevant Incorrect	100.00	100.00	75.00	90.00
	Reversal	100.00	100.00	100.00	90.00
	Reversal Incorrect	55.00	100.00	100.00	80.00
	Reversal + Irrelevant	100.00	100.00	100.00	100.00
	Reversal Inc. + Irrelevant	35.00	100.00	95.00	75.00

Appendix D.3. Analogy

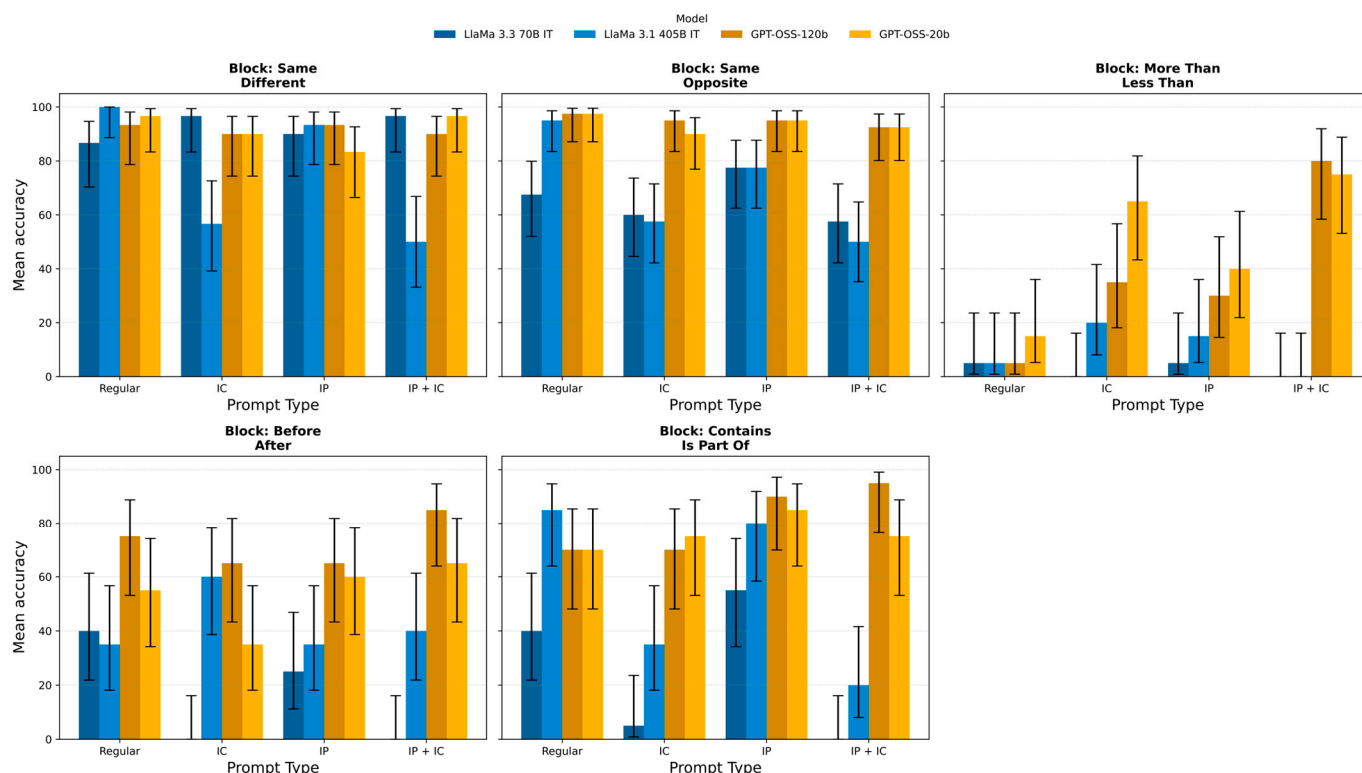


Figure A9. Replication study performance on different variations in analogy problems, grouped by relations involved. Error bars are 95% Wilson confidence intervals. IC = incorrect Conclusion, IP = Irrelevant Premise.

Table A30. Model Accuracy (%) For Different Analogy Problem Variants in the Replication Study,, Grouped by Relations Involved.

