A Survey on Large Language Model Reasoning Failures

Anonymous authors
Paper under double-blind review

Abstract

Large Language Models (LLMs) have exhibited remarkable reasoning capabilities, achieving impressive results across a wide range of tasks. Despite these advances, significant reasoning failures persist, occurring even in seemingly simple scenarios. To systematically understand and address these shortcomings, we present the ***first comprehensive survey dedicated to reasoning failures in LLMs***. We introduce a novel categorization framework that distinguishes ***reasoning*** into embodied and non-embodied types, with the latter further subdivided into informal (intuitive) and formal (logical) reasoning. In parallel, we classify reasoning ***failures*** along a complementary axis into three types: fundamental failures intrinsic to LLM architectures that broadly affect downstream tasks; application-specific limitations that manifest in particular domains; and robustness issues characterized by inconsistent performance across minor variations. For each reasoning failure, we provide a clear definition, analyze existing studies, explore root causes, and present mitigation strategies. By unifying fragmented research efforts, our survey provides a structured perspective on systemic weaknesses in LLM reasoning, offering valuable insights and guiding future research towards building stronger, more reliable, and robust reasoning capabilities.

1 Introduction

"Failure is success if we learn from it." – Malcolm Forbes

With the rise of powerful architectures (Vaswani et al., 2023; Jiang et al., 2024a; Gu & Dao, 2024; Hasani et al., 2020), efficient algorithms (Hu et al., 2021; Zhao et al., 2024b; Gretsch et al., 2024; 2025; Dao et al., 2022), and massive data (Cai et al., 2024; Raffel et al., 2020; Gao et al., 2020), Large Language Models (LLMs) have recently shown significant success across diverse domains. These range from traditional linguistic tasks such as machine translation (Zhu et al., 2024b; Tang et al., 2024), to mathematical (Shao et al., 2024; Yang et al., 2023a; 2024a) and even scientific (Zhang et al., 2024b; Wang et al., 2023b; Brodeur et al., 2024) discoveries. Among these achievements, reasoning as an emergent capability of LLMs (Wei et al., 2022a) has attracted particular interest (Huang & Chang, 2023; Yu et al., 2023b; Qiao et al., 2023).

LLMs have set impressive records in reasoning (Wu et al., 2025a; Kıcıman et al., 2024; Plaat et al., 2024), though it remains controversial whether LLMs really leverage a human-like reasoning procedure when attempting these tasks (Jiang et al., 2024b; Fedorenko et al., 2024; Amirizaniani et al., 2024b; Zhang et al., 2022). This survey bears no aim to settle this hot debate; rather we focus on an important area of study in LLM reasoning that has long been overlooked – LLM reasoning failures.

Extensive psychological research (Cannon & Edmondson, 2005; Maxwell, 2007; Coelho & McClure, 2004) underscores the importance of identifying and learning from failures in human development[1]. Given that AI systems have historically drawn inspiration from human cognition (Schmidgall et al., 2023; Xu & Poo, 2023; Woźniak et al., 2020), we believe the same principle of learning from failures could similarly benefit the study of LLMs, since such failures can usually be traced back to fundamental elements and bring valuable insights to ultimate improvements (Dreyfus, 1992; Karl et al., 2024; An et al., 2024).

Despite some existing works that prospectively realized this importance and investigated LLM reasoning failures on a case-by-case basis (Williams & Huckle, 2024; Tie et al., 2024; Helwe et al., 2021; Borji, 2023), the topic remains fragmented, and underexplored as a unified research area. This fragmentation limits broader

[1] In fact, this theory has been confirmed even more broadly, in non-human animals (Spence, 1936).

understanding, which is however a prerequisite for common patterns to be noticed, and thereby meaningful lessons to be derived. To close this gap, we present the first comprehensive survey dedicated to unifying studies on LLM reasoning failures. We identify meaningful patterns across failures, analyze underlying causes, and discuss potential mitigation strategies. We hope this work not only organize the field but also stimulate further research and increased attention, toward improving the robustness and reliability of LLM reasoning.

2 Definition and Formulation

2.1 Fundamentals of Reasoning

Human reasoning broadly refers to the ability to draw conclusions and make decisions based on available knowledge (Lohman & Lakin, 2011; Ribeiro et al., 2020). Within cognitive science and philosophy, reasoning has been studied through various frameworks. To systematically survey reasoning failures in LLMs, we propose a comprehensive taxonomy distinguishing reasoning along two primary axes: ***embodied*** versus ***non-embodied***, with the latter further subdivided into ***informal*** and ***formal*** reasoning.

Non-embodied reasoning. Non-embodied reasoning comprises cognitive processes not requiring physical interaction with environments. Within this category, *informal reasoning* encompasses intuitive judgments driven by inherent biases and heuristics, common in everyday decision-making and social activities (Piaget, 1952; Vygotsky, 1978; Kail, 1990). By contrast, *formal reasoning* involves explicit, rule-based manipulation of symbols, grounded in logic, mathematics, code, etc. (Copi et al., 2016; Mendelson, 2009; Liu et al., 2023b).

Embodied reasoning. Embodied reasoning depends on physical interaction with environments, fundamentally relying on spatial intelligence and real-time feedback (Shapiro, 2019; Barsalou, 2008). This includes predicting and interpreting physical dynamics, and performing goal-directed behaviors constrained by real-world physical laws (Huang et al., 2022b; Lee-Cultura & Giannakos, 2020).

2.2 LLM Reasoning Failures & Common Research Practice

Despite advances in interpretability research (Dwivedi et al., 2023; Li et al., 2024e), LLMs remain largely *black-box* systems (Luo & Specia, 2024), reflecting the inherent complexity of human cognition they emulate (Castelvecchi, 2016). As such, reasoning abilities are typically assessed behaviorally by examining model outputs on carefully designed prompts and tasks (Ribeiro et al., 2020). We define ***LLM reasoning failures*** as cases where model responses significantly diverge from expected logical coherence, contextual relevance, or factual correctness. Failures can manifest in two broad ways. The first type is straightforward poor performance — the model fails decisively on a task, exposing clear deficiencies. The second, subtler type involves apparently adequate performance that is in fact unstable, indicating a ***robustness*** issue that reveals hidden vulnerabilities. The former category – straightforward failure – can be sub-divided into two, based on scope and nature. ***Fundamental*** failures are usually intrinsic to LLM architectures, manifesting broadly and universally across diverse downstream tasks. In contrast, application-specific ***limitations*** reflect shortcomings tied to particular domains of importance, where models underperform despite human expectations of competence. Together, these taxonomies — for reasoning and for failures — offer a comprehensive and mutually consistent framework. Figure 1 uses this framework to visualize a clear organization of topics in this survey.

Current research in this space typically begins with ***simple, intuitive tests*** that reveal glaring reasoning failures. These initial observations motivate ***larger-scale systematic evaluations***, to confirm the generality and impact of identified failure modes. By explicitly defining and categorizing LLM reasoning failures according to our framework, this survey unifies fragmented research findings, highlights shared patterns, and directs focused efforts toward understanding and mitigating critical reasoning weaknesses. To help visualize the failure cases, we provide a few most representative examples for each of the failure case presented in this survey. The examples can be found in Appendix E.

3 Reasoning Informally in Intuitive Applications

Humans naturally develop the capacity for informal reasoning early in life, relying on intuitive judgments shaped by cognitive processes and social experiences. Though often taken for granted, this forms the foundation of human reasoning and decision-making. In this section, we focus on failures exhibited by LLMs

