

Beyond Accuracy: Evaluating the Reasoning Behavior of Large Language Models - A Survey

Philipp Mondorf & Barbara Plank

MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
Munich Center for Machine Learning (MCML), Munich, Germany
{p.mondorf, b.plank}@lmu.de

Abstract

Large language models (LLMs) have recently shown impressive performance on tasks involving reasoning, leading to a lively debate on whether these models possess reasoning capabilities similar to humans. However, despite these successes, the depth of LLMs’ reasoning abilities remains uncertain. This uncertainty partly stems from the predominant focus on task *performance*, measured through shallow accuracy metrics, rather than a thorough investigation of the models’ reasoning *behavior*. This paper seeks to address this gap by providing a comprehensive review of studies that go beyond task accuracy, offering deeper insights into the models’ reasoning processes. Furthermore, we survey prevalent methodologies to evaluate the reasoning behavior of LLMs, emphasizing current trends and efforts towards more nuanced reasoning analyses. Our review suggests that LLMs tend to rely on surface-level patterns and correlations in their training data, rather than on sophisticated reasoning abilities. Additionally, we identify the need for further research that delineates the key differences between human and LLM-based reasoning. Through this survey, we aim to shed light on the complex reasoning processes within LLMs.

1 Introduction

“These models are castles in the air. They have no foundations whatsoever.”
— Jitendra Malik (2021)

Reasoning is an integral aspect of human intelligence and deliberate, rational thought (Holyoak & Morrison, 2005). It allows individuals to draw conclusions from available information and move beyond their current knowledge (Lohman & Lakin, 2011). As such, reasoning plays a fundamental role in problem-solving and decision-making, and has been a long-standing goal within the field of artificial intelligence (Robinson & Voronkov, 2001).

In recent years, large language models have demonstrated remarkable performance on tasks that require reasoning (Bubeck et al., 2023; Wei et al., 2022; Kojima et al., 2022). This has sparked a vigorous debate about the extent to which these models possess reasoning abilities similar to humans (Mitchell & Krakauer, 2023; Mitchell, 2023; Borji, 2023). While proponents argue that reasoning capabilities *emerge* with scale, referring to LLMs as *foundation* models (Bommasani et al., 2021; Kaplan et al., 2020), skeptics contend that the models’ performance primarily reflects their capacity to memorize the vast amount of data they have been trained on (Wu et al., 2024; Dziri et al., 2023; Razeghi et al., 2022; Zhang et al., 2023). Thus, the question arises: are these models simply “*castles in the air*” with “*no foundations whatsoever*,” as Malik (2021) once stated, or do they possess genuine reasoning capacities? One of the major challenges in this debate is the immense size, complexity, and closed-source nature of popular LLMs and their underlying training data. Moreover, the focus often lies on the models’ *performance* on downstream reasoning tasks (Fu et al., 2023; Liu et al., 2023), overshadowing in-depth analyses of their reasoning *behavior*. In this study, we seek to shed light on the ongoing debate by providing a comprehensive overview of research that goes

beyond task accuracy, offering more nuanced insights into the models’ reasoning processes. Specifically, we address the following research questions:

RQ1: How do current LLMs behave across diverse reasoning tasks?

RQ2: What are the prevalent evaluation methods employed to assess the reasoning behavior of large language models?

Other Surveys. While various surveys on reasoning within large language models have emerged in recent years (Huang & Chang, 2023; Yu et al., 2024; Sun et al., 2024; Yang et al., 2023; Qiao et al., 2023; Chu et al., 2023; Mialon et al., 2023; Liu et al., 2024; Ahn et al., 2024), these studies predominantly focus on techniques that seek to enhance or benchmark the *performance* of LLMs on downstream reasoning tasks. However, there seems to exist no review of work yet that assesses the reasoning *behavior* of LLMs, rather than their final task performance. Hence, the contributions of this survey are as follows: (i) To the best of our knowledge, we present the first comprehensive review of literature that evaluates LLM-based reasoning beyond task accuracy, offering deeper insights into the models’ reasoning *behavior*, and (ii) we suggest a taxonomy that categorizes prevalent evaluation methods designed to analyze the reasoning dynamics of LLMs.

2 Terminology

We begin this survey with an introduction to fundamental concepts and terminologies pertinent to reasoning in both humans and LLMs. The study of reasoning has captivated scholarly interest for centuries, with insights from diverse disciplines such as philosophy, logic, psychology, neuroscience and artificial intelligence (Holyoak & Morrison, 2005). Over the years, various definitions of reasoning have been proposed, with each discipline offering its unique perspective. In this survey, we define reasoning in the broadest sense (Leighton, 2003; Holyoak & Morrison, 2005; Byrne et al., 1993), and direct the interested reader to the work by Yu et al. (2024) and Sun et al. (2024) for more domain-specific definitions.

Definition 2.1 (Reasoning). *The process of drawing conclusions based on available information (usually a set of premises).*

It is imperative to note that reasoning is a fundamentally process-oriented activity, rather than a singular endpoint (Leighton, 2003; Johnson-Laird, 2006). Although this process often remains hidden in humans, manifesting predominantly through the final conclusions inferred, cognitive psychology leverages methodologies like “think aloud” protocols to unveil the cognitive mechanisms underpinning reasoning (Wolcott & Lobczowski, 2021; Van der Henst et al., 2002; Rips, 1989). Similarly, to understand the reasoning capabilities of large language models, it is crucial to consider not merely the end result, but the reasoning process itself. Thus, we differentiate between reasoning *behavior* and reasoning *performance*. Drawing from behavioral psychology, which views behavior as an organism’s response to stimuli (American Psychological Association, n.d.), we define reasoning behavior as follows:

Definition 2.2 (Reasoning Behavior). *The system’s computed response to a reasoning task (the stimulus), particularly its actions, expressions and underlying mechanisms exhibited during the reasoning process.*

Our working definition highlights the procedural aspects of reasoning, rather than its final outcome. Understanding a model’s reasoning behavior involves analyzing *how* it arrives at its conclusions. Conversely, reasoning *performance* is outcome-oriented. It is typically measured in terms of accuracy or the efficiency with which relevant conclusions are drawn (Barredo Arrieta et al., 2020). While performance can be helpful in evaluating a system’s capacities to tackle specific reasoning tasks, an analysis of reasoning *behavior* can yield deeper insights into the process itself, thereby offering a more comprehensive understanding.

2.1 A Categorization of Reasoning Tasks

Analogous to the assessment of human reasoning within the field of cognitive psychology (WOOD, 1969; Byrne et al., 1993; Dewey, 2022), the evaluation of reasoning capabilities in

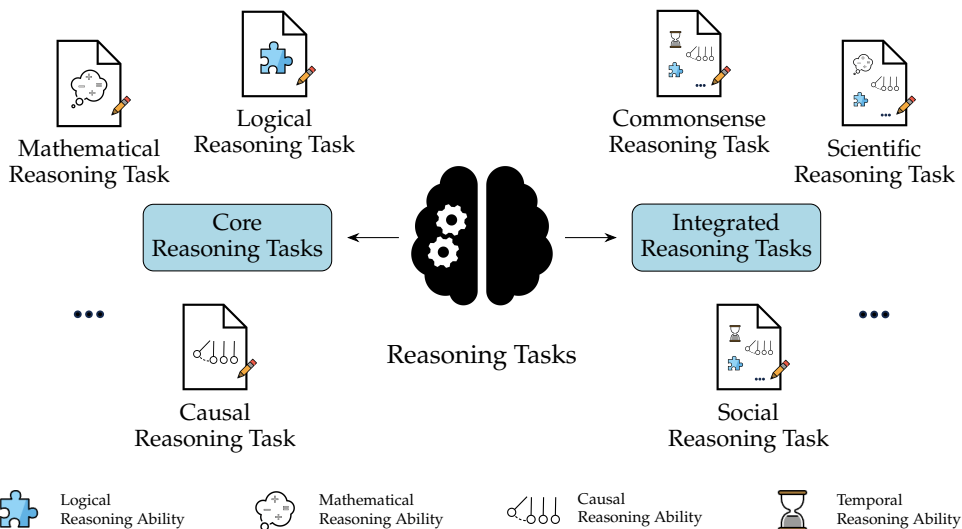


Figure 1: Schematic overview of the two types of reasoning tasks distinguished in this survey. *Core* reasoning tasks are designed to assess a particular reasoning ability within a given context. Conversely, *integrated* reasoning tasks involve the concurrent use of various reasoning skills. Tasks and abilities listed are not exhaustive.

large language models predominantly occurs through their engagement with designated reasoning tasks (Sun et al., 2024). These tasks are designed to elicit the system’s capability of drawing conclusions relevant to the problem at hand. In this survey, we distinguish between *core* and *integrated* reasoning tasks. *Core* reasoning tasks are designed to assess fundamental reasoning skills in an isolated manner. They typically aim to test a single type of reasoning, such as logical, mathematical or causal reasoning. Examples of such tasks may include syllogisms, basic arithmetic problems or structured causal-graph predictions. Conversely, *integrated* reasoning tasks require the concurrent use of various reasoning types, thereby assessing a combination of fundamental reasoning skills. Examples are commonsense or scientific reasoning tasks. Such problems often reflect the complex cognitive challenges humans encounter in everyday life and professional settings. A schematic representation distinguishing these task categories is provided in Figure 1.