Block	Problem Variant	LLaMa 3.3 70B	LLaMa 3.3 405B	GPT OSS 20B	GPT OSS 120B
Same and Different	Regular	86.67	100.00	96.67	93.33
	Incorrect Conclusion	99.67	56.67	90.00	90.00
	Irrelevant Premise	90.00	93.33	83.33	93.33
	Irrelevant Incorrect	96.67	50.00	96.67	90.00
Same and Opposite	Regular	67.50	95.00	97.50	97.50
	Incorrect Conclusion	60.00	57.50	90.00	95.00
	Irrelevant Premise	77.50	77.50	95.00	95.00
	Irrelevant Incorrect	57.50	50.00	92.50	92.50
More Than and Less Than	Regular	5.00	5.00	15.00	5.00
	Incorrect Conclusion	0.00	20.00	65.00	35.00
	Irrelevant Premise	5.00	15.00	40.00	30.00
	Irrelevant Incorrect	0.00	0.00	75.00	80.00
Before and After	Regular	40.00	35.00	55.00	75.00
	Incorrect Conclusion	0.00	60.00	35.00	65.00
	Irrelevant Premise	25.00	35.00	60.00	65.00
	Irrelevant Incorrect	0.00	40.00	65.00	85.00
Contains and Is Part Of	Regular	40.00	85.00	70.00	70.00
	Incorrect Conclusion	5.00	35.00	75.00	70.00
	Irrelevant Premise	55.00	80.00	85.00	90.00
	Irrelevant Incorrect	0.00	20.00	75.00	95.00

Appendix D.4. Transformation of Function

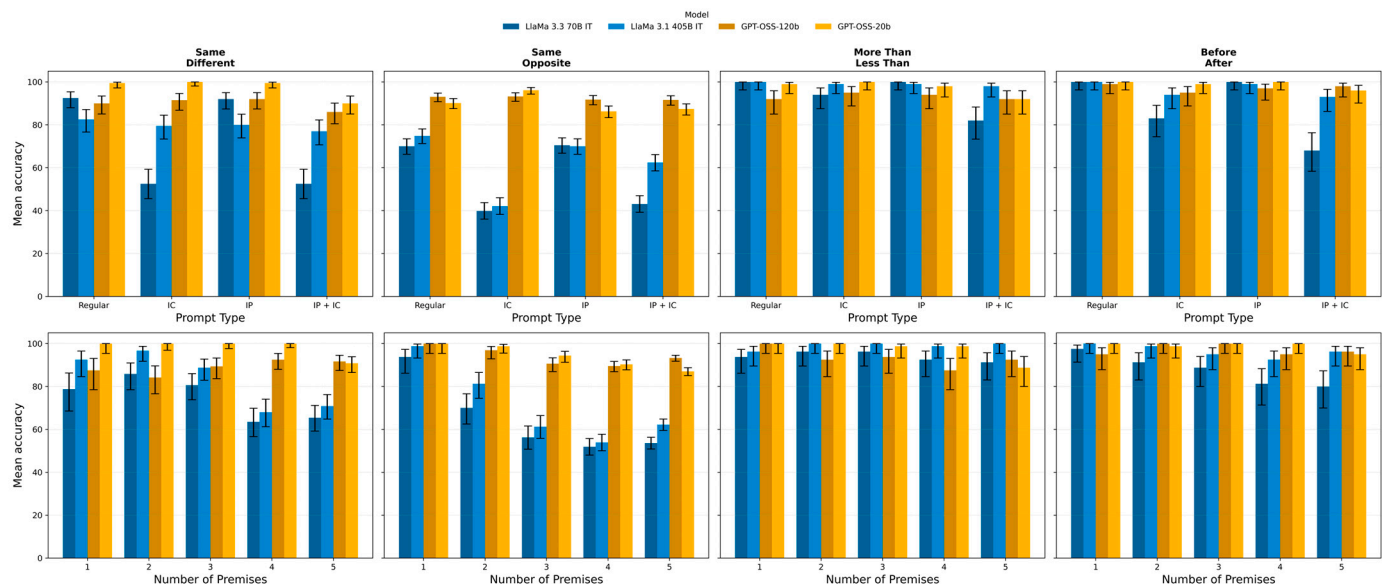


Figure A10. Replication study performance on transformation of function trials for different problem variation and number of premises, grouped by relations involved. Error bars are 95% Wilson confidence intervals. IC = incorrect Conclusion, IP = Irrelevant Premise.

Table A31. Model Accuracy (%) For Different Transformation of Function Problem Variants in the Replication Study, Grouped by Relations Involved.