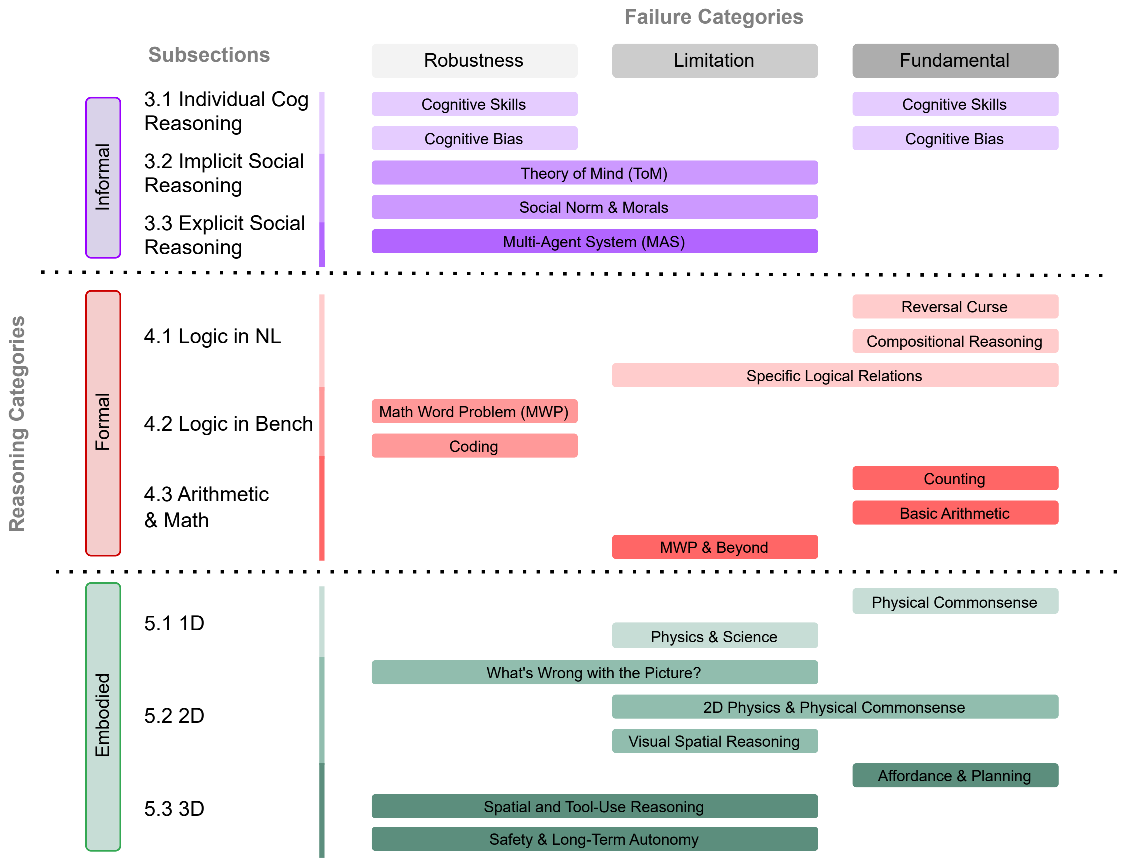


Figure 1: **A Taxonomy of LLM Reasoning Failures.** We adopt a nuanced 2-axis structure (reasoning type × failure type), with each row representing a reasoning category and each column a failure category. A more detailed explanation is presented in Section 2.

in such informal reasoning. We begin by examining reasoning failures in core cognitive abilities reflected in individual LLM behaviors; then explore those exposed in social contexts, expressed implicitly or explicitly.

3.1 Individual Cognitive Reasoning

Many reasoning failures exhibited by LLMs can be traced back to core human cognitive phenomena (Han et al., 2024b; Gong et al., 2024; Galatzer-Levy et al., 2024; Suri et al., 2024). These failures arise either because LLMs lack certain fundamental cognitive abilities possessed by humans – leading to errors that humans typically avoid (Han et al., 2024b) – or because LLMs replicate human-like cognitive biases and heuristics, resulting in analogous mistakes (Suri et al., 2024; Lampinen et al., 2024). In both cases, these failures relate closely to ***well-documented human cognitive phenomena*** and psychological evidence.

Fundamental Cognitive Skills. Humans naturally possess a set of fundamental cognitive skills indispensable for reasoning. LLMs demonstrate systematic failures due to deficiencies in these areas. A prominent example is the set of ***core executive functions*** – working memory (Baddeley, 2020), inhibitory control (Diamond, 2013; Williams et al., 1999), and cognitive flexibility (Canas et al., 2006) – essential in human reasoning (Diamond, 2013). ***Working memory*** is the capacity to hold and manipulate information over short periods. LLMs' limited working memory leads to failures when task demands exceed their capacity (Gong et al., 2024; Zhang et al., 2024a; Gong & Zhang, 2024; Upadhayay et al., 2025; Huang et al., 2025a). In particular, LLMs suffer from "proactive interference" to a much larger extent than humans, where earlier information significantly disrupts retrieval of newer updates (Wang & Sun, 2025). ***Inhibitory control*** – the ability to suppress impulsive or default responses when contexts demand – is also weak in LLMs, with them often sticking to previously learned patterns even when contexts shift (Han et al., 2024b; Patel et al., 2025). Lastly, ***cognitive flexibility***, the skill of adapting to new rules or switching tasks efficiently, remains a challenge, especially in rapid task switching and adaptation to new instructions (Kennedy & Nowak, 2024).

Another key aspect is ***abstract reasoning*** (Guinungco & Roman, 2020), the cognitive ability to recognize patterns and relationships in intangible concepts. Even advanced LLMs struggle with abstract reasoning tasks, such as inferring underlying rules from limited examples, understanding implicit conceptual relationships, and reliably handling symbolic or temporal abstractions (Xu et al., 2023c; Gendron et al., 2023; Galatzer-Levy et al., 2024; Saxena et al., 2025).

These phenomena are **fundamental** reasoning failures that are from intrinsic limitations of LLM architectures and training dynamics, and often manifest as robustness vulnerabilities across a wide range of tasks. Recent work attributes these failures to the underlying self-attention mechanism's dispersal of focus under complex tasks (Gong & Zhang, 2024; Patel et al., 2025), and to the next token prediction training objective, which prioritizes statistical pattern completion over deliberate reasoning (Han et al., 2024b; Enström et al., 2024). Some also point out that unlike humans – who develop fundamental cognitive functions through embodied, goal-driven interactions with the physical and social world (Pearce & Miller, 2025; Rodríguez, 2022; Jin et al., 2018) – LLMs learn passively from text alone, lacking grounding and experiential feedback to support the development. Efforts to enhance these skills correspondingly include advanced prompting like Chain-of-Thought (CoT) (Wei et al., 2022b), retrieval augmentation (Xu et al., 2023b), fine-tuning with deliberately injected interference (Li et al., 2022), multimodality (Hao et al., 2025), and architectural innovations to mimic human attention mechanisms (Wu et al., 2024d).

Cognitive Biases. Cognitive biases – systematic deviations from rational judgment – are well-studied in human reasoning (Tversky & Kahneman, 1974; 1981). They arise from mental shortcuts, limited cognitive resources, or contextual influences, often leading to predictable errors. LLMs exhibit similar biases that systematically affect their reasoning across diverse tasks (Hagendorff, 2023; Bubeck et al., 2023). Since these biases are deeply ingrained from training data and model architecture, they permeate a wide range of downstream applications, necessitating careful identification and mitigation.

In humans, these biases become evident only when information is presented and their responses observed – similarly, in LLMs, cognitive biases manifest also through the processing of information. Here lie two interrelated factors: ***the content of information*** and ***the presentation of that information***. Regarding content, LLMs struggle more with abstract or unfamiliar topics – a phenomenon known as the "content effect" (Lampinen et al., 2024) – and tend to favor information that aligns with prior context or assumptions, reflecting human-like confirmation bias (O'Leary, 2025b; Shi et al., 2024; Malberg et al., 2024; Wan et al., 2025b; Zhu et al., 2024c). Social cognitive biases also influence LLM outputs, including group attribution bias (Hamilton & Gifford, 1976; Allison & Messick, 1985; Raj et al., 2025) and negativity bias (Rozin & Royzman, 2001), which prioritize popular content (Echterhoff et al., 2024; Lichtenberg et al., 2024; Jiang et al., 2025a) and negative inputs (Yu et al., 2024c; Malberg et al., 2024) respectively.

Equally important is how the same content is presented. LLMs are highly sensitive to the order in which information is given, exhibiting order bias (Koo et al., 2023; Pezeshkpour & Hruschka, 2023; Jayaram et al., 2024; Guan et al., 2025; Cobbina & Zhou, 2025), and show anchoring bias (Lieder et al., 2018; Rastogi et al., 2022), where early inputs disproportionately shape their reasoning (Lou & Sun, 2024; O'Leary, 2025a; Huang et al., 2025e; Wang et al., 2025b). Framing effects further influence outputs: logically equivalent but differently phrased prompts can lead to different results (Jones & Steinhardt, 2022; Suri et al., 2024; Nguyen, 2024; Lior et al., 2025; Robinson & Burden, 2025; Shafiei et al., 2025). Additionally, factors like narrative perspective (e.g., first-person vs. third-person) (Cohn et al., 2024; Lin et al., 2024b), prompt length or verbosity (Koo et al., 2023; Saito et al., 2023), and irrelevant or distracting information (Shi et al., 2023) further derail logical reasoning.

Cognitive biases constitute **fundamental** reasoning failures rooted in LLM training paradigms and architectures, and they manifest as **robustness** vulnerabilities across a wide range of downstream applications. The root causes of these cognitive biases in LLMs are threefold. First, biases are *inherited from the pre-training data*, where the linguistic patterns in human languages reflect cognitive errors (Itzhak et al., 2025). Second, architectural features of the model – such as the Transformer's causal masking – introduce predispositions toward order-based biases independent of data (Wu et al., 2025b; Dufter et al., 2022). Third, *alignment* processes like Reinforcement Learning from Human Feedback (RLHF) amplify biases by aligning model behavior with human raters who are themselves biased (Sumita et al., 2025; Perez et al., 2023).

Mitigation strategies fall into three categories. *Data-centric* approaches focus on curating training data to reduce biased content (Sun et al., 2025a; Schmidgall et al., 2024; Han et al., 2024a). *In-processing* techniques, such as adversarial training, aim to prevent biased associations during model learning (Yang et al., 2023b; Cantini et al., 2024). Lastly, *post-processing* methods leverage prompt engineering or output filtering to steer model responses after training (Sumita et al., 2025; Lin & Ng, 2023). In this category, indirect methods like inducing specific model personalities have also shown promise in modulating biases (Shi et al., 2024; He & Liu, 2025). Nonetheless, even when mitigated in one context, cognitive biases often re-emerge when contexts shift. The diverse and penetrative nature of cognitive biases makes them difficult to be fully eliminated.