For the purpose of this survey, our focus is confined to examining the behaviors of LLMs in the context of *core* reasoning tasks, specifically logical, mathematical, and causal reasoning tasks. We leave a review of the models’ reasoning behaviors within the context of *integrated* reasoning tasks to future work.

3 Reasoning Behavior of LLMs

This section reviews studies that extend beyond mere task accuracy, focusing instead on evaluating the reasoning *behavior* of large language models.¹ Through this review, we intend to address RQ1 and shed light on how these models currently behave across three *core* reasoning tasks: logical reasoning (Section 3.1), mathematical reasoning (Section 3.2), and causal reasoning (Section 3.3).

¹We consistently specify when a particular technique, other than standard prompting, is utilized. For a review of prompting approaches designed to improve the models’ reasoning performance, we recommend consulting prior surveys by Huang & Chang (2023), Yu et al. (2024), or Sun et al. (2024).

3.1 Behavior in Logical Reasoning Tasks

The study of logical reasoning evolves around the question of how individuals infer valid conclusions from a set of given premises within a structured framework of logical rules and principles (Holyoak & Morrison, 2005). Based on how conclusions are inferred, logical reasoning can be classified into three types: *deductive*, *inductive*, and *abductive* reasoning (Johnson-Laird, 2008; Flach & Kakas, 2000). While in *deductive* reasoning, conclusions necessarily follow from the premises' truth, *inductive* and *abductive* reasoning are considered defeasible, i.e. conclusions are at most probable, but never necessary (Koons, 2022). *Inductive* reasoning is concerned with deriving general conclusions from specific instances, whereas *abductive* reasoning entails formulating plausible hypotheses that explain the data observed. For additional details on the distinction between the three types of logical reasoning, we refer to prior surveys on reasoning performance within LLMs, such as the ones by Yu et al. (2024) or Sun et al. (2024).

3.1.1 Behavior in Deductive Reasoning Tasks

Extensive research has been dedicated to examining the reasoning behavior of LLMs in the context of deductive reasoning tasks. Various studies analyze the validity and consistency of logical proofs generated by LLMs within deductive reasoning problems. For instance, Saparov & He (2023) conduct a systematic assessment of the rationales produced by various GPT-3 iterations (Brown et al., 2020), through chain-of-thought (CoT) prompting (Wei et al., 2022). By parsing each step of the generated reasoning trace into first-order-logic, its *validity*, *atomicity*, and *utility*² is measured. Findings indicate that, in comparison to their smaller counterparts, larger models are more adept at generating both valid and atomic steps. However, the utility of these steps is often low, resulting in misleading steps from which the models struggle to recover. In a similar vein, Dziri et al. (2023) study rationales of LLMs by parsing them into computation graphs, where each node represents a partial solution of the multi-step reasoning process. Evaluations on Einstein's puzzle (Vassberg & Vassberg, 2009) indicate that GPT-3, ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI et al., 2024) solve the task by reducing multi-step reasoning into shortcut pattern matching. This shortcut behavior can yield correct answers when the model has been exposed to similar patterns during training, but falls short in generalizing to novel or more complex instances. Furthermore, theoretical findings suggest that due to the models' autoregressive nature, mistakes in early stages of the reasoning process can lead to compounding errors that exponentially diminish the likelihood of arriving at the correct solution in subsequent steps.

Further studies highlight that LLMs tend to rely on superficial statistical features, rather than engaging in systematic reasoning. Chen et al. (2024b) illustrate that the premise order markedly influences the LLMs' behavior in propositional reasoning tasks. Specifically, when premises are presented in an order that does not align with the ground-truth proof, models such as ChatGPT, GPT-4, PaLM 2-L (Anil et al., 2023) and Gemini Pro (Team et al., 2023) encounter significant difficulties within their reasoning, even though such an ordering does not change the underlying logic. Zhang et al. (2023) show that an over-reliance on statistical features in the training data can hinder a model's reasoning and generalization capacity. By eliminating certain statistical cues from its training dataset, BERT (Devlin et al., 2019) demonstrates enhanced generalization capabilities in propositional logic. Similarly, Pirozelli et al. (2023) indicate that various fine-tuned language models show difficulties in transferring their logical reasoning ability when cross-probed on unseen deductive reasoning tasks.

Additional research points to the difficulties of LLMs in understanding specific logical operators. Sanyal et al. (2022) show that GPT-3, RoBERTa (Liu et al., 2019), and models from the T5 series (Raffel et al., 2020) exhibit deficiencies in comprehending logical negations, often failing to correctly deduce the implications of negated statements and rules. Similar findings have been reported by Truong et al. (2023), who demonstrate that models lack a sensitivity to negations within a broad range of natural language inference tasks. Wan

²*Validity* denotes whether the proof step logically follows from preceding steps, *atomicity* reflects whether it can be proven with exactly one application of a deduction rule, and *utility* measures its direct contribution towards the derivation of the final conclusion.

et al. (2024) comprehensively assess a suite of formal reasoning scenarios, subjecting GPT-3, ChatGPT, GPT-4, Bard (PaLM 2), Vicuna (Vicuna, 2023) and Guanaco (Dettmers et al., 2023) to minimum functionality tests (MFTs), serving as logical reasoning unit tests that gauge the models’ inherent logic comprehension. Their analysis uncovers a common difficulty among models in identifying logical fallacies. In addition, GPT-4 appears to struggle with De Morgan’s Laws, which relate conjunctions and disjunctions through logical negation.

A growing body of research investigates the extent to which LLMs exhibit human-like reasoning patterns, particularly in deductive reasoning tasks. For example, Eisape et al. (2024) explore the parallels between human reasoning and that of LLMs in syllogisms. Their findings suggest that LLMs, much like their human counterparts, are susceptible to common logical fallacies and cognitive biases. Similarly, Dasgupta et al. (2022) find that LLMs, akin to humans, display content effects, indicating that the problem’s semantic content can significantly influence the models’ reasoning behavior. Further research corroborates the manifestation of human reasoning patterns in LLMs, as outlined in Appendix A.

Mechanistic Evaluation. Additional research evaluates the reasoning behavior of LLMs by inspecting the models’ internal mechanisms during the reasoning process. For example, Hou et al. (2023) analyze the models’ attention patterns, specifically those of GPT-2 (Radford et al., 2019) and LLaMA (Touvron et al., 2023), to uncover if and how these models perform multi-step reasoning internally. Findings indicate a structured, step-wise information processing within the models. Furthermore, layer-wise probing reveals that LLMs tend to focus on identifying relevant information from the task in lower layers, transitioning to more intricate reasoning processes in higher layers. Pirozelli et al. (2023) similarly probe individual layers of a fine-tuned RoBERTa-large model, corroborating the pivotal role of higher layers in the reasoning process. Dutta et al. (2024) present an in-depth investigation into the internal mechanisms of LLaMA 2-7B (Touvron et al., 2023) when instructed via chain-of-thought prompting. Findings indicate a notable functional rift in the model’s middle layers, where token representations in the initial half are biased towards pre-training priors, and an abrupt shift to in-context priors occurs in the latter half. As opposed to the findings by Hou et al. (2023), the study suggests that LLaMA 2-7B employs multiple, concurrent pathways, instead of following a singular path of reasoning.

3.1.2 Behavior in Inductive Reasoning Tasks

In contrast to the well-examined domain of deductive reasoning, the reasoning behavior of LLMs in inductive reasoning tasks remains comparatively underexplored. Nevertheless, research exists that seeks to evaluate the capability of LLMs to infer general conclusions from specific examples. For instance, Yang et al. (2024) investigate how LLMs, such as GPT-J (Wang & Komatsuzaki, 2021) and LLaMA 7B, induce general rules from given facts. While in principle the models seem able to infer general rules from the data provided, challenges arise in ensuring that the rules are consistent with the premises, extend beyond the given information, align with real-world knowledge, and are pertinent to the given task. Similarly, Qiu et al. (2024) find that LLMs such as GPT-3.5, GPT-4, Claude 2 (Anthropic, 2023), and LLaMA 2-70B are capable of inferring rules from given data. However, the models frequently err in the application of these rules, highlighting a gap between their ability to generate and apply rules. Moreover, the rules derived often diverge significantly from those humans might produce, exhibiting a tendency towards verbosity and an inability to concentrate on the fundamental patterns for generalization. In addition, the study finds that LLMs display a pronounced sensitivity to alterations of the task descriptions. Han et al. (2024) examine the reasoning behaviors of GPT-3.5 and GPT-4 in property induction tasks, where properties common among different categories need to be induced. Findings suggest that GPT-4’s behavior closely aligns with human judgments. However, challenges arise when the models need to handle premise non-monotonicity, a scenario in which adding more information to an argument can actually decrease its likelihood. For instance, Han et al. (2024) illustrate that the three-premise argument {crow, peacock, rabbit} → bird is considered weaker by humans than the two-premise argument {crow, peacock} → bird, a line of reasoning that both GPT-3.5 and GPT-4 seem to struggle with.