Block	Problem Variant	LlaMa 3.3 70B	LlaMa 3.3 405B	GPT OSS 20B	GPT OSS 120B
Same and Different	Regular	92.50	82.50	99.50	90.00
	Incorrect Conclusion	52.50	79.50	100.00	91.50
	Irrelevant Premise	92.00	80.00	99.50	92.00
	Irrelevant Incorrect	52.50	77.00	90.00	86.00
Same and Opposite	Regular	70.00	74.84	90.16	93.07
	Incorrect Conclusion	39.84	42.10	96.13	93.23
	Irrelevant Premise	70.48	70.00	86.29	91.77
	Irrelevant Incorrect	43.07	62.42	87.42	91.61
More Than and Less Than	Regular	100.00	100.00	99.00	92.00
	Incorrect Conclusion	94.00	99.00	100.00	95.00
	Irrelevant Premise	100.00	99.00	98.00	94.00
	Irrelevant Incorrect	82.00	98.00	92.00	92.00
Before and After	Regular	100.00	100.00	100.00	99.00
	Incorrect Conclusion	83.00	94.00	99.00	95.00
	Irrelevant Premise	100.00	99.00	100.00	97.00
	Irrelevant Incorrect	68.00	93.00	96.00	98.00

Table A32. Model Accuracy (%) for Transformation of Function Problems with Increasing Number of Premises Grouped by Relations Involved.

Block	Number of Premises	LlaMa 3.3 70B	LlaMa 3.3 405B	GPT OSS 20B	GPT OSS 120B
Same and Different	One	78.75	92.50	100.00	87.50
	Two	85.83	69.67	100.00	84.17
	Three	80.63	88.75	100.00	89.38
	Four	63.50	0.0068	100.00	92.50
	Five	65.42	70.83	90.83	91.67

Same and Opposite	One	93.75	98.75	100.00	100.00
	Two	70.00	81.25	98.75	96.88
	Three	56.25	91.25	94.38	90.63
	Four	51.88	53.91	90.31	89.53
	Five	53.59	62.19	97.03	93.28
More Than and Less Than	One	93.75	96.25	100.00	100.00
	Two	96.25	100.00	100.00	92.50
	Three	96.25	100.00	98.75	93.75
	Four	92.50	98.75	98.75	87.50
	Five	91.25	100.00	88.75	92.50
Before and After	One	97.50	100.00	100.00	95.00
	Two	91.25	98.75	98.75	100.00
	Three	88.75	95.00	100.00	100.00
	Four	81.25	92.50	100.00	95.00
	Five	80.00	96.25	95.00	96.25

Notes

- More information, as well as the illustrations of the derivation tables and scripts to create or implement them can be found on the OSF: https://osf.io/bjxqg/overview?view_only=76aea06fae764230b8fa5dea3a5c4728 (accessed on 29 October 2025).
- https://osf.io/78u36/overview?view_only=8f2df70d8ff845e9ad393d407f4c27c1 (accessed on 29 October 2025).
- The preregistered (https://osf.io/78u36/overview?view_only=8f2df70d8ff845e9ad393d407f4c27c1, accessed on 29 October 2025) model sample also contained two models from Google’s Gemma 3 model family and from DeepSeek AI. However, for technical (we could not test Gemma models via Together AI’s API-service), practical (querying the DeepSeek models took a very long time compared to other models) and financial (the cost of running DeepSeek models) reasons, we did not test these models.
- <https://docs.together.ai/reference/chat-completions-1> (accessed on 29 October 2025).