3.2 Implicit Social Reasoning

Certain cognitive reasoning failures manifest only within social contexts. We define ***implicit social reasoning*** as an individual model's capacity to internally infer and reason about (1) others' mental states (e.g., beliefs, emotions, intentions) and (2) shared social norms *without requiring direct interaction*.

Theory of Mind (ToM). ToM is the cognitive ability to attribute mental states – beliefs, intentions, emotions – to oneself and others, and to understand that others' mental states may differ from one's own (Frith & Frith, 2005). ToM enables humans to interpret behaviors, predict actions, and navigate complex interpersonal interactions, central in social reasoning. Typically emerging in early childhood with milestones like passing false belief tasks (understand that others' beliefs may be incorrect or different) (Wimmer & Perner, 1983), ToM has been a central focus in human psychology and cognitive science.

Under this inspiration, recent research evaluates the ToM capacity of LLMs, to gauge their ability to engage in social reasoning. Early studies focused on classic ToM tasks, such as false-belief (van Duijn et al., 2023; Kim et al., 2023), perspective-taking (infer what another individual perceives) (Sap et al., 2022), and unexpected content tasks (predicting what others would believe is inside a mislabeled unopened container) (Pi et al., 2024). Surprisingly, even advanced models such as GPT-4 struggle with these tasks trivial for human children. Moreover, minor modifications in task phrasing lead to drastic drops in performance, showing LLM ToM reasoning is unstable (Ullman, 2023; Kosinski, 2023; Pi et al., 2024; Shapira et al., 2023).

While there has been clear progress from early models like GPT-3 – which largely failed at ToM tasks – to newer models such as GPT-4o and reasoning models like o1-mini, which can solve many standard ToM tests, their underlying reasoning remains brittle under simple perturbations (Gu et al., 2024; Zhou et al., 2023d). Also, LLMs still struggle with higher-order, more complex aspects of ToM, such as predicting others' behaviors, making appropriate moral or social judgments, and translating this understanding into coherent actions (He et al., 2023; Gu et al., 2024; Marchetti et al., 2025; Amirizaniani et al., 2024a; Strachan et al., 2024). Particularly, on dynamic, conversational benchmarks (Xiao et al., 2025; Kim et al., 2023), even state-of-the-art models fail to demonstrate consistent ToM capabilities and perform significantly worse than humans. Furthermore, current models exhibit deficits in *emotional* reasoning. This includes difficulties in emotional intelligence (EI) (Sabour et al., 2024; Hu et al., 2025; Amirizaniani et al., 2024b; Vzorinab et al., 2024), susceptibility to affective bias (Chochlakis et al., 2024), and limited understanding of cultural variations in emotional expression and interpretation (Havaldar et al., 2023).

While prompting techniques like CoT offer some improvements (Gandhi et al., 2024), fundamental gaps remain due to deeper limitations from the LLM architecture, training paradigms, and a lack of embodied cognition (Strachan et al., 2024; Sclar et al., 2023). Failures in this domain constitute important application-specific **limitations**, and because ToM underlies many socially grounded tasks, such failures often result in significant **robustness** vulnerabilities. Given ToM's centrality to social reasoning, future work should move beyond prompting, to probe deeper root causes and general mitigation.

Social Norms and Moral Values. LLMs also struggle with reasoning about social norms, moral values, and ethical principles that govern human behavior. Unlike humans, who develop moral and ethical reasoning through experience, LLMs, trained purely on text, often exhibit inconsistent and unreliable social, moral, and ethical reasoning (Ji et al., 2024; Jain et al., 2024b).

One key limitation is that LLMs cannot reason and apply ***moral values*** (Ji et al., 2024) and ***social norms*** (Jain et al., 2024b) consistently. They often produce contradictory ethical judgments or varied moral reasoning performance when questions are slightly reworded (Bonagiri et al., 2024), generalized (Tanmay et al., 2023), or presented in a different language (Agarwal et al., 2024). Fine-tuning further exacerbates these inconsistencies, leading to sometimes prioritizing task-specific optimization over ethical coherence (Yu et al., 2024a).

Beyond ***inconsistencies***, LLMs show notable ***disparities*** compared to humans in reasoning with social norms and moral values. These models fail significantly in understanding real-world social norms (Rezaei et al., 2025), aligning with human moral judgments (Garcia et al., 2024; Takemoto, 2024), and adapting to cultural differences (Jiang et al., 2025b). Without consistent and reliable moral reasoning, LLMs are not fully ready for real-world decision-making involving ethical considerations.

These inconsistencies and disparities constitute important application-specific **limitations** for safety, privacy, sensitivity, and morality-related tasks, and such failures often create severe **robustness vulnerabilities**, including susceptibility to jailbreaks and other forms of manipulation. Many argue that these failures stem from a fundamental absence of robust, internalized representations of ethical principles, normative frameworks, and moral intentionality (Chakraborty et al., 2025; Wang et al., 2025a; Pock et al., 2023; Almeida et al., 2024). Although training procedures such as RLHF and instruction fine-tuning introduce alignment signals, they often operate superficially and fail to produce coherent moral behavior in complex contexts (Dahlgren Lindström et al., 2025; Wang et al., 2025a; Barnhart et al., 2025; Han et al., 2025). Current efforts to address these limitations mainly include prompt-based interventions (Chakraborty et al., 2025; Ma et al., 2023), internal activation steering (Tlaie, 2024; Turner et al., 2023), and direct fine-tuning on curated moral reasoning benchmarks (Senthilkumar et al., 2024; Karpov et al., 2024). However, in practice, these methods often suffer from the same limitations as RLHF, offering surface-level and task-specific improvements that remain vulnerable against prompt manipulations and jailbreaks.

3.3 Explicit Social Reasoning

In reasoning, "society" can refer to not just an abstract concept but real-world settings involving interactions among multiple agents. In Multi-Agent Systems (MAS), explicit social reasoning is ***the capacity of AI systems to collaboratively plan and solve complex tasks***, an area challenging for current LLMs.

Currently, key challenges include (1) ***long-horizon planning***, (2) ***communications and ToM***, and (3) ***robustness and adaptability***. Long-horizon planning is the ability to maintain coherent and coordinated strategies over extended interactions, where LLMs frequently fail (Li et al., 2023a; Cross et al., 2024; Guo et al., 2024c; Han et al., 2024c; Zhou et al., 2025) as they rely excessively on local or recent information (Piatti et al., 2024; Zhang et al., 2023; Han et al., 2024c). Furthermore, individual agents' social reasoning failures (discussed in Section 3.2, e.g., inefficient communication and ToM) (Guo et al., 2024c; Agashe et al., 2024; Zhou et al., 2025), lead to misinterpretations and inaccurate representations of other agents, causing strategic misalignment (Pan et al., 2025; Li et al., 2023a; Cross et al., 2024; Han et al., 2024c). Lastly, MAS face robustness and adaptability issues (Li et al., 2023a; Cross et al., 2024), lacking resilience to disruptive or malicious disturbances (Huang et al., 2024) and struggling with task verification and termination (Pan et al., 2025; Baker et al., 2025).

These failures stem from both *individual LLM capabilties* and *MAS system design* (Pan et al., 2025), representing key application-specific failures, and they frequently manifest as **robustness** vulnerabilities in multi-agent settings. Standard LLMs, optimized for next-token prediction, lack the explicit reasoning depth needed for multi-step, jointly conditioned objectives, and their fragile ToM representations cause coordination breakdowns. Individual limitations in cognitive skills, such as working memory, and cognitive biases, such as the anchoring effect, can also lead to MAS failures like difficulties with long-horizon planning. On the system level, many MAS often lack effective robustness layers – clear role specifications, cross-verification among agents, and reliable termination checks – allowing errors to cascade (Huang et al., 2024; Pan et al., 2025).

Mitigation research thus targets (i) richer internal models like belief tracking and hypothesis testing (Li et al., 2023a; Cross et al., 2024), (ii) structured communication protocols with mandatory verification phases (Pan et al., 2025), and (iii) dedicated inspector or challenger agents that monitor and contest questionable outputs (Huang et al., 2024; Baker et al., 2025). While these approaches reduce errors, none eliminate them and all require significant task-specific engineering that is difficult to generalize. In parallel, the recent rise of context engineering (Mei et al., 2025) – which focuses on a systematic optimization of the entire information payload fed to an LLM during inference – is increasingly seen as a more robust alternative to traditional prompt engineering in MAS. Real-world deployment will hence require an integrated stack combining all three strands with domain fine-tuning and formal safety guarantees (Lindemann & Dimarogonas, 2025; de Witt, 2025).