3.1.3 Behavior in Abductive Reasoning Tasks

Some efforts have been made to evaluate the behavior of LLMs in abductive reasoning tasks. [Collins et al. \(2022\)](#) ask both humans and GPT-3 to generate plausible explanations for unexpected counterfactual scenarios, for instance why a window did not break despite being struck by a rock. Analyses indicate that GPT-3's ability to generate plausible and coherent explanations for scenarios that require reasoning beyond established patterns is limited. In scenarios demanding inventive, coherent, and context-sensitive responses, especially when the usage of common nouns is restricted, human performance distinctly surpasses that of GPT-3, underscoring a pronounced deficiency in the model's ability to reason beyond its training distribution. Further, [Rudinger et al. \(2020\)](#) reveal that language models such as GPT-2, BART, and various T5 variants exhibit difficulties in identifying or generating arguments that either weaken or strengthen a given hypothesis. Even after fine-tuning, models tend to struggle with the task, often contradicting themselves by producing identical statements to strengthen and weaken the same premise-hypothesis pair. [Xu et al. \(2023\)](#) explore the behavior of GPT-3.5, ChatGPT, and PaLM 2 on various abductive reasoning tasks, paying particular attention to the models' rationales and errors manifested during reasoning. The paper highlights a tendency for LLMs to incorporate redundant information in their explanations and to generate content not grounded in the context, resulting in hallucinations. In comparison to deductive reasoning scenarios, findings suggest that models seem to particularly struggle with multi-hop reasoning in abductive reasoning tasks.

3.2 Behavior in Mathematical Reasoning Tasks

Mathematical reasoning encompasses the structured process of arriving at conclusions based on established mathematical principles and logical deduction ([Horsten, 2023](#)). Several studies explore the behavior of LLMs in the context mathematical reasoning tasks, especially in arithmetic and math word problems (MWP). For instance, [Srivastava et al. \(2024\)](#) evaluate a set of LLMs on a *functional* variant of the MATH dataset ([Hendrycks et al., 2021](#)), where the underlying mathematical principles of each problem are captured rather than a static problem formulation. This allows for evaluating models on different dataset *snapshots*, each comprising unique problem formulations but the same underlying reasoning process. Results reveal inconsistencies across varying snapshots, indicating a tendency of models to rely on memorization rather than reasoning. [Razeghi et al. \(2022\)](#) further highlight the impact of how a problem is formulated, noting a marked correlation between the frequency of terms in the models' pre-training data and their ability to solve arithmetic tasks. Similarly, [Wu et al. \(2024\)](#) observe that models struggle with two-digit additions expressed in mathematical bases that are less represented in the models' training data.

Further studies underline the models' susceptibility to perturbations of the reasoning task. [Shi et al. \(2023\)](#) find that LLMs such as code-davinci-002 and GPT-3.5 can be distracted by context irrelevant to the MWP solution, especially when the irrelevant context shares lexical similarities with the original problem formulation. In a similar vein, [Stolfo et al. \(2023\)](#) report that LLMs are sensitive to interventions on MWPs, such as changing numerical values or altering the textual framing of the problem.

Other studies investigate whether human-like reasoning behavior in mathematical tasks manifests in LLMs. [Hagendorff et al. \(2023\)](#) show that, akin to humans, GPT-3 variants tend to offer intuitive yet incorrect answers to cognitive reflection tests ([Frederick, 2005](#)). Notably, GPT-3.5 and GPT-4 provide more deliberate responses, outperforming humans in avoiding intuitive errors. At the same time, [McKenzie et al. \(2023\)](#) demonstrate a form of goal misgeneralization, where models, tasked with rounding numbers to a specific number of significant digits, consistently round based on the number of decimal places. This reflects a cognitive bias known as attribute substitution ([Morewedge & Kahneman, 2010](#)), where a more challenging task is replaced with a simpler, related task. In a mechanistic evaluation, [Chen et al. \(2024a\)](#) conduct a layer-wise analysis of LLaMA's mathematical reasoning capabilities, discovering that higher layers exhibit superior mathematical problem-solving abilities, while lower layers seem to lack basic arithmetic and factual knowledge.

3.3 Behavior in Causal Reasoning Tasks

Causal reasoning is the process of discerning the cause-and-effect relationships that govern the dynamics of our environment (Sloman, 2005). Extending beyond correlation, it offers a more nuanced understanding of how changes in one variable can bring about changes in another (Pearl, 2009). For a comprehensive introduction to causal reasoning, we strongly recommend the work of Pearl (2009). While research on the behavior of LLMs in causal reasoning tasks is still in its early stages, the field is receiving growing attention. Various studies to date suggest that LLMs are capable of reciting causal facts from their training data, but lack an intrinsic ability to comprehend or construct causal relationships. For instance, Zečević et al. (2023) indicate that while LLMs, including GPT-3, Meta AI’s Opt, and AlephAlpha’s Luminous (AlephAlpha, 2022), demonstrate some proficiency in causal reasoning tasks that align with causal facts seen during training, they exhibit limited ability to accurately discern and apply causal relationships in scenarios that demand more than associative recalls from their training data. Jin et al. (2023) probe the models’ behavior across three levels of causation: (1) the *associational*, (2) the *interventional*, and (3) the *counterfactual*, as outlined by Pearl & Mackenzie (2018)’s *Ladder of Causation*. Similarly, their findings indicate that LLMs are more adept at answering *associational* queries than tackling *interventional* or *counterfactual* tasks. Models seem to particularly struggle with causal relationships that deviate from commonsense or are unlikely to be part of their training corpora. Jin et al. (2024) evaluate the ability of LLMs to infer causation from statements that describe correlations between variables. Analyses reveal significant challenges in solving the given task across seventeen LLMs. Although fine-tuning offers substantial improvements, models still fail to generalize in out-of-distribution scenarios. Kosoy et al. (2023)’s evaluation of GPT-3 and PaLM on the *blicket detector* task (Gopnik & Sobel, 2000), where models need to infer which objects cause a light to switch on, further highlights challenges in causal reasoning with LLMs. While models can identify the correct causal structure when a set of causal hypotheses is provided, they struggle in conditions where the hypotheses are not given, underscoring a limitation in their ability to infer causal relationships from limited data.

Behavior in Counterfactual Reasoning Tasks. Positioned at the last level of Pearl & Mackenzie (2018)’s *Ladder of Causation*, counterfactual tasks assess the models’ ability to reason about hypothetical scenarios. Research into the behavior of LLMs in counterfactual scenarios reveals notable challenges in current models. Studies, like those conducted by Frohberg & Binder (2022), focus on the ability of LLMs to predict outcomes in hypothetical setups, identifying a significant gap between the capabilities of humans and LLMs, particularly in highly unrealistic scenarios. Li et al. (2023) further illustrate that while models like GPT-3 can produce outcomes that seem to align with counterfactual propositions, these models heavily rely on simple lexical cues within the given context, rather than demonstrating a profound grasp of the scenarios’ hypothetical essence. Yu et al. (2023) evaluate LLMs on questions embedded in counterfactual presuppositions, pushing the models beyond simple fact retrieval. Their findings suggest that “closed-book” models such as GPT-3 and code-davinci-002 may fabricate facts or base their responses on flawed premises when answering counterfactual questions. Huang et al. (2024) explore a different angle by asking models to modify a given piece of argumentative text so that it upholds a specified logical relationship under new premises. Findings indicate that while LLMs like GPT-3.5 and GPT-4 demonstrate a capacity for solving such tasks, they struggle with modifying arguments such that a new premise makes the argument less likely to be true, an observation also made by Rudinger et al. (2020) in the context of abductive reasoning.

Summary. Regarding RQ1, our review suggests that the reasoning behaviors of LLMs are nuanced and diverse, yet a significant trend emerges: while current LLMs demonstrate proficiency in reasoning problems that align with their training distribution, they frequently encounter substantial conceptual difficulties in out-of-distribution scenarios. Notably, slight alterations in the task context can markedly impair the models’ reasoning capabilities. Multi-step reasoning is often reduced to shortcut pattern matching, and fundamental conceptual errors highlight the models’ deficiency in understanding basic principles of logic, mathematics, and causal reasoning, particularly in counterfactual setups.