References

- Alexander, P. A., Dumas, D., Grossnickle, E. M., List, A., & Firetto, C. M. (2016). Measuring relational reasoning. *The Journal of Experimental Education*, 84(1), 119–151. <https://doi.org/10.1080/00220973.2014.963216>.
- Ando, R., Morishita, T., Abe, H., Mineshima, K., & Okada, M. (2023). Evaluating large language models with NeuBAROCO: Syllogistic reasoning ability and human-like biases. *arXiv*, arXiv:2306.12567.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March 3–10). *On the dangers of stochastic parrots: Can language models be too big?* 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623), Virtual.
- Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., ... Anca Dragan, A. (2024). Managing extreme AI risks amid rapid progress. *Science*, 384(6698), 842–845. <https://doi.org/10.1126/science.adn0117>.
- Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., & Evans, O. (2023). The Reversal Curse: LLMs trained on “A is B” fail to learn “B is A”. *arXiv*, arXiv:2309.12288.
- Bertolazzi, L., Gatt, A., & Bernardi, R. (2024). A systematic analysis of large language models as soft reasoners: The case of syllogistic inferences. *arXiv*, arXiv:2406.11341.
- Birney, D. P., Halford, G. S., & Andrews, G. (2006). Measuring the influence of complexity on relational reasoning: The development of the latin square task. *Educational and Psychological Measurement*, 66(1), 146–171.
- Borji, A. (2023). A categorical archive of chatgpt failures. *arXiv*, arXiv:2302.03494.
- Bowman, S. R., & Dahl, G. E. (2021). What will it take to fix benchmarking in natural language understanding? *arXiv*, arXiv:2104.02145.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & Agarwal, S. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Cassidy, S., Roche, B., & Hayes, S. C. (2011). A relational frame training intervention to raise intelligence quotients: A pilot study. *The Psychological Record*, 61(2), 173–198. <https://doi.org/10.1007/bf03395755>.
- Chollet, F. (2019). On the measure of intelligence. *arXiv*, arXiv:1911.01547.

- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., & Schuh, P. (2023). Palm: Scaling language modelling with pathways. *Journal of Machine Learning Research*, 24(240), 1–113.
- Colbert, D., Dobutowitsch, M., Roche, B., & Brophy, C. (2017). The proxy-measurement of intelligence quotients using a relational skills abilities index. *Learning and Individual Differences*, 57, 114–122. <https://doi.org/10.1016/j.lindif.2017.03.010>.
- Colbert, D., Malone, A., Barrett, S., & Roche, B. (2020). The relational abilities index+: Initial validation of a functionally understood proxy measure for intelligence. *Perspectives on Behavior Science*, 43(1), 189–213. <https://doi.org/10.1007/s40614-019-00197-z>.
- Colbert, D., Tyndall, L., Roche, B., & Cassidy, S. (2018). Can SMART training really increase intelligence? A replication study. *Journal of Behavioral Education*, 27(4), 509–531. <https://doi.org/10.1007/s10864-018-9302-2>.
- Crockett, M., & Messeri, L. (2023). Should large language models replace human participants? *PsyArXiv preprint*. <https://doi.org/10.31234/osf.io/4zdx9>.
- Cummins, J. (2023). On the measurement of relational responding. *Journal of Contextual Behavioral Science*, 30, 155–168. <https://doi.org/10.1016/j.jcbs.2023.10.003>.
- Cummins, J., Nevejans, M., Colbert, D., & De Houwer, J. (2023). On the structure of relational responding. *Journal of Contextual Behavioral Science*, 27, 16–25. <https://doi.org/10.1016/j.jcbs.2022.11.006>.
- Dixon, M. R., Yi, Z., & Chastain, A. N. (2022). PEAK relational training system. In *Handbook of applied behavior analysis interventions for autism: Integrating research into practice* (pp. 341–360). Springer International Publishing. https://doi.org/10.1007/978-3-030-96478-8_18.
- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., & Ganapathy, R. (2024). The llama 3 herd of models. *arXiv*, <https://doi.org/10.48550/arXiv.2407.21783>.
- Eisape, T., Tessler, M. H., Dasgupta, I., Sha, F., van Steenkiste, S., & Linzen, T. (2023). A systematic comparison of syllogistic reasoning in humans and language models. *arXiv*, arXiv:2311.00445.
- Finn, M., & De Houwer, J. (2021). The selective action of Cfunc control. *Journal of the Experimental Analysis of Behavior*, 116(3), 314–331. <https://doi.org/10.1002/jeab.717>.
- Finn, M., Raemaekers, M., & De Houwer, J. (2023). Instructing via relations: Function transformations of response and consequence functions of upcoming contingencies. *Journal of Contextual Behavioral Science*, 30, 203–209. <https://doi.org/10.1016/j.jcbs.2023.11.005>.
- Frank, M. C. (2023a). Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology*, 2(8), 451–452. <https://doi.org/10.1038/s44159-023-00211-x>.
- Frank, M. C. (2023b). Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, 27(11), 990–992. <https://doi.org/10.1016/j.tics.2023.08.007>.
- Gemma Team, Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., & Rouillard, L. (2025). Gemma 3 technical report. *arXiv*, arXiv:2503.19786.
- Gentner, D., & Smith, A. L. (2013). Analogical learning and reasoning. In D. Reisberg (Ed.), *The oxford handbook of cognitive psychology*. Oxford Library of Psychology. <https://doi.org/10.1093/oxfordhb/9780195376746.013.0042>.
- Goel, V., & Dolan, R. J. (2001). Functional neuroanatomy of three-term relational reasoning. *Neuropsychologia*, 39(9), 901–909. [https://doi.org/10.1016/s0028-3932\(01\)00024-0](https://doi.org/10.1016/s0028-3932(01)00024-0).
- Goodwin, G. P., & Johnson-Laird, P. N. (2005). Reasoning about relations. *Psychological Review*, 112(2), 468. <https://doi.org/10.1037/0033-295x.112.2.468>.
- Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., & Li, E. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv*, arXiv:2501.12948.
- Halford, G. S., Wilson, W. H., & Phillips, S. (2010). Relational knowledge: The foundation of higher cognition. *Trends in Cognitive Sciences*, 14(11), 497–505. <https://doi.org/10.1016/j.tics.2010.08.005>.
- Hayes, S. C., Barnes-Holmes, D., & Roche, B. (Eds.). (2001). *Relational frame theory: A post-Skinnerian account of human language and cognition*. Springer Science & Business Media.
- Holyoak, K. J. (2012). Analogy and relational reasoning. In *The oxford handbook of thinking and reasoning* (pp. 234–259). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199734689.013.0013>.
- Hughes, S., & Barnes-Holmes, D. (2015). Relational frame theory: The basic account. In *The Wiley handbook of contextual behavioral science* (pp. 129–178). Wiley Online Library. <https://doi.org/10.1002/9781118489857.ch9>.
- Hummel, J. E., & Holyoak, K. J. (2001). A process model of human transitive inference. *Spatial Schemas and Abstract Thought*, 279–306. <https://doi.org/10.7551/mitpress/6392.003.0015>.

- Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., & Antoniak, S. (2024). Mixtral of experts. *arXiv*, arXiv:2401.04088.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *arXiv*, arXiv:2001.08361.
- Kaufman, A. S., & Lichtenberger, E. O. (2006). *Assessing adolescent and adult intelligence* (3rd ed.). John Wiley & Sons, Inc.
- Legg, S., & Hutter, M. (2007). A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and applications*, 157, 17.
- Lewis, M., & Mitchell, M. (2024). Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. *arXiv*, arXiv:2402.08955.
- Li, L. H., Hessel, J., Yu, Y., Ren, X., Chang, K. W., & Choi, Y. (2023). Symbolic chain-of-thought distillation: Small models can also “think” step-by-step. *arXiv*, arXiv:2306.14050.
- Lin, Z. (2025). Six fallacies in substituting large language models for human participants. *Advances in Methods and Practices in Psychological Science*, 8(3), 25152459251357566. <https://doi.org/10.1177/25152459251357566>.
- Lionello-DeNolf, K. M. (2009). The search for symmetry: 25 years in review. *Learning & Behavior*, 37(2), 188–203. <https://doi.org/10.3758/lb.37.2.188>.
- Lionello-DeNolf, K. M. (2021). An update on the search for symmetry in nonhumans. *Journal of the Experimental Analysis of Behavior*, 115(1), 309–325. <https://doi.org/10.1002/jeab.647>.
- May, R. J., Tyndall, I., McTiernan, A., Roderique-Davies, G., & McLoughlin, S. (2022). The impact of the SMART program on cognitive and academic skills: A systematic review and meta-analysis. *British Journal of Educational Technology*, 53(5), 1244–1261. <https://doi.org/10.1111/bjet.13192>.
- McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., & Griffiths, T. L. (2024). Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41), e2322420121.
- McHugh, L., Barnes-Holmes, Y., & Barnes-Holmes, D. (2004). Perspective-taking as relational responding: A developmental profile. *The Psychological Record*, 54(1), 115–144. <https://doi.org/10.1007/bf03395465>.
- McLoughlin, S., Tyndall, I., & Pereira, A. (2020). Convergence of multiple fields on a relational reasoning approach to cognition. *Intelligence*, 83, 101491. <https://doi.org/10.1016/j.intell.2020.101491>.
- Meta AI (2024, July 23). *Introducing Llama 3.1: Our most capable models to date*. Available online: <https://ai.meta.com/blog/meta-llama-3-1/> (accessed on 29 October 2025).
- Morris, M. R., Sohl-dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., & Legg, S. (2023). Levels of AGI for operationalizing progress on the path to AGI. *arXiv*, arXiv:2311.02462.
- Open AI (2025, August 5). *Introducing gpt-oss*. Available online: <https://openai.com/index/introducing-gpt-oss/> (accessed on 29 October 2025).
- Open AI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., ... Bello, I. (2023). Gpt-4 technical report. *arXiv*, arXiv:2303.08774.
- Open AI, Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., Iftimie, A., Karpenko, A., Passos, A. T., Neitz, A., Prokofiev, A., Wei, A., Tam, A., Bennett, A., ... Kumar, A. (2024). Openai o1 system card. *arXiv*, arXiv:2412.16720.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., & Schulman, J. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
- Ozeki, K., Ando, R., Morishita, T., Abe, H., Mineshima, K., & Okada, M. (2024). Exploring reasoning biases in large language models through syllogism: Insights from the NeuBAROCO dataset. *arXiv*, arXiv:2408.04403.
- Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin’s mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, 31(2), 109–130. <https://doi.org/10.1017/s0140525x08003543>.
- Premack, D. (1983). The codes of man and beasts. *Behavioral and Brain Sciences*, 6(1), 125–136. <https://doi.org/10.1017/s0140525x00015077>.
- Raemaekers, M. (in preparation). Open-source tools for relational network derivation, visualization and task generation.
- Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *arXiv*, Available online: <https://arxiv.org/abs/2305.18290> (accessed on 29 October 2025).