4 Reasoning Formally in Logic

When reasoning goes beyond intuition, a formal framework is needed to ensure rigor. As introduced in Section 2, ***logic*** is directly about ***doing "correct" reasoning***, ensuring ***premises support conclusions*** (Jaakko & Sandu, 2002). LLM failures in logical reasoning (Liu et al., 2025) thus pose serious risks, potentially leading to flawed thought processes and harmful decisions. Logic spans a continuum from implicit structures in natural languages (Iwańska, 1993), to explicit symbolic (Lewis et al., 1959) and mathematical (Shoenfield, 2018) representations. This section follows that progression, examining failures in increasingly formal reasoning paradigms.

4.1 Logic in Natural Languages

Reversal Curse. While natural languages are not fully logical structures (Fedorenko et al., 2024), they do hold simple logical relations (Sampson, 1979; Stich, 1975) that humans trivially grasp. A representative failure of LLMs is ***reversal curse***: despite being trained on "A is B," models often fail to infer the equivalent "B is A" – a trivial bidirectional equivalence for humans. Such failures occur even when a factual sentence from training data is just restated as a question during inference. First observed by Berglund et al. (2023) as a **fundamental** failure that occur widely across tasks on GPT-based (Radford & Narasimhan, 2018) models, this phenomenon is later shown in Wu et al. (2024a) not to affect BERT (Devlin et al., 2019).

This failure has been attributed to uni-directional training objectives of Transformer-based LLMs (Lv et al., 2024; Lin et al., 2024c), which induce structural asymmetry in model weights (Zhu et al., 2024a) and inability to predict antecedent words within training data (Guo et al., 2024b; Youssef et al., 2024). Golovneva et al. (2024) further argues that scaling alone cannot resolve the issue due to Zipf's law (Newman, 2005). Mitigation efforts accordingly center on reducing directional bias through training data augmentation. Early approaches syntactically reverse facts (Lu et al., 2024; Ma et al., 2024b), while later methods introduce substring-preserving reversals (Golovneva et al., 2024) and permuting semantic units in training data (Guo et al., 2024b). Despite differing in complexity, all methods share a common goal: ***exposing models to bidirectional formulations to restore logical symmetry***.

Compositional Reasoning. Compositional reasoning requires combining *multiple* pieces of knowledge or arguments into a coherent inference. **Fundamental** failures arise when LLMs are ***capable*** of each component but fail in ***integrating*** them. Studies show systematic failures in basic two-hop reasoning – combining only two facts across documents – and even worsening performance with increased compositional depth and the addition of distractors (Zhao & Zhang, 2024; Xu et al., 2024b; Guo et al., 2025a). This fundamental weakness extends beyond basic tasks, to compositions of math problems (Zhao et al., 2024c; Hosseini et al., 2024; Sun et al., 2025b) (i.e., LLMs succeed in individual problems but fail in composed ones), multi-fact claim verification (Dougrez-Lewis et al., 2024), and other inherently compositional tasks (Dziri et al., 2023).

This failure is attributed to an inability of **holistic planning and in-depth thinking.** CoT prompting improves on this by making reasoning steps explicit at inference time. Still, latent compositionality is more efficient in practice yet harder to achieve (Yang et al., 2024c). Toward this, Li et al. (2024f) identifies faulty implicit reasoning in mid-layer multi-head self-attention (MHSA) modules and edit them, while Zhou et al. (2024a) enhances training with graph-structured reasoning path data, similar to distilling CoT reasoning process into training data (Yu et al., 2024b). Both converge in spirit to improving latent compositional reasoning by explicitly guiding models' internal reasoning mechanisms.

Specific Logical Relations. Both reversal curse and compositional reasoning reflect **fundamental** failures affecting a broad range of reasoning tasks, exposed across general corpora or arbitrary logical statements. In contrast, another line of work focuses on ***specific logical relations***, uncovering targeted LLM reasoning failures, which requires ***purpose-built datasets*** for quantitative analysis at scale. Using this approach, studies reveal LLM weaknesses in specific types of logic such as converse binary relations (Qi et al., 2023), syllogistic reasoning (Ando et al., 2023), causal inference (Joshi et al., 2024), and even shallow yes/no questions (Clark et al., 2019). Those weaknesses appear as both **fundamental** inabilities in reasoning with certain logic, and **limitations** in specific corresponding downstream applications: more complexities are added by testing divergences between factual inference and logical entailment (Chan et al., 2024), or putting causal reasoning in contexts (Zhao et al., 2024d). To scale up, some synthetically generate natural language data from symbolic templates (Wan et al., 2024; Wang et al., 2024; Gui et al., 2024). Alternatively, Chen et al. (2024d) seed known failures and leverage LLMs to synthetically expand the dataset. While root causes are harder to isolate for those specific logic, the curated datasets offer a natural mitigation by direct fine-tuning.

4.2 Logic in Benchmarks

While Section 4.1 studies LLM reasoning failures directly within natural language logic, another growing body of work ***leverages logical structures implicit in benchmarks to systematically uncover robustness issues in LLM reasoning***. Motivated by rising concerns about the reliability of static benchmarks (Zhou et al., 2023c; Zheng et al., 2024b; Xu et al., 2024a; Patel et al., 2021), these studies introduce ***logic-preserving*** transformations based on particular task structures, such as reordering options in multiple-choice questions (MCQs) (Zheng et al., 2023; Pezeshkpour & Hruschka, 2023; Alzahrani et al., 2024; Gupta et al., 2024; Ni et al., 2024), rearranging parallel premises and events (Chen et al., 2024c; Yamin et al., 2024), or superficially editing unimportant contexts (e.g., character names) (Jiang et al., 2024b; Mirzadeh et al., 2024; Shi et al., 2023; Wang & Zhao, 2024). Such modifications keep the tasks semantically the same. Performance drops thus point to reduced trustworthiness and reveal critical **robustness** issues: despite strong static benchmark scores, the model's reasoning must remain *consistent* on the reasoning tasks being tested.

Math Word Problem (MWP) Benchmarks. Certain benchmarks inherently possess richer logical structures that facilitate targeted perturbations. MWPs exemplify this, as their logic can be readily abstracted into reusable templates. Researchers use this property to generate variants by sampling numeric values (Gulati et al., 2024; Qian et al., 2024; Li et al., 2024b) or substituting irrelevant entities (Shi et al., 2023; Mirzadeh et al., 2024). Structural transformations – such as exchanging known and unknown components (Deb et al., 2024; Guo et al., 2024a) or applying small alterations that change the logic needed to solve problems (Huang et al., 2025b) – further highlight deeper robustness limitations.

Coding Benchmarks. Another example is coding benchmarks, which ask to generate code snippets based on function definitions, doc strings specifying coding tasks, and optional starter code. Common transformations include syntactically editing doc strings (Xia et al., 2024; Wang et al., 2022; Sarker et al., 2024; Roh et al., 2025), renaming functions and variables (Wang et al., 2022; Hooda et al., 2024), and altering control-flow logic such as swapping *if-else* cases (Hooda et al., 2024). Beyond preserving the task logic, some studies introduce adversarial code changes to test whether LLMs identify and adapt to them (Miceli-Barone et al., 2023; Dinh et al., 2023), thereby evaluating deeper reliability. Beyond perturbations, a rising approach utilizes meta-theorems such as the Monadic Second-Order logic from CS theory to synthesize algorithmic coding problems at scale (Beniamini et al., 2025), posing a significant challenge even for state-of-the-art large reasoning models (LRMs) (Xu et al., 2025a).

Mitigation & Extensions. These failures are attributed to ***a lack of robustness*** or ***overfitting to public datasets***. Robustness-related issues are commonly mitigated by applying perturbations to diversify training data (Patel et al., 2021), thus enhancing resilience to variations. Though effective, these approaches are expensive in compute and limited in domain, making them hard to generalize. Overfitting concerns are addressed through dynamically evolving (Jain et al., 2024a; White et al., 2024) or privately maintained datasets (Rajore et al., 2024). They ensure rigorous evaluation, a necessary first step for steering LLM improvement toward better reasoning in the benchmark subjects.

Beyond *individual* benchmarks, Hong et al. (2024) automates a set of transformations across math and coding benchmarks, and Wu et al. (2024e) alters common assumptions of well-known tasks. Shojaee et al. (2025) further moves beyond standard math and coding benchmarks – which assess models solely by final-answer accuracy – by evaluating them on logic puzzles like the Tower of Hanoi, where both reasoning steps and solutions can be systematically assessed. The study finds that even state-of-the-art LRMs suffer an "accuracy collapse" as puzzle complexity increases, though Lawsen (2025) criticizes aspects of the experimental design, suggesting these may unfairly impact reported accuracy.

4.3 Arithmetic & Mathematics

Mathematics, historically a universal framework for rigorous reasoning (Shoenfield, 2018), has exposed fundamental limits in LLM reasoning, particularly in arithmetic-related tasks.