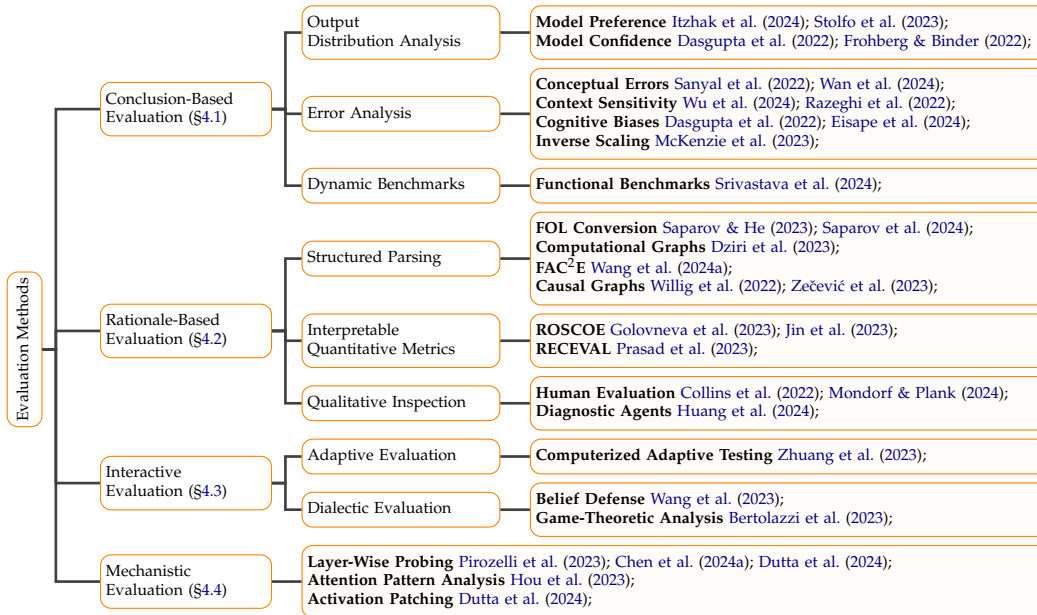


Figure 2: A taxonomy of evaluation methods to analyze the reasoning behaviors of LLMs. Only representative approaches for each method are listed.

4 Evaluation Methods

So far, we have provided an overview of studies that evaluate the behavior of LLMs on three core reasoning tasks. However, to date, a standardized methodology for assessing the reasoning capabilities of large language models is lacking. To address RQ2, we review and categorize predominant evaluation frameworks for analyzing the *behavior* of LLMs in reasoning tasks. As depicted in Figure 2, we categorize evaluation methodologies into four distinct groups: (i) *conclusion-based*, (ii) *rationale-based*, (iii) *interactive*, (iv) and *mechanistic* evaluations. In the following sections, we further delineate each category by discussing prevalent techniques. An overview of the advantages and disadvantages of each evaluation procedure can be found in Table 1. Additional details are provided in Appendix B.

4.1 Conclusion-Based Evaluation

In conclusion-based evaluation schemes, emphasis is placed on the model’s final answer rather than the process by which the model arrives at its conclusion. This outcome-oriented approach, despite its limitation of overlooking the model’s rationales, can nonetheless provide insights into the model’s reasoning behavior. For instance, a thorough *error analysis* can unveil conceptual errors (Sanyal et al., 2022), reflections of cognitive biases (Dasgupta et al., 2022), or sensitivities to the task context (Wu et al., 2024). Similarly, an examination of the model’s *output distribution* may reveal inherent predispositions towards certain outcomes (Itzhak et al., 2024), or serve as an indicator of the model’s confidence in specific conclusions (Frohberg & Binder, 2022). However, relying solely on the model’s final conclusion can be less reliable than also considering how the model arrives at its conclusion. For example, candidate answers derived from first token probabilities in multiple-choice setups are often not robust (Wang et al., 2024b), and final answers might not always align with the model’s verbalized reasoning (Mondorf & Plank, 2024; Ye & Durrett, 2022). Moreover, answers to benchmark datasets might have been compromised by data leakage during the model’s training process, limiting the insights that can be drawn from a correct conclusion (Balocco et al., 2024; Xu et al., 2024). To address the issue of data contamination and test the model’s capacity to generalize, *dynamic benchmarks* can be employed that update over time. For instance, functional benchmarks capture the underlying reasoning process rather than a

static problem formulation, probing the model’s reasoning capabilities on dataset snapshots released quarterly (Srivastava et al., 2024).

4.2 Rationale-Based Evaluation

In contrast to conclusion-based evaluation schemes, rationale-based evaluation methods focus on examining the reasoning trace generated by the model, typically assessing its logical validity and coherence. While this approach allows for more nuanced insights into the model’s reasoning behavior, rationale-based evaluation schemes are often more challenging to automate and scale. Given the variability in the model’s reasoning, a spectrum of assessment techniques exist. For rationales following a highly consistent format—either through structured contexts or methods that guide the model’s response style (Wan et al., 2024)—rationales can be *parsed* into more formalized representations such as first-order logic (Saparov & He, 2023), computation graphs (Dziri et al., 2023), or causal graphs (Willig et al., 2022), facilitating a more granular examination. Alternatively, *interpretable quantitative metrics*, such as ROSCOE (Golovneva et al., 2023) or RECEVAL (Prasad et al., 2023), may be utilized to evaluate the rationales’ semantic alignment with the reasoning task, their coherence, factual consistency, and logical validity. In instances where rationales are less structured, qualitative inspections are commonly employed, either through diagnostic agents (Huang et al., 2024), or human judgments (Mondorf & Plank, 2024).

4.3 Interactive Evaluation

Similar to the principle of dynamic assessment within psychology (Haywood & Tzuriel, 2002), interactive evaluation offers a framework to engage with LLMs during the evaluation. This approach allows for more flexible assessments tailored to the model’s specific reasoning behavior. Although such evaluations are often costly and challenging to scale, several techniques have been developed to automate the process. For instance, *adaptive evaluations* dynamically select reasoning tasks based on the model’s responses, thus providing deeper insights into its capabilities and limitations beyond what static questionnaires can reveal (Zhuang et al., 2023). *Dialectic evaluation* methods assess the model’s reasoning in dialogue form, for example, by challenging the model’s conclusions (Wang et al., 2023), or engaging it in game-theoretical scenarios (Bertolazzi et al., 2023). While interactive evaluations yield nuanced insights, they lack the structure of traditional methods, posing challenges in terms of standardization and reproducibility.

4.4 Mechanistic Evaluation

Mechanistic evaluations of LLMs delve into the underlying processes that drive the model’s responses, aiming to uncover the “*how*” and “*why*” within their reasoning processes. By analyzing the internal mechanisms such as attention patterns (Hou et al., 2023), activation flows (Dutta et al., 2024), and the functional attributes of individual layers (Pirozelli et al., 2023), deeper insights into the model’s operational logic can be gained, as illustrated in Section 3.1.1. Focusing on the model’s intrinsic processes, this framework contrasts with previous approaches, drawing parallels to the study of human reasoning from a neuroscientific perspective (Papo, 2015). Nonetheless, current methods remain compute-intensive, and their findings may not always generalize across different models and tasks (Bereska & Gavves, 2024; Ferrando et al., 2024).

5 Discussion

Despite the notable performance of large language models in prominent reasoning tasks (Bubeck et al., 2023; Fu et al., 2023), our review suggests that current models more closely resemble *stochastic parrots* (Bender et al., 2021) than systematic reasoners. As discussed in Section 3, we find that although many LLMs demonstrate proficiency in reasoning problems that align with their training data, the models’ reasoning behavior reveals significant conceptual errors and limitations in out-of-distribution scenarios. As highlighted by Mahowald et al. (2024), this suggests a limited functional linguistic competence in LLMs. It is likely

Evaluation Method	Advantages	Disadvantages
Conclusion-based evaluation	Allows for controlled setups Provides metrics for comparison Easy to automate and scale Easy to reproduce	Limited insights Less reliable
Rationale-based evaluation	Offers more nuanced insights More robust in certain scenarios	Difficult to automate and scale Might require expert interpretation
Interactive evaluation	Highly flexible Customizable to model behavior	Expensive Difficult to automate and scale Less standardized and reproducible
Mechanistic evaluation	Identifies features or circuits responsible for specific behaviors Supports direct interventions on model internals	Findings may not generalize across tasks or models Results may be hard to interpret Compute-intensive

Table 1: Comparison of strengths and limitations of the different evaluation methods.

that the apparent success of LLMs in reasoning tasks predominantly reflects their ability to memorize the extensive data they have been trained on (Wu et al., 2024; Dziri et al., 2023). Recent studies indicate that a substantial amount of benchmark datasets has been leaked to current LLMs (Balloccu et al., 2024; Xu et al., 2024), raising concerns about the insights derived from their performance on such benchmarks. Therefore, we advocate for more nuanced analyses of the models’ reasoning behavior, particularly in novel scenarios that the models have not previously encountered.