- Raven, J., & Raven, J. (2003). Raven progressive matrices. In R. S. McCallum (Ed.), *Handbook of nonverbal assessment* (pp. 223–237). Kluwer Academic/Plenum Publishers. https://doi.org/10.1007/978-1-4615-0153-4_11.
- Shiffrin, R., & Mitchell, M. (2023). Probing the psychology of AI models. *Proceedings of the National Academy of Sciences*, 120(10), e2300963120. <https://doi.org/10.1073/pnas.2300963120>.
- Sourati, Z., Ilievski, F., Sommerauer, P., & Jiang, Y. (2024). Arn: Analogical reasoning on narratives. *Transactions of the Association for Computational Linguistics*, 12, 1063–1086. https://doi.org/10.1162/tacl_a_00688.
- Srivastava, A., Kleyko, D., & Wu, Z. (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv*. Available online: <https://arxiv.org/abs/2206.04615> (accessed on 29 October 2025).
- Sternberg, R. J., & Detterman, D. K. (1987). What is intelligence? Contemporary viewpoints on its nature and definition. *The American Journal of Psychology*, 100(1), 141. <https://doi.org/10.2307/1422652>.
- Todd, J. A. M., Andrews, G., & Conlon, E. G. (2019). Relational thinking in later adulthood. *Psychology and Aging*, 34(4), 486. <https://doi.org/10.1037/pag0000346>.
- Wang, P. (2019). On defining artificial intelligence. *Journal of artificial general intelligence*, 10(2), 1–37. <https://doi.org/10.2478/jagi-2019-0002>.
- Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9), 1526–1541. <https://doi.org/10.1038/s41562-023-01659-w>.
- Wechsler, D. (2008). *Wechsler adult intelligence scale-fourth edition (WAIS-IV)* [Database record]. APA PsycTests. <https://doi.org/10.1037/t15169-000>.
- Wu, W., & Deng, W. (2025, April 6–11). *Transitive Inference in Large Language Models and Prompting Intervention*. ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (pp. 1–5). Hyderabad, India. <https://doi.org/10.1109/ICASSP49660.2025.10889585>.
- Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., Kim, N., Andreas, J., & Kim, Y. (2024). Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (Volume 1: Long Papers)* (pp. 1819–1862). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.naacl-long.102>.
- Zador, A., Escola, S., Richards, B., Ölveczky, B., Bengio, Y., Boahen, K., Botvinick, M., Chklovskii, D., Churchland, A., Clopath, C., DiCarlo, J., Ganguli, S., Hawkins, J., Körding, K., Koulakov, A., LeCun, Y., Lillicrap, T., Marblestone, A., Olshausen, B., ... Pouget, A. (2023). Catalyzing next-generation artificial intelligence through neuroai. *Nature Communications*, 14(1), 1597. <https://doi.org/10.1038/s41467-023-37180-x>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.