Counting. Despite its simplicity, counting poses a notable fundamental challenge for LLMs (Xu & Ma, 2024; Chang & Bisk, 2024; Zhang & He, 2024; Fu et al., 2024; Conde et al., 2025; Yehudai et al., 2024), even the reasoning ones (Malek et al., 2025), which extend to basic character-level operations like reordering or replacement (Shin & Kaneko, 2024) and affect a wide range of downstream reasoning applications (Vo et al., 2025; Guo et al., 2025b; Parcalabescu et al., 2021). Although the failures manifest at the application level, much work suggest that they originate primarily from architectural and representational limits, including tokenization (Zhang et al., 2024f; Shin & Kaneko, 2024), positional encoding (Chang & Bisk, 2024), and training data composition (Allen-Zhu & Li, 2024), rather than from superficial prompting or task framing on the application-level. Mitigation via supervised fine-tuning (Zhang & He, 2024) and engaged reasoning (Xu & Ma, 2024) have been proposed, yet robust counting remains elusive for current models. Since the limitations largely arise from current LLM architectures, future work should consider deeper mitigation through architectural innovations.

Basic Arithmetic. Another fundamental failure is that LLMs quickly fail in arithmetic as operands increase (Yuan et al., 2023; Testolin, 2024), especially in ***multiplication***. Research shows models rely on superficial pattern-matching rather than arithmetic algorithms, thus struggling notably in middle-digits (Deng et al., 2024). Surprisingly, LLMs fail at simpler tasks (determining the last digit) but succeed in harder ones (first digit identification) (Gambardella et al., 2024). Those fundamental inconsistencies lead to failures for practical tasks like temporal reasoning (Su et al., 2024).

These issues stem from heuristic-driven reasoning strategies (Nikankin et al., 2024) and limited numerical precision (Feng et al., 2024a). Proposed solutions include detailed step-by-step training datasets (Yang et al., 2023c), digit-order reversals to focus attention on least significant digits – mirroring human multiplication strategies (Zhang-Li et al., 2024; Shen et al., 2024), LLM self-improvement methods (Lee et al., 2025), and neuro-symbolic augmentations that enable internal arithmetic reasoning (Dugan et al., 2024). Despite these advances, fundamental research on intrinsic arithmetic capabilities is increasingly overshadowed by the prevalent reliance on external tool use.

Math Word Problems & Beyond. Beyond counting and basic arithmetic – two fundamental failures that propagate into many downstream reasoning applications – Math Word Problems (MWPs) represent a more specific yet highly consequential application domain. Math Word Problems (MWPs) combine arithmetic with contextual logical reasoning, making them a prominent application for assessing LLM capabilities. Beyond using transformations to expose reasoning flaws (Section 4.2), research identifies challenges ranging from specific simple tasks (Nezhurina et al., 2024) to large-scale evaluations on a domain of math (Wei et al., 2023b; Boye & Moell, 2025; Fan et al., 2024; Sun et al., 2025b). Additionally, LLMs exhibit susceptibility when faced with unsolvable or faulty MWPs (Ma et al., 2024a; Rahman et al., 2024; Tian et al., 2024). LLMs struggle even in ***assessing*** reasoning process on MWPs (Zhang et al., 2024g), an arguably easier task than ***generation***. Given these persistent challenges, current efforts in MWPs prioritize developing general methods to improve overall reasoning performance rather than investigating and addressing individual failures.

5 Reasoning in Embodied Environments

Reasoning is not merely an abstract process; it is ***deeply grounded in reality*** (Shapiro & Spaulding, 2024), requiring the ability to perceive, interpret, predict, and act within the physical world, with accurate understanding of spatial relationships, object dynamics, and physical laws (Lee-Cultura & Giannakos, 2020). While humans (Varela et al., 2017) – and even many animals (Andrews & Monsó, 2021) – develop such embodied reasoning naturally through sensory and motor experiences, LLMs remain fundamentally limited by their lack of true physical grounding in the physical world. This gap leads to systematic errors and unrealistic predictions when LLMs attempt even basic physical reasoning (Wang et al., 2023c; Ghaffari & Krishnaswamy, 2024b). Despite growing interest in spatial intelligence, research into LLMs' physical reasoning failures is still sparse. In this section, we survey failures across three progressively complex embodied reasoning modalities: (1) 1D text-based, (2) 2D perception-based, and (3) 3D real-world physical reasoning.

5.1 1D – Text-Based Physical Reasoning Failures

Text-Based Physical Commonsense Reasoning. Physical commonsense reasoning refers to the intuitive understanding of how objects interact in the physical world. Failures of LLMs include lack of knowledge about ***object attributes*** (e.g., size, weight, softness) (Wang et al., 2023c; Liu et al., 2022b; Shu et al., 2023;

Kondo et al., 2023), ***spatial relationships*** (e.g., above, inside, next to) (Liu et al., 2022b; Shu et al., 2023; Kondo et al., 2023), simple physical laws (e.g., gravity, motion, and force) (Gregorcic & Pendrill, 2023), and object affordance (possible actions/reactions an object can make) (Aroca-Ouellette et al., 2021; Adak et al., 2024; Pensa et al., 2024). Humans acquire this kind of reasoning effortlessly through embodied experience, whereas LLMs struggle in it, as they rely solely on textual data without direct perceptual or embodied experience. Even in purely text-based settings, when tasks require more than semantic comprehension, demanding real-world understanding, LLMs exhibit systematic failures. These failures are**fundamental** of current LLMs. While their architectures and training paradigms support impressive language-based learning, they lack the physical grounding.

Physics & Scientific Reasoning. Beyond basic physical commonsense, LLMs struggle with formal physics reasoning and scientific problem-solving, which require not just factual recall and intuition but multi-step logical deduction, quantitative reasoning, and correct use of physical laws – areas where even state-of-the-art models like o1 (Jaech et al., 2024) and o3-mini (OpenAI, 2025) have notable deficits (Zhang et al., 2025a; Xu et al., 2025b; Gupta, 2023; Chung et al., 2025; Zhang et al., 2025b; Qiu et al., 2025). Notably, even when LLMs possess these scientific skills, they often fail to *apply* them effectively in complex problems and real-world scientific discovery (Jaiswal et al., 2024; Ouyang et al., 2023; Chen et al., 2025). Failure in this results in strong **limitations** in LLM's application in scientific domains.

Text-Based Mitigation. These failures largely reflect limitations inherent to the text modality, since semantic and linguistic understanding alone cannot guarantee grounded physical insight (Wang et al., 2023c; Zhang et al., 2025b). Text-based mitigation strategies focus on three fronts: training, prompting, and integration with external tools. First, LLMs are fine-tuned on corpora that explicitly encode structured physical knowledge – such as object attributes, spatial relationships, or physical laws – to better align model priors with real-world dynamics (Lyu et al., 2024; Wang et al., 2023c). Second, prompting methods like CoT encourage models to reason explicitly, reducing reliance on shallow text-based pattern-matching and enabling discovery of more nuanced causal and spatial relationships (Wei et al., 2022b; Ding et al., 2023). Third, LLMs are increasingly paired with external tools – such as code executors or physics engines – that allow models to verify, simulate, or compute outcomes directly and tangibly (Ma et al., 2024c; Cherian et al., 2024).

5.2 2D – Perception-Based Physical Reasoning Failures

What's Wrong with the Picture? The classic "What's Wrong with the Picture?" visual reasoning game challenges participants to spot anomalies in static images. Applied to vision-language models (VLMs), similar tasks reveal surprising failures in simple tasks such as anomaly detection (Bitton-Guetta et al., 2023; Zhou et al., 2023b), object counting and overlap identification (Rahmanzadehgervi et al., 2024), and spatial relation understanding from the image content (Liu et al., 2023a; Zhao et al., 2024a). These failures constitute key perception-related **limitations** and **robustness** vulnerabilities.

2D Physics and Physical Commonsense. Extending beyond detecting simple anomalies or object properties in static images, VLMs face deeper challenges reasoning about the physics in visual contexts. Despite the addition of visual inputs, VLMs still struggle with physical commonsense (Li et al., 2024d; Ghaffari & Krishnaswamy, 2024a; Schulze Buschoff et al., 2025; Dagan et al., 2023; Balazadeh et al., 2024b; Chow et al., 2025; Bear et al., 2021; Xu et al., 2025c) and advanced physics (Ates et al., 2020; Anand et al., 2024; Shen et al., 2025), exhibiting performance gaps similar to those seen in text-only settings discussed in Section 5.1. Similar to the 1D setting, these weaknesses reflect **fundamental** failures of current models and lead to significant **limitations** in applying them to scientific and perception-based domains.

Visual Input for Spatial Reasoning. Real-world spatial reasoning requires understanding *evolving spatial relationships* rather than isolated snapshots. Recent works use 2D simulated environments to test models' grasp of ***motion and object interactions*** (e.g., predicting post-impact trajectories) (Cherian et al., 2024), ***spatial prediction and manipulation*** (e.g., object placement for stability) (Ghaffari & Krishnaswamy, 2024b), ***spatial communication and alignment*** (e.g., conveying location information) (Kar et al., 2025), and ***embodied planning*** in multi-step tasks (Chia et al., 2024; Paglieri et al., 2024; Xu et al., 2025c). While VLMs exhibit some basic spatial knowledge, they consistently fail to compose and apply it in dynamic, interactive tasks, revealing a gap in structured spatial reasoning. This failure is an indication of **limitations** on 2D relevant applications.