While human reasoning is not infallible (Johnson-Laird, 2010), the human capacity for robust reasoning and generalization from limited data remains unmatched by current LLMs. Various studies point to the fundamental differences between human reasoning and that of LLMs, especially the models’ restrictive autoregressive pre-training objective (McCoy et al., 2023; Shanahan, 2024; Lenci, 2023). We call for further research—particularly on reasoning behavior—within both humans and LLMs to better discern and comprehend the essential components missing in LLMs, which are crucial for robust and systematic reasoning.

6 Conclusion

This survey provides a comprehensive review of literature that evaluates LLM-based reasoning beyond mere task accuracy, offering deeper insights into the reasoning behavior of large language models. We discuss the behavior of LLMs across three core reasoning tasks (RQ1), assessing logical, mathematical and causal reasoning. Furthermore, we outline predominant evaluation frameworks and compare their respective strengths and limitations (RQ2). Our findings indicate that although LLMs demonstrate proficiency in reasoning problems that align with their training data, they often encounter significant conceptual challenges in out-of-distribution scenarios. Considering these insights and the recent issue of benchmark dataset leakage (Balloccu et al., 2024; Xu et al., 2024), we caution against relying solely on shallow accuracy metrics for static reasoning tasks. Instead, we advocate for more sophisticated reasoning assessments, such as those discussed in Section 4.

Acknowledgments

We express our gratitude to the anonymous reviewers for their insightful comments and suggestions. Furthermore, we would like to acknowledge that the emojis featured in Figure 1 are designed by [OpenMoji](#) – the open-source emoji and icon project (License: [CC BY-SA 4.0](#)). Finally, we gratefully recognize the support for BP through the ERC Consolidator Grant DIALECT 101043235.

References

- Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. Large language models for mathematical reasoning: Progresses and challenges. In Neele Falk, Sara Papi, and Mike Zhang (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pp. 225–237, St. Julian’s, Malta, March 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.eacl-srw.17>.
- AlephAlpha. Luminous language model, 2022.
- American Psychological Association. Behavior. In *APA Dictionary of Psychology*, n.d. Retrieved March 15, 2024, from <https://dictionary.apa.org/behavior>.
- Risako Ando, Takanobu Morishita, Hirohiko Abe, Koji Mineshima, and Mitsuhiro Okada. Evaluating large language models with NeuBAROCO: Syllogistic reasoning ability and human-like biases. In Stergios Chatzikyriakidis and Valeria de Paiva (eds.), *Proceedings of the 4th Natural Logic Meets Machine Learning Workshop*, pp. 1–11, Nancy, France, June 2023. Association for Computational Linguistics. URL <https://aclanthology.org/2023.naloma-1.1>.
- Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.
- Anthropic. Model card and evaluations for claude models, 2023. Technical Report. Available at: <https://www-cdn.anthropic.com/files/4zrzovbb/website/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226.pdf>.
- Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 67–93, St. Julian’s, Malta, March 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.eacl-long.5>.
- Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bénézet, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020. ISSN 1566-2535. doi: <https://doi.org/10.1016/j.inffus.2019.12.012>. URL <https://www.sciencedirect.com/science/article/pii/S1566253519308103>.

- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, pp. 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL <https://doi.org/10.1145/3442188.3445922>.
- Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety—a review. *arXiv preprint arXiv:2404.14082*, 2024.
- Leonardo Bertolazzi, Davide Mazzaccara, Filippo Merlo, and Raffaella Bernardi. ChatGPT’s information seeking strategy: Insights from the 20-questions game. In C. Maria Keet, Hung-Yi Lee, and Sina Zarrieß (eds.), *Proceedings of the 16th International Natural Language Generation Conference*, pp. 153–162, Prague, Czechia, September 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.inlg-main.11. URL <https://aclanthology.org/2023.inlg-main.11>.
- Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Muniyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogun, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R’e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *ArXiv*, 2021. URL <https://crfm.stanford.edu/assets/report.pdf>.
- Ali Borji. Stochastic parrots or intelligent systems? a perspective on true depth of understanding in llms, July 2023. Available at SSRN: <https://ssrn.com/abstract=4507038> or <http://dx.doi.org/10.2139/ssrn.4507038>.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.
- R. M. J. Byrne, J. S. B. T. Evans, and S. E. Newstead. *Human Reasoning: The Psychology Of Deduction*. Psychology Press, 1st edition, 1993. doi: 10.4324/9781315785028. URL <https://doi.org/10.4324/9781315785028>.

- Nuo Chen, Ning Wu, Shining Liang, Ming Gong, Linjun Shou, Dongmei Zhang, and Jia Li. Is bigger and deeper always better? probing llama across scales and layers. *arXiv preprint arXiv:2312.04333*, 2024a.
- Xinyun Chen, Ryan A Chi, Xuezhi Wang, and Denny Zhou. Premise order matters in reasoning with large language models. *arXiv preprint arXiv:2402.08939*, 2024b.
- Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. A survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402*, 2023.
- Kelsey M Collins, Catherine Wong, Jessica Feng, Michelle Wei, and Joshua Tenenbaum. Structured, flexible, and robust: Benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44. Cognitive Science Society, 2022. URL <https://escholarship.org/uc/item/3qq6w5kx>.
- Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*, 2022.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient fine-tuning of quantized llms. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423>.
- Aliya R. Dewey. Arbitrating norms for reasoning tasks. *Synthese*, 200(6):502, Nov 2022. ISSN 1573-0964. doi: 10.1007/s11229-022-03981-8. URL <https://doi.org/10.1007/s11229-022-03981-8>.
- Subhabrata Dutta, Joykirat Singh, Soumen Chakrabarti, and Tanmoy Chakraborty. How to think step-by-step: A mechanistic understanding of chain-of-thought reasoning. *arXiv preprint arXiv:2402.18312*, 2024.
- Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang (Lorraine) Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 70293–70332. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/deb3c28192f979302c157cb653c15e90-Paper-Conference.pdf.
- Tiwalayo Eisape, Michael Tessler, Ishita Dasgupta, Fei Sha, Sjoerd Steenkiste, and Tal Linzen. A systematic comparison of syllogistic reasoning in humans and language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8425–8444, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.466. URL <https://aclanthology.org/2024.naacl-long.466>.
- Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R Costa-jussà. A primer on the inner workings of transformer-based language models. *arXiv preprint arXiv:2405.00208*, 2024.

- Peter A. Flach and Antonis C. Kakas. *Abductive and Inductive Reasoning: Background and Issues*, pp. 1–27. Springer Netherlands, Dordrecht, 2000. ISBN 978-94-017-0606-3. doi: 10.1007/978-94-017-0606-3_1. URL https://doi.org/10.1007/978-94-017-0606-3_1.
- Shane Frederick. Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4):25–42, December 2005. doi: 10.1257/089533005775196732. URL <https://www.aeaweb.org/articles?id=10.1257/089533005775196732>.
- Jörg Froberg and Frank Binder. CRASS: A novel data set and benchmark to test counterfactual reasoning of large language models. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, H el ene Mazo, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 2126–2140, Marseille, France, June 2022. European Language Resources Association. URL <https://aclanthology.org/2022.lrec-1.229>.
- Yao Fu, Litu Ou, Mingyu Chen, Yuhao Wan, Hao Peng, and Tushar Khot. Chain-of-thought hub: A continuous effort to measure large language models’ reasoning performance. *arXiv preprint arXiv:2305.17306*, 2023.
- Olga Golovneva, Moya Peng Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. ROSCOE: A suite of metrics for scoring step-by-step reasoning. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=xYlJRpzZtsY>.
- Alison Gopnik and David M. Sobel. Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development*, 71(5):1205–1222, 2000. ISSN 00093920, 14678624. URL <http://www.jstor.org/stable/1131970>.
- Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chat-gpt. *Nature Computational Science*, 3(10):833–838, Oct 2023. ISSN 2662-8457. doi: 10.1038/s43588-023-00527-x. URL <https://doi.org/10.1038/s43588-023-00527-x>.
- Simon Jerome Han, Keith J. Ransom, Andrew Perfors, and Charles Kemp. Inductive reasoning in humans and large language models. *Cognitive Systems Research*, 83:101155, 2024. ISSN 1389-0417. doi: <https://doi.org/10.1016/j.cogsys.2023.101155>. URL <https://www.sciencedirect.com/science/article/pii/S1389041723000839>.
- H. Carl Haywood and David Tzuriel. Applications and challenges in dynamic assessment. *Peabody Journal of Education*, 77(2):40–63, 2002. doi: 10.1207/S15327930PJE7702_5. URL https://doi.org/10.1207/S15327930PJE7702_5.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL <https://openreview.net/forum?id=7Bywt2mQsCe>.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- K.J. Holyoak and R.G. Morrison. *The Cambridge Handbook of Thinking and Reasoning*. Cambridge Handbooks in Psychology. Cambridge University Press, 2005. ISBN 9780521824170. URL <https://books.google.de/books?id=znbkHaC8QeMC>.
- Leon Horsten. Philosophy of Mathematics. In Edward N. Zalta and Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition, 2023.