Perception-Based Mitigation. These errors arise from three key sources. First, models often over-rely on text or common scenarios from their training data, rather than accurately interpreting visual inputs (Deng et al., 2025a; Bitton-Guetta et al., 2023; Zhou et al., 2023b). Second, some failures may be explained by the binding problem from cognitive science, where the brain – or a model – struggles to process multiple distinct objects simultaneously due to limited shared resources (Campbell et al., 2025). Third, just as language alone does not guarantee grounded physical understanding, visual inputs alone may also lack sufficient spatial semantics; plain image recognition does not automatically confer an understanding of spatial object dynamics and causality (Chen et al., 2024a; Qi et al., 2025). To mitigate, recent work focuses on curating balanced, augmented datasets to reduce bias toward text inputs, or directly using 2D physics data to improve physical understanding (Deng et al., 2025a; Balazadeh et al., 2024a). Another strategy targets training and model architecture (Cheng et al., 2024), by introducing spatially grounded, sequential attention mechanisms (Izadi et al., 2025) and leveraging reinforcement learning to align models with spatial commonsense (Sarch et al., 2025). Finally, beyond end-to-end learning, integration with external physical simulation tools has also emerged, to enable explicit trial-and-error (Liu et al., 2022a; Cherian et al., 2024; Zhu et al., 2025).

5.3 3D – Real-World Physical Reasoning Failures

Real embodied reasoning requires agents to actively interact with their environment, through robotics or interactive simulations that go beyond static images or simple 2D snapshots. Such agents must process real-time goals and feedback, and execute physical actions. Unlike 1D (text-only) and 2D (image-based) tasks, 3D embodied reasoning centers on ***action*** rather than passive analysis. Despite advances in robotics and embodied AI, LLMs and VLMs face persistent challenges including inaccurate spatial modeling, unrealistic affordance prediction, tool-use failures, and unsafe behaviors. This subsection highlights these key failure cases from both simulated and real-world studies.

Real-World Failures in Affordance and Planning. A key failure is models' inability to generate feasible and coherent action plans. LLMs and VLMs often produce physically impossible or inefficient actions due to affordance errors (incorrect reasoning about possible object actions) (Ahn et al., 2022; Li et al., 2025; Hu et al., 2024; Huang et al., 2022a; Jin et al., 2024) and causal real-world reasoning limitations that cause illogical or looping behaviors (Jin et al., 2024; Hu et al., 2024). These failures indicate important application-specific **limitations** for safety-aware and flexible real-world applications, and they often lead to significant **robustness** vulnerabilities.

Spatial and Tool-Use Reasoning. Even when LLMs successfully decompose tasks and generate seemingly valid plans, failures arise due to poor spatial reasoning (Dao & Vu, 2025; Mecattaf et al., 2024) and the inability to generalize tool-use strategies (Xu et al., 2023a). Concretely, LLMs often struggle with 3D distance estimation (Mecattaf et al., 2024; Chen et al., 2024a), object localization (Mecattaf et al., 2024), and multi-step manipulation (Guran et al., 2024), leading to systematic failures in both spatial awareness and interaction with physical environments. Failures in this is an indication of **limitations** in trustworthy, flexible real-world application, and result in robustness vulnerability.

Safety and Long-Term Autonomy. Safety and reliability of LLM-driven embodied agents are ongoing concerns. LLM-generated robotic task plans are highly sensitive to prompt phrasing (Liang et al., 2023) and vulnerable to adversarial manipulation (Zhang et al., 2024c). Moreover, these systems fail to align with human ethical requirements and are easily jailbroken to perform harmful actions, such as recording private information (Rezaei et al., 2025; Zhang et al., 2024c). These findings on **limitations** and **robustness** concerns underscore the urgent need for robust, self-correcting, and safety-aware embodied AI systems before real-world deployment.

Embodied Mitigation. A critical factor underlying these failures is the auto-regressive nature of LLMs. Naive LLMs and VLMs generate plans step by step, lacking mechanisms to detect and correct earlier mistakes or execution errors (Liang et al., 2023; Huang et al., 2022b; Duan et al., 2024). Incorporating feedback mechanisms or explicit error-handling strategies significantly reduces these errors (Liang et al., 2023; Wang et al., 2023a). Another major factor is the absence of a robust *internal world model* (Dao & Vu, 2025; Wu et al., 2025a), which often forces LLMs to rely on external aids – such as explicit spatial prompts – to compensate for their limited spatial and real-world reasoning. To advance embodied intelligence, future research should focus on strengthening LLMs' internal representations of space, including spatial memory, real-world causal dynamics, and quantitative spatial understanding.

6 Discussions & Conclusion

Along the Failure Axis. While our main taxonomy organizes failures by *reasoning type*, examining them along the complementary *failure axis* reveals cross-cutting patterns. **Fundamental failures** – stemming from intrinsic architectural or training constraints – manifest across all reasoning types. For example, the *reversal curse* (Section 4.1), cognitive biases such as confirmation bias (Section 3.1), and working memory limitations that cause proactive interference (Section 3.1) appear in informal reasoning, formal logic, and embodied settings alike. Root cause analysis in those categories are particularly rich, suggesting meaningful methods not only for mitigating the specific failures, but for generally improving the architecture and our understanding of it. **Application-specific limitations** cluster in certain domains: Theory of Mind instability in implicit social reasoning (Section 3.2), inability to generalize to novel Math Word Problem structures in formal reasoning (Section 4.2), or systematic affordance prediction errors in 3D embodied reasoning (Section 5.3). These typically require domain-specific mitigation strategies, such as integrating physics simulators for embodied tasks or symbolic augmentation for mathematics. Tracing the failure cases back to fundamental elements in data or architecture has, on the other hand, attracted less attention from existing literature. **Robustness issues** cut across domains but are particularly well-studied in benchmark-based evaluations (Section 4.2) and social reasoning (Section 3.2, where minor, semantically-preserving perturbations – such as reordering options in multiple-choice questions, renaming variables in code, or paraphrasing moral dilemmas – can lead to large and inconsistent shifts in model outputs). Approaches to detect robustness issues largely revolve around applying such perturbations at scale, often automatically, to stress-test model stability. This perturbation-based paradigm has proven transferable across domains, from coding benchmarks to ToM evaluations, suggesting its utility as a unified detection methodology.

Suggestions for Future Directions. Our survey highlights several gaps and opportunities. First, **root cause analyses** remain incomplete for some failures, including compositional reasoning breakdowns (Section 4.1), higher-order ToM failures (Section 3.2), physical commonsense gaps in 2D and 3D environments (Sections 5.2, 5.3), and brittle multi-agent planning (Section 3.3). Bridging these requires connecting behavioral errors to specific internal mechanisms, e.g., faulty attention head coordination or insufficient intermediate representation alignment. Second, the field would benefit from **unified, persistent failure benchmarks** that span all failure types, akin to the very recent effort Malek et al. (2025), updated regularly to test the *latest* general-purpose and reasoning-specialized models. Such benchmarks should preserve historically challenging cases while incorporating newly discovered ones, enabling longitudinal tracking of failure persistence. Third, **failure-injection principles** could be applied not only to dedicated robustness benchmarks but also to general reasoning benchmarks – by adding adversarial sections, multi-level task difficulty, or cross-domain compositions designed to trigger known weaknesses. Fourth, **dynamic and event-driven benchmarks** could combat overfitting and encourage continual improvement. Promising strategies include (i) (partially) private benchmarks (Phan et al., 2025; Rajore et al., 2024; Zhang et al., 2024d), (ii) dynamically evolving suites (Jain et al., 2024a; White et al., 2024; Zheng et al., 2025), and (iii) adapting regularly occurring events into benchmarks, such as annual competitions (e.g., AIMO [2] for mathematical reasoning), which naturally provide fresh, unseen evaluation items. In combination, these approaches would make reasoning evaluation both more comprehensive and more resistant to short-term overfitting.

A Broad Picture. Admittedly, existing literature, and therefore this survey, may over-represent certain reasoning or failure types, leaving some areas less explored. In particular, multi-turn and interactive contexts remain closer to real-world deployment conditions but are underrepresented in current literature; persistent coordination breakdowns in multi-agent simulations (Section 3.3) illustrate the complexity and significance of these scenarios. Future work should expand benchmark diversity to better capture reasoning failures that arise in such realistic, interactive settings. Overall, the systematic study of reasoning failures in LLMs parallels fault-tolerance research in early computing and incident analysis in safety-critical industries: understanding and categorizing failure is a prerequisite for building resilient systems. By unifying fragmented observations into a structured, two-axis taxonomy, this survey lays a foundation for a mature subfield dedicated to anticipating, detecting, and mitigating reasoning failures. As reasoning-specialized models become more prevalent, sustained attention to failure modes will be essential to ensure that future LLMs not only perform better in reasoning tasks, but *fail better* (gracefully, transparently, recoverably).

[2]AIMO Prize: `https://aimoprize.com/`.

References

Sayantan Adak, Daivik Agrawal, Animesh Mukherjee, and Somak Aditya. Text2afford: Probing object affordance prediction abilities of language models solely from text. *arXiv preprint arXiv:2402.12881*, 2024.

Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. Ethical reasoning and moral value alignment of llms depend on the language we prompt them in. *arXiv preprint arXiv:2404.18460*, 2024.

Saaket Agashe, Yue Fan, Anthony Reyna, and Xin Eric Wang. Llm-coordination: Evaluating and analyzing multi-agent coordination abilities in large language models, 2024. URL https://arxiv.org/abs/2310.03903.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction, 2024. URL https://arxiv.org/abs/2309.14316.

Scott T Allison and David M Messick. The group attribution error. *Journal of Experimental Social Psychology*, 21(6):563–579, 1985.

Guilherme F.C.F. Almeida, José Luiz Nunes, Neele Engelmann, Alex Wiegmann, and Marcelo de Araújo. Exploring the psychology of llms' moral and legal reasoning. *Artificial Intelligence*, 333:104145, August 2024. ISSN 0004-3702. doi: 10.1016/j.artint.2024.104145. URL http://dx.doi.org/10.1016/j.artint.2024.104145.

Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Altwairesh, Areeb Alowisheq, et al. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. *arXiv preprint arXiv:2402.01781*, 2024.

Maryam Amirizaniani, Elias Martin, Maryna Sivachenko, Afra Mashhadi, and Chirag Shah. Can llms reason like humans? assessing theory of mind reasoning in llms for open-ended questions. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 34–44, 2024a.

Maryam Amirizaniani, Elias Martin, Maryna Sivachenko, Afra Mashhadi, and Chirag Shah. Do llms exhibit human-like reasoning? evaluating theory of mind in llms for open-ended responses. *arXiv preprint arXiv:2406.05659*, 2024b.

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. Learning from mistakes makes llm better reasoner, 2024. URL https://arxiv.org/abs/2310.20689.

Avinash Anand, Janak Kapuriya, Apoorv Singh, Jay Saraf, Naman Lal, Astha Verma, Rushali Gupta, and Rajiv Shah. Mm-phyqa: Multimodal physics question-answering with multi-image cot prompting. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 53–64. Springer, 2024.

Risako Ando, Takanobu Morishita, Hirohiko Abe, Koji Mineshima, and Mitsuhiro Okada. Evaluating large language models with neubaroco: Syllogistic reasoning ability and human-like biases, 2023. URL https://arxiv.org/abs/2306.12567.

Kristin Andrews and Susana Monsó. Animal Cognition. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2021 edition, 2021.

Stéphane Aroca-Ouellette, Cory Paik, Alessandro Roncone, and Katharina Kann. Prost: Physical reasoning about objects through space and time. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4597–4608, 2021. doi: 10.18653/v1/2021.findings-acl.404.

Tayfun Ates, M Samil Atesoglu, Cagatay Yigit, Ilker Kesen, Mert Kobas, Erkut Erdem, Aykut Erdem, Tilbe Goksun, and Deniz Yuret. Craft: A benchmark for causal reasoning about forces and interactions. *arXiv preprint arXiv:2012.04293*, 2020.

Alan Baddeley. Working memory. *Memory*, pp. 71–111, 2020.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. Measuring implicit bias in explicitly unbiased large language models. *arXiv preprint arXiv:2402.04105*, 2024.

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

Vahid Balazadeh, Mohammadmehdi Ataei, Hyunmin Cheong, Amir Hosein Khasahmadi, and Rahul G Krishnan. Physics context builders: A modular framework for physical reasoning in vision-language models. *arXiv preprint arXiv:2412.08619*, 2024a.

Vahid Balazadeh, Mohammadmehdi Ataei, Hyunmin Cheong, Amir Hosein Khasahmadi, and Rahul G Krishnan. Synthetic vision: Training vision-language models to understand physics. *arXiv preprint arXiv:2412.08619*, 2024b.

Logan Barnhart, Reza Akbarian Bafghi, Stephen Becker, and Maziar Raissi. Aligning to what? limits to rlhf based alignment, 2025. URL https://arxiv.org/abs/2503.09025.

Lawrence W Barsalou. Grounded cognition. *Annu. Rev. Psychol.*, 59(1):617–645, 2008.

Daniel M Bear, Elias Wang, Damian Mrowca, Felix J Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021.

Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, Hoda Heidari, Anson Ho, Sayash Kapoor, Leila Khalatbari, Shayne Longpre, Sam Manning, Vasilios Mavroudis, Mantas Mazeika, Julian Michael, Jessica Newman, Kwan Yee Ng, Chinasa T. Okolo, Deborah Raji, Girish Sastry, Elizabeth Seger, Theodora Skeadas, Tobin South, Emma Strubell, Florian Tramèr, Lucia Velasco, Nicole Wheeler, Daron Acemoglu, Olubayo Adekanmbi, David Dalrymple, Thomas G. Dietterich, Edward W. Felten, Pascale Fung, Pierre-Olivier Gourinchas, Fredrik Heintz, Geoffrey Hinton, Nick Jennings, Andreas Krause, Susan Leavy, Percy Liang, Teresa Ludermir, Vidushi Marda, Helen Margetts, John McDermid, Jane Munga, Arvind Narayanan, Alondra Nelson, Clara Neppel, Alice Oh, Gopal Ramchurn, Stuart Russell, Marietje Schaake, Bernhard Schölkopf, Dawn Song, Alvaro Soto, Lee Tiedrich, Gaël Varoquaux, Andrew Yao, Ya-Qin Zhang, Fahad Albalawi, Marwan Alserkal, Olubunmi Ajala, Guillaume Avrin, Christian Busch, André Carlos Ponce de Leon Ferreira de Carvalho, Bronwyn Fox, Amandeep Singh Gill, Ahmet Halit Hatip, Juha Heikkilä, Gill Jolly, Ziv Katzir, Hiroaki Kitano, Antonio Krüger, Chris Johnson, Saif M. Khan, Kyoung Mu Lee, Dominic Vincent Ligot, Oleksii Molchanovskyi, Andrea Monti, Nusu Mwamanzi, Mona Nemer, Nuria Oliver, José Ramón López Portillo, Balaraman Ravindran, Raquel Pezoa Rivera, Hammam Riza, Crystal Rugege, Ciarán Seoighe, Jerry Sheehan, Haroon Sheikh, Denise Wong, and Yi Zeng. International ai safety report, 2025. URL https://arxiv.org/abs/2501.17805.

Gal Beniamini, Yuval Dor, Alon Vinnikov, Shir Granot Peled, Or Weinstein, Or Sharir, Noam Wies, Tomer Nussbaum, Ido Ben Shaul, Tomer Zekharya, Yoav Levine, Shai Shalev-Shwartz, and Amnon Shashua. Formulaone: Measuring the depth of algorithmic reasoning beyond competitive programming, 2025. URL https://arxiv.org/abs/2507.13337.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv preprint arXiv:2309.12288*, 2023.

Apratim Bhattacharyya, Sunny Panchal, Mingu Lee, Reza Pourreza, Pulkit Madan, and Roland Memisevic. Look, remember and reason: Grounded reasoning in videos with language models, 2024. URL https://arxiv.org/abs/2306.17778.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pp. 1493–1504. ACM, June 2023. doi: 10.1145/3593013.3594095. URL http://dx.doi.org/10.1145/3593013.3594095.

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional imagesbreaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2616–2627, 2023.

Vamshi Krishna Bonagiri, Sreeram Vennam, Manas Gaur, and Ponnurangam Kumaraguru. Measuring moral inconsistencies in large language models. *arXiv preprint arXiv:2402.01719*, 2024.

Angana Borah and Rada Mihalcea. Towards implicit bias detection and mitigation in multi-agent llm interactions. *arXiv preprint arXiv:2410.02584*, 2024.

Ali Borji. A categorical archive of chatgpt failures, 2023. URL https://arxiv.org/abs/2302.03494.

Johan Boye and Birger Moell. Large language models and mathematical reasoning failures, 2025. URL https://arxiv.org/abs/2502.11574.

Peter G. Brodeur, Thomas A. Buckley, Zahir Kanjee, Ethan Goh, Evelyn Bin Ling, Priyank Jain, Stephanie Cabral, Raja-Elie Abdulnour, Adrian Haimovich, Jason A. Freed, Andrew Olson, Daniel J. Morgan, Jason Hom, Robert Gallo, Eric Horvitz, Jonathan Chen, Arjun K. Manrai, and Adam Rodman. Superhuman performance of a large language model on the reasoning tasks of a physician, 2024. URL https://arxiv.org/abs/2412.10849.

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.

Tianchi Cai, Xierui Song, Jiyan Jiang, Fei Teng, Jinjie Gu, and Guannan Zhang. Ulma: Unified language model alignment with human demonstration and point-wise preference, 2024. URL https://arxiv.org/abs/2312.02554.

Declan Campbell, Sunayana Rane, Tyler Giallanza, Camillo Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven Frankland, Tom Griffiths, Jonathan D Cohen, et al. Understanding the limits of vision language models through the lens of the binding problem. *Advances in Neural Information Processing Systems*, 37:113436–113460, 2025.

Jose J Canas, Inmaculada Fajardo, and Ladislao Salmeron. Cognitive flexibility. *International encyclopedia of ergonomics and human factors*, 1(3):297–301, 2006.

Mark D. Cannon and Amy C. Edmondson. Failing to learn and learning to fail (intelligently): How great organizations put failure to work to innovate and improve. *Long Range Planning*, 38(3):299–319, 2005. ISSN 0024-6301. doi: https://doi.org/10.1016/j.lrp.2005.04.005. URL https://www.sciencedirect.com/science/article/pii/S0024630105000580. Organizational Failure.

Riccardo Cantini, Giada Cosenza, Alessio Orsino, and Domenico Talia. Are large language models really bias-free? jailbreak prompts for assessing adversarial robustness to bias elicitation. In *International Conference on Discovery Science*, pp. 52–68. Springer, 2024.

Davide Castelvecchi. Can we open the black box of ai? *Nature News*, 538(7623):20, 2016.

Mohna Chakraborty, Lu Wang, and David Jurgens. Structured moral reasoning in language models: A value-grounded evaluation framework, 2025. URL https://arxiv.org/abs/2506.14948.

Jason Chan, Robert Gaizauskas, and Zhixue Zhao. Rulebreakers challenge: Revealing a blind spot in large language models' reasoning with formal logic, 2024. URL https://arxiv.org/abs/2410.16502.

Yingshan Chang and Yonatan Bisk. Language models need inductive biases to count inductively, 2024. URL https://arxiv.org/abs/2405.20131.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024a.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17754–17762, 2024b.

Tingting Chen, Srinivas Anumasa, Beibei Lin, Vedant Shah, Anirudh Goyal, and Dianbo Liu. Auto-bench: An automated benchmark for scientific discovery in llms. *arXiv preprint arXiv:2502.15224*, 2025.

Xinyun Chen, Ryan A. Chi, Xuezhi Wang, and Denny Zhou. Premise order matters in reasoning with large language models, 2024c. URL https://arxiv.org/abs/2402.08939.

Yulong Chen, Yang Liu, Jianhao Yan, Xuefeng Bai, Ming Zhong, Yinghao Yang, Ziyi Yang, Chenguang Zhu, and Yue Zhang. See what llms cannot answer: A self-challenge framework for uncovering llm weaknesses, 2024d. URL https://arxiv.org/abs/2408.08978.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

Anoop Cherian, Radu Corcodel, Siddarth Jain, and Diego Romeres. Llmphy: Complex physical reasoning using large language models and world models. *arXiv preprint arXiv:2411.08027*, 2024.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios, 2023. *URL https://arxiv. org/abs/2307.13528*, 2023.

Yew Ken Chia, Qi Sun, Lidong Bing, and Soujanya Poria. Can-do! a dataset and neuro-symbolic grounded framework for embodied planning with large multimodal models. *arXiv preprint arXiv:2409.14277*, 2024.

Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models, 2023. URL https://arxiv.org/abs/2202.04053.

Georgios Chochlakis, Alexandros Potamianos, Kristina Lerman, and Shrikanth Narayanan. The strong pull of prior knowledge in large language models and its impact on emotion recognition. *arXiv preprint arXiv:2403.17125*, 2024.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*, 2025.

Zhibo Chu, Zichong Wang, and Wenbin Zhang. Fairness in large language models: A taxonomic survey. *ACM SIGKDD explorations newsletter*, 26(1):34–48, 2024.

Daniel JH Chung, Zhiqi Gao, Yurii Kvasiuk, Tianyi Li, Moritz Münchmeyer, Maja Rudolph, Frederic Sala, and Sai Chaitanya Tadepalli. Theoretical physics benchmark (tpbench)–a dataset and study of ai reasoning capabilities in theoretical physics. *arXiv preprint arXiv:2502.15815*, 2025.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions, 2019. URL https://arxiv.org/abs/1905.10044.

Kwesi Cobbina and Tianyi Zhou. Where to show demos in your prompt: A positional bias of in-context learning. *arXiv preprint arXiv:2507.22887*, 2025.

Philip R. P. Coelho and James E. McClure. Learning from Failure. Working Papers 200402, Ball State University, Department of Economics, January 2004. URL https://ideas.repec.org/p/bsu/wpaper/200402.html.

Michelle Cohn, Mahima Pushkarna, Gbolahan O Olanubi, Joseph M Moran, Daniel Padgett, Zion Mengesha, and Courtney Heldreth. Believing anthropomorphism: examining the role of anthropomorphic cues on trust in large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2024.

Javier Conde, Gonzalo Martínez, Pedro Reviriego, Zhen Gao, Shanshan Liu, and Fabrizio Lombardi. Can chatgpt learn to count letters? *Computer*, 58(3):96–99, 2025.

Irving M Copi, Carl Cohen, and Kenneth McMahon. *Introduction to logic*. Routledge, 2016.

Logan Cross, Violet Xiang, Agam Bhatia, Daniel LK Yamins, and Nick Haber. Hypothetical minds: Scaffolding theory of mind for multi-agent tasks with large language models. *arXiv preprint arXiv:2407.07086*, 2024.

Gautier Dagan, Frank Keller, and Alex Lascarides. Learning the effects of physical actions in a multi-modal environment. *arXiv preprint arXiv:2301.11845*, 2023.

Adam Dahlgren Lindström, Leila Methnani, Lea Krause, Petter Ericson, Íñigo Martínez de Rituerto de Troya, Dimitri Coelho Mollo, and Roel Dobbe. Helpful, harmless, honest? sociotechnical limits of ai alignment and safety through reinforcement learning from human feedback: Ad lindström et al. *Ethics and Information Technology*, 27(2):28, 2025.

Alan Dao and Dinh Bach Vu. Alphamaze: Enhancing large language models' spatial intelligence via grpo. *arXiv preprint arXiv:2502.14669*, 2025.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness, 2022. URL https://arxiv.org/abs/2205.14135.

Badhan Chandra Das, M. Hadi Amini, and Yanzhao Wu. Security and privacy challenges of large language models: A survey. *ACM Comput. Surv.*, 57(6), February 2025. ISSN 0360-0300. doi: 10.1145/3712001. URL https://doi.org/10.1145/3712001.

David "davidad" Dalrymple, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, Alessandro Abate, Joe Halpern, Clark Barrett, Ding Zhao, Tan Zhi-Xuan, Jeannette Wing, and Joshua Tenenbaum. Towards guaranteed safe ai: A framework for ensuring robust and reliable ai systems, 2024. URL https://arxiv.org/abs/2405.06624.

Christian Schroeder de Witt. Open challenges in multi-agent security: Towards secure systems of interacting ai agents, 2025. URL https://arxiv.org/abs/2505.02077.

Aniruddha Deb, Neeva Oza, Sarthak Singla, Dinesh Khandelwal, Dinesh Garg, and Parag Singla. Fill in the blank: Exploring and enhancing llm capabilities for backward reasoning in math word problems, 2024. URL https://arxiv.org/abs/2310.01991.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? *arXiv preprint arXiv:2503.02199*, 2025a.

Chunyuan Deng, Zhiqi Li, Roy Xie, Ruidi Chang, and Hanjie Chen. Language models are symbolic learners in arithmetic, 2024. URL https://arxiv.org/abs/2410.15580.

Xun Deng, Sicheng Zhong, Andreas Veneris, Fan Long, and Xujie Si. Verifythisbench: Generating code, specifications, and proofs all at once, 2025b. URL https://arxiv.org/abs/2505.19271.

Soham Deshmukh, Shuo Han, Hazim Bukhari, Benjamin Elizalde, Hannes Gamper, Rita Singh, and Bhiksha Raj. Audio entailment: Assessing deductive reasoning for audio understanding, 2024. URL https://arxiv.org/abs/2407.18062.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Adele Diamond. Executive functions. *Annual review of psychology*, 64(1):135–168, 2013.

Jingzhe Ding, Yan Cen, and Xinyuan Wei. Using large language model to solve and explain physics word problems approaching human level. *arXiv preprint arXiv:2309.08182*, 2023.

Tuan Dinh, Jinman Zhao, Samson Tan, Renato Negrinho, Leonard Lausen, Sheng Zha, and George Karypis. Large language models of code fail at completing code with potential bugs, 2023. URL https://arxiv.org/abs/2306.03438.

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. Lp-musiccaps: Llm-based pseudo music captioning, 2023. URL https://arxiv.org/abs/2307.16372.

Kefan Dong and Tengyu Ma. Stp: Self-play llm theorem provers with iterative conjecturing and proving, 2025. URL https://arxiv.org/abs/2502.00212.

John Dougrez-Lewis, Mahmud Elahi Akhter, Yulan He, and Maria Liakata. Assessing the reasoning abilities of chatgpt in the context of claim verification, 2024. URL https://arxiv.org/abs/2402.10735.

Hubert L. Dreyfus. *What Computers Still Can?T Do: A Critique of Artificial Reason*. MIT Press, 1992.

Jiafei Duan, Wilbert Pumacay, Nishanth Kumar, Yi Ru Wang