- Yifan Hou, Jiaoda Li, Yu Fei, Alessandro Stolfo, Wangchunshu Zhou, Guangtao Zeng, Antoine Bosselut, and Mrinmaya Sachan. Towards a mechanistic interpretation of multi-step reasoning capabilities of language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4902–4919, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.299. URL <https://aclanthology.org/2023.emnlp-main.299>.
- Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1049–1065, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.67. URL <https://aclanthology.org/2023.findings-acl.67>.
- Yinya Huang, Ruixin Hong, Hongming Zhang, Wei Shao, Zhicheng Yang, Dong Yu, Changshui Zhang, Xiaodan Liang, and Linqi Song. Clomo: Counterfactual logical modification with large language models. *arXiv preprint arXiv:2311.17438*, 2024.
- Itay Itzhak, Gabriel Stanovsky, Nir Rosenfeld, and Yonatan Belinkov. Instructed to Bias: Instruction-Tuned Language Models Exhibit Emergent Cognitive Bias. *Transactions of the Association for Computational Linguistics*, 12:771–785, 06 2024. ISSN 2307-387X. doi: 10.1162/tacl.a.00673. URL <https://doi.org/10.1162/tacl.a.00673>.
- Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng LYU, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. CLadder: A benchmark to assess causal reasoning capabilities of language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=e2wtjx0Yqu>.
- Zhijing Jin, Jiarui Liu, Zhiheng LYU, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona T. Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation? In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=vqIH00bdqL>.
- P. N. Johnson-Laird. Mental models and human reasoning. *Proc Natl Acad Sci U S A*, 107(43):18243–18250, 2010. doi: 10.1073/pnas.1012933107.
- Philip Johnson-Laird. *How We Reason*. Oxford University Press, 10 2008. ISBN 9780199551330. doi: 10.1093/acprof:oso/9780199551330.001.0001. URL <https://doi.org/10.1093/acprof:oso/9780199551330.001.0001>.
- P.N. Johnson-Laird. *How We Reason*. Oxford University Press, 2006. ISBN 9780198569763. URL <https://books.google.de/books?id=UjYsJN0krNYC>.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- Karl Christoph Klauer, Jochen Musch, and Barbara Naumer. On belief bias in syllogistic reasoning. *Psychological Review*, 107(4):852–884, 2000. doi: 10.1037/0033-295X.107.4.852. URL <https://doi.org/10.1037/0033-295X.107.4.852>.
- Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.
- Robert Koons. Defeasible Reasoning. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2022 edition, 2022.

- E. Kosoy, D. M. Chan, A. Liu, J. Collins, J. Hamrick, S. H. Huang, et al. Towards understanding how machines can learn causal overhypotheses. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45. Cognitive Science Society, 2023. URL <https://escholarship.org/uc/item/9q29w1xh>.
- Jacqueline P. Leighton. *Defining and Describing Reason*, pp. 3–11. Cambridge University Press, 2003.
- Alessandro Lenci. Understanding natural language understanding systems. a critical analysis. *arXiv preprint arXiv:2303.04229*, 2023.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL <https://aclanthology.org/2020.acl-main.703>.
- Jiaxuan Li, Lang Yu, and Allyson Ettinger. Counterfactual reasoning: Testing language models’ understanding of hypothetical scenarios. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 804–815, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.70. URL <https://aclanthology.org/2023.acl-short.70>.
- Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439*, 2023.
- Wentao Liu, Hanglei Hu, Jie Zhou, Yuyang Ding, Junsong Li, Jiayi Zeng, Mengliang He, Qin Chen, Bo Jiang, Aimin Zhou, and Liang He. Mathematical language models: A survey. *arXiv preprint arXiv:2312.07622*, 2024.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- David F. Lohman and Joni M. Lakin. *Intelligence and Reasoning*, pp. 419–441. Cambridge Handbooks in Psychology. Cambridge University Press, 2011.
- Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 2024.
- Jitendra Malik. Workshop on foundation models. <https://www.youtube.com/watch?v=dG628PEN1fY&t=15830s>, August 2021. YouTube video.
- R Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L Griffiths. Embers of autoregression: Understanding large language models through the problem they are trained to solve. *arXiv preprint arXiv:2309.13638*, 2023.
- Ian R. McKenzie, Alexander Lyzhov, Michael Martin Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Xudong Shen, Joe Cavanagh, Andrew George Gritsevskiy, Derik Kauffman, Aaron T. Kirtland, Zhengping Zhou, Yuhui Zhang, Sicong Huang, Daniel Wurgaft, Max Weiss, Alexis Ross, Gabriel Recchia, Alisa Liu, Jiacheng Liu, Tom Tseng, Tomasz Korbak, Najoung Kim, Samuel R. Bowman, and Ethan Perez. Inverse scaling: When bigger isn’t better. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL <https://openreview.net/forum?id=DwgRm72GQF>. Featured Certification.
- Grégoire Mialon, Roberto Dessi, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Roziere, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. Augmented language models: a survey. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL <https://openreview.net/forum?id=jh7wH2AzKK>. Survey Certification.

- Melanie Mitchell. Can large language models reason? AI: A Guide for Thinking Humans, September 2023. URL <https://aiguide.substack.com/p/can-large-language-models-reason>. Accessed on: 13th of March 2024.
- Melanie Mitchell and David C. Krakauer. The debate over understanding in ai’s large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120, 2023. doi: 10.1073/pnas.2215907120. URL <https://www.pnas.org/doi/abs/10.1073/pnas.2215907120>.
- Philipp Mondorf and Barbara Plank. Comparing inferential strategies of humans and large language models in deductive reasoning. *arXiv preprint arXiv:2402.14856*, 2024.
- Carey K. Morewedge and Daniel Kahneman. Associative processes in intuitive judgment. *Trends in Cognitive Sciences*, 14(10):435–440, 2010. doi: 10.1016/j.tics.2010.07.004.
- OpenAI. Chatgpt: Optimizing language models for dialogue, 2022. Available at: <https://openai.com/blog/chatgpt>.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry

- Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.
- David Papo. How can we study reasoning in the brain? *Frontiers in Human Neuroscience*, 9: 222, 2015. doi: 10.3389/fnhum.2015.00222.
- Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.
- Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.
- Paulo Pirozelli, Marcos M José, Anarosa AF Brandão, Fabio G Cozman, et al. Assessing logical reasoning capabilities of encoder-only transformer models. *arXiv preprint arXiv:2312.11720*, 2023.
- Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. ReCEval: Evaluating reasoning chains via correctness and informativeness. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10066–10086, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.622. URL <https://aclanthology.org/2023.emnlp-main.622>.
- Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. Reasoning with language model prompting: A survey. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5368–5393, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.294. URL <https://aclanthology.org/2023.acl-long.294>.
- Linlu Qiu, Liwei Jiang, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, and Xiang Ren. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=bNt7oajl2a>.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019. URL <https://api.semanticscholar.org/CorpusID:160025533>.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL <http://jmlr.org/papers/v21/20-074.html>.
- Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. Impact of pre-training term frequencies on few-shot numerical reasoning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics*:

- EMNLP 2022*, pp. 840–854, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.59. URL <https://aclanthology.org/2022.findings-emnlp.59>.
- Lance J. Rips. The psychology of knights and knaves. *Cognition*, 31(2):85–116, 1989. ISSN 0010-0277. doi: [https://doi.org/10.1016/0010-0277\(89\)90019-X](https://doi.org/10.1016/0010-0277(89)90019-X). URL <https://www.sciencedirect.com/science/article/pii/001002778990019X>.
- John Alan Robinson and Andrei Voronkov (eds.). *Handbook of Automated Reasoning (in 2 volumes)*. Elsevier and MIT Press, 2001. ISBN 0-444-50813-9. URL <https://www.sciencedirect.com/book/9780444508133/handbook-of-automated-reasoning>.
- Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. Thinking like a skeptic: Defeasible inference in natural language. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4661–4675, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.418. URL <https://aclanthology.org/2020.findings-emnlp.418>.
- Soumya Sanyal, Zeyi Liao, and Xiang Ren. RobustLR: A diagnostic benchmark for evaluating logical robustness of deductive reasoners. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9614–9631, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.653. URL <https://aclanthology.org/2022.emnlp-main.653>.
- Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *International Conference on Learning Representations*, 2023.
- Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Mehran Kazemi, Najoung Kim, and He He. Testing the general deductive reasoning capacity of large language models using ood examples. *Advances in Neural Information Processing Systems*, 36, 2024.
- Murray Shanahan. Talking about large language models. *Commun. ACM*, 67(2):68–79, jan 2024. ISSN 0001-0782. doi: 10.1145/3624724. URL <https://doi.org/10.1145/3624724>.
- Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Huai hsien Chi, Nathanael Scharli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, 2023. URL <https://api.semanticscholar.org/CorpusID:256459776>.
- Steven Sloman. *Causal Models: How People Think about the World and Its Alternatives*. Oxford University Press, 08 2005. ISBN 9780195183115. doi: 10.1093/acprof:oso/9780195183115.001.0001. URL <https://doi.org/10.1093/acprof:oso/9780195183115.001.0001>.
- Saurabh Srivastava, Anto PV, Shashank Menon, Ajay Sukumar, Alan Philipose, Stevin Prince, Sooraj Thomas, et al. Functional benchmarks for robust evaluation of reasoning performance, and the reasoning gap. *arXiv preprint arXiv:2402.19450*, 2024.
- Alessandro Stolfo, Zhijing Jin, Kumar Shridhar, Bernhard Schoelkopf, and Mrinmaya Sachan. A causal framework to quantify the robustness of mathematical reasoning with language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 545–561, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.32. URL <https://aclanthology.org/2023.acl-long.32>.
- Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, Yue Wu, Wenhai Wang, Junsong Chen, Zhangyue Yin, Xiaozhe Ren, Jie Fu, Junxian He, Wu Yuan, Qi Liu, Xihui Liu, Yu Li, Hao Dong, Yu Cheng, Ming Zhang, Pheng Ann Heng, Jifeng Dai, Ping Luo, Jingdong Wang, Ji-Rong Wen, Xipeng Qiu, Yike Guo, Hui Xiong, Qun Liu, and Zhenguo Li. A survey of reasoning with foundation models. *arXiv preprint arXiv:2312.11562*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayanan Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Yujing Zhang, Ravi Addanki, Antoine Miech, Annie Louis, Laurent El Shafey, Denis Teplyashin, Geoff Brown, Elliot Catt, Nithya Attaluri, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaly Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, Hanzhao Lin, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yong Cheng, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovska, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao

Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjöstrand, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlias, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, YaGuang Li, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Gamaleldin Elsayed, Ed Chi, Mahdis Mahdieh, Ian Tenney, Nan Hua, Ivan Petrychenko, Patrick Kane, Dylan Scandinaro, Rishub Jain, Jonathan Uesato, Romina Datta, Adam Sadovsky, Oskar Bunyan, Dominik Rabiej, Shimu Wu, John Zhang, Gautam Vasudevan, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Betty Chan, Pam G Rabinovitch, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajt Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Sahitya Potluri, Jane Park, Elnaz Davoodi, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Chris Gorgolewski, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Paul Suganthan, Evan Palmer, Geoffrey Irving, Edward Loper, Manaal Faruqui, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Michael Fink, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marin Georgiev, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnappalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Minnie Lui, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Lam Nguyen Thiet, Daniel Andor, Pedro Valenzuela, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Sarmishta Velury, Sebastian Krause, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Tejasi Latkar, Mingyang Zhang, Quoc Le, Elena Allica Abellan, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Sid Lall, Ken Franko, Egor Filonov, Anna Bulanova, Rémi Leblond, Vikas Yadav, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Hao Zhou, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Jeremiah Liu, Mark Omernick, Colton Bishop, Chintu Kumar, Rachel Sterneck, Ryan Foley, Rohan Jain, Swaroop Mishra, Jiawei Xia, Taylor Bos, Geoffrey Cideron, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Petru Gurita, Hila Noga, Premal Shah, Daniel J. Mankowitz, Alex Polozov, Nate Kushman, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Anhad Mohananey, Matthieu Geist, Sidharth

Mudgal, Sertan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Quan Yuan, Sumit Bagri, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Aliaksei Severyn, Jonathan Lai, Kathy Wu, Heng-Tze Cheng, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Mark Geller, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Andrei Sozanschi, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Abhimanyu Goyal, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Sabaer Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Tao Zhu, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Dustin Tran, Yeqing Li, Nir Levine, Ariel Stolovich, Norbert Kalb, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Balaji Lakshminarayanan, Charlie Deck, Shyam Upadhyay, Hyo Lee, Mike Dusenberry, Zonglin Li, Xuezhi Wang, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Summer Yue, Sho Arora, Eric Malmi, Daniil Mirylenka, Qijun Tan, Christy Koh, Soheil Hassas Yeganeh, Siim Pöder, Steven Zheng, Francesco Pongetti, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Ragha Kotikalapudi, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Chenkai Kuang, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Pei Sun, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Ishita Dasgupta, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Yuan Liu, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimota, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Ivo Penchev, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Adam Kurzrok, Lynette Webb, Sahil Dua, Dong Li, Preethi Lahoti, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Taylan Bilal, Evgenii Eltyshev, Daniel Balle, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Adams Yu, Christof Angermueller, Xiaowei Li, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Kevin Brooks, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Komal Jalan, Dinghua Li, Ginger Perng, Blake Hechtman, Parker Schuh, Milad Nasr, Mia Chen, Kieran Milan, Vladimir Mikulik, Trevor Strohman, Juliana Franco, Tim Green, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Think Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. Language models are not naysayers: an analysis of language models on negation benchmarks. In Alexis Palmer and Jose Camacho-collados (eds.), *Proceedings of the 12th Joint Conference on*

- Lexical and Computational Semantics (*SEM 2023)*, pp. 101–114, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.starsem-1.10. URL <https://aclanthology.org/2023.starsem-1.10>.
- Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974. doi: 10.1126/science.185.4157.1124. URL <https://www.science.org/doi/abs/10.1126/science.185.4157.1124>.
- Jean-Baptiste Van der Henst, Yingrui Yang, and P.N. Johnson-Laird. Strategies in sentential reasoning. *Cognitive Science*, 26(4):425–468, 2002. ISSN 0364-0213. URL <https://www.sciencedirect.com/science/article/pii/S0364021302000745>.
- Dylan Vassberg and John Vassberg. Is einstein’s puzzle over-specified? In *AJ75th Symposium: 21st Century Challenges in Computational Engineering & Science*, Princeton University, Princeton, NJ, November 2009.
- Vicuna. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. <https://vicuna.lmsys.org/>, 2023. Accessed: 2024-03-15.
- Yuxuan Wan, Wenxuan Wang, Yiliu Yang, Youliang Yuan, Jen tse Huang, Pinjia He, Wenxiang Jiao, and Michael R. Lyu. A & b == b & a: Triggering logical reasoning failures in large language models. *arXiv preprint arXiv:2401.00757*, 2024.
- Ben Wang and Aran Komatsuzaki. Gpt-j-6b: A 6 billion parameter autoregressive language model. <https://github.com/kingoflolz/mesh-transformer-jax>, May 2021. URL <https://github.com/kingoflolz/mesh-transformer-jax>.
- Boshi Wang, Xiang Yue, and Huan Sun. Can ChatGPT defend its belief in truth? evaluating LLM reasoning via debate. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11865–11881, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.795. URL <https://aclanthology.org/2023.findings-emnlp.795>.
- Xiaoqiang Wang, Bang Liu, and Lingfei Wu. Fac²e: Better understanding large language model capabilities by dissociating language and cognition. *arXiv preprint arXiv:2403.00126*, 2024a.
- Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Röttger, and Barbara Plank. Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think. *arXiv preprint arXiv:2404.08382*, 2024b.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.
- Moritz Willig, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. Can foundation models talk causality? In *UAI 2022 Workshop on Causal Representation Learning*, 2022. URL <https://openreview.net/forum?id=DbJXEqU0kaM>.
- Michael D. Wolcott and Nikki G. Lobczowski. Using cognitive interviews and think-aloud protocols to understand thought processes. *Currents in Pharmacy Teaching and Learning*, 13(2):181–188, 2021. ISSN 1877-1297. doi: <https://doi.org/10.1016/j.cptl.2020.09.005>. URL <https://www.sciencedirect.com/science/article/pii/S1877129720303026>.
- D. J. WOOD. Approach to the study of human reasoning. *Nature*, 223(5201):101–102, Jul 1969. ISSN 1476-4687. doi: 10.1038/223101a0. URL <https://doi.org/10.1038/223101a0>.

- Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najaoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *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, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.102. URL <https://aclanthology.org/2024.naacl-long.102>.
- Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. Are large language models really good logical reasoners? a comprehensive evaluation from deductive, inductive and abductive views. *arXiv preprint arXiv:2306.09841*, 2023.
- Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*, 2024.
- Zonglin Yang, Xinya Du, Rui Mao, Jinjie Ni, and Erik Cambria. Logical reasoning over natural language as knowledge representation: A survey. *arXiv preprint arXiv:2303.12023*, 2023.
- Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. Language models as inductive reasoners. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 209–225, St. Julian’s, Malta, March 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.eacl-long.13>.
- Xi Ye and Greg Durrett. The unreliability of explanations in few-shot prompting for textual reasoning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30378–30392. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c402501846f9fe03e2cac015b3f0e6b1-Paper-Conference.pdf.
- Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. Natural language reasoning, a survey. *ACM Comput. Surv.*, may 2024. ISSN 0360-0300. doi: 10.1145/3664194. URL <https://doi.org/10.1145/3664194>. Just Accepted.
- Wenhao Yu, Meng Jiang, Peter Clark, and Ashish Sabharwal. IfQA: A dataset for open-domain question answering under counterfactual presuppositions. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL <https://openreview.net/forum?id=V49Jx2Lj04>.
- Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL <https://openreview.net/forum?id=tv46tCzs83>.
- Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van Den Broeck. On the paradox of learning to reason from data. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI ’23*, 2023. ISBN 978-1-956792-03-4. doi: 10.24963/ijcai.2023/375. URL <https://doi.org/10.24963/ijcai.2023/375>.
- Yan Zhuang, Qi Liu, Yuting Ning, Weizhe Huang, Rui Lv, Zhenya Huang, Guanhao Zhao, Zheng Zhang, Qingyang Mao, Shijin Wang, et al. Efficiently measuring the cognitive ability of llms: An adaptive testing perspective. *arXiv preprint arXiv:2306.10512*, 2023.

A Human-Like Reasoning Behavior in LLMs

Recent research has begun to explore the extent to which human reasoning behaviors are reflected in large language models, given their training on human-generated data. Various studies delve into the manifestation of human-like reasoning behaviors within LLMs in the context of deductive reasoning tasks. For instance, [Eisape et al. \(2024\)](#) compare the behaviors of humans and LLMs, particularly those from the PaLM 2 series, in syllogistic reasoning tasks. The findings reveal that, similar to humans, LLMs are prone to logical fallacies and cognitive biases such as ordering effects. This susceptibility persists across model sizes, though larger models tend to engage in more deliberate reasoning, showing a reduced sensitivity towards these errors. A notable finding indicates that LLMs, unlike humans, rarely produce the “nothing follows” response, even when it represents the accurate deduction. In a similar vein, [Dasgupta et al. \(2022\)](#) demonstrate that LLMs, akin to humans, exhibit content effects in predicate logic problems, i.e. their reasoning is influenced by the semantic content of the task. This effect is evident in various LLMs including ChatGPT, PaLM 2, and Chinchilla ([Hoffmann et al., 2022](#)). In particular, analyses indicate that the models’ reasoning in syllogisms is biased by the believability of the conclusion, a behavior known as belief bias ([Klauer et al., 2000](#)). Further research corroborates the influence of semantic content on the models’ reasoning. [Ando et al. \(2023\)](#) demonstrate a belief-bias in models like ChatGPT, RoBERTa and BART ([Lewis et al., 2020](#)) by comparing the models’ behavior on abstract, belief-consistent, and belief-inconsistent syllogisms. Their analyses also uncover a predisposition towards conversion errors and atmosphere effects, where logical quantifiers are either misinterpreted or lead to uninformed inferences, respectively ([Tversky & Kahneman, 1974](#)). [Itzhak et al. \(2024\)](#) further show that methods such as instruction tuning and reinforcement learning from human feedback (RLHF) ([Ouyang et al., 2022](#)) may amplify cognitive biases within LLMs. [Mondorf & Plank \(2024\)](#) delve into the inferential strategies of open-access LLMs in problems of propositional logic. Comprehensive manual evaluations of the models’ rationales reveal that LLMs utilize inferential strategies similar to those employed by human reasoners. Their evaluations further underline difficulties of LLMs with logical negations and a vulnerability to logical fallacies commonly observed in human reasoning. Similarly, [McKenzie et al. \(2023\)](#) indicate that LLMs tend to replicate human-like logical errors when engaging with the logical principle of modus tollens, suggesting an imitation of flawed reasoning patterns from their training data. Notably, this trend becomes more pronounced with increasing model size, a phenomenon denoted as *inverse scaling*.

B Further Details on Evaluation Methods

In this section, we offer a more detailed overview of the evaluation frameworks commonly employed to examine the reasoning behavior of large language models. Expanding on the taxonomy depicted in Figure 2, we present *conclusion-based* evaluation methods in Table 2, provide an overview of *rationale-based* evaluation approaches in Table 3, discuss *interactive* evaluation techniques in Table 4, and delineate *mechanistic* evaluation strategies in Table 5.

Category	Evaluation Method	Description	Exemplary Reference
Output Distribution Analysis	Model Preference	Evaluate the likelihood of various candidate answers from a predefined set to assess the model’s tendency towards specific answers.	Itzhak et al. (2024)
	Model Confidence	Interpret the probability assigned to an answer as a measure of confidence towards that answer.	Dasgupta et al. (2022)
Error Analysis	Conceptual Errors	Assess the model’s errors with respect to fundamental principles and concepts within the domain of reasoning.	Sanyal et al. (2022)
	Context Sensitivity	Evaluate the model’s robustness towards perturbations of the task’s context.	Wu et al. (2024)
	Cognitive Biases	Probe the model with respect to cognitive biases or heuristics commonly encountered in human reasoning.	Eisape et al. (2024)
	Inverse Scaling	Analyze scenarios in which larger models tend to exhibit more pronounced errors than smaller models.	McKenzie et al. (2023)
Dynamic Benchmarks	Functional Benchmarks	Transform static benchmarks in question-answer format into functional form, where triplets of (problem, solution, input) are defined that represent the underlying reasoning process of the task, and thus allow to produce flexible question-answer pairs.	Srivastava et al. (2024)

Table 2: Overview of conclusion-based evaluation methods that assess the model’s reasoning behavior by focusing on the final answers they produced.

Category	Evaluation Method	Description	Exemplary Reference
Structured Parsing	FOL Conversion	Translate the model’s rationale into first-order-logic (FOL) language and evaluate the logical expressions.	Saparov & He (2023)
	Computational Graphs	Parse the rationale into a computation graph where vertices represent intermediate results and edges denote function mappings.	Dziri et al. (2023)
	FAC ² E	Use capability-specific instructions to elicit intermediate structured reasoning steps (crystallized and fluid reasoning). Evaluate each step separately, using ground truth responses and the BARTScore-Recall metric.	Wang et al. (2024a)
Interpretable Quantitative Metrics	ROSCOE	Assess the model’s rationale using a suite of nuanced metrics, quantifying its semantic alignment, semantic similarity, logical inference, and language coherence.	Golovneva et al. (2023)
	ReCEval	Analyze the model’s rationale using metrics that quantify its intra-step and inter-step logical validity, as well as the informativeness of each reasoning step.	Prasad et al. (2023)
Qualitative Inspection	Human Evaluator	Inspect rationales manually through human annotators.	Mondorf & Plank (2024)
	Diagnostic Agents	Assess rationales by means of additional language models, acting as diagnostic agents.	Huang et al. (2024)

Table 3: Rationale-based evaluation methods that examine the model’s reasoning behavior through an analysis of its underlying rationale.

Category	Evaluation Method	Description	Exemplary Reference
Adaptive Evaluation	Computerized Adaptive Testing (CAT)	Dynamically adapt questions presented during the evaluation procedure based on the model's responses.	Zhuang et al. (2023)
Dialectic Evaluation	Belief Defense	Test the model's reasoning through challenging its prior response in a conversational discourse.	Wang et al. (2023)
	Game-Theoretic Analysis	Evaluate the model's reasoning by engaging with it in a game-theoretic scenario.	Bertolazzi et al. (2023)

Table 4: Interactive evaluation techniques designed to assess the model's reasoning behavior through dynamic interaction.

Evaluation Method	Description	Exemplary Reference
Layer Probing	Probe the functional role of different layers of the model's architecture.	Pirozelli et al. (2023)
Attention Pattern Analysis	Analyze the model's attention matrices to gain insights into the underlying information flow within the model.	Hou et al. (2023)
Activation Patching	Alter specific neuron activations within a model and observe its impact on the model's output.	Dutta et al. (2024)

Table 5: Overview of mechanistic evaluation methods that assess the model's reasoning by delving into its internal mechanisms throughout the reasoning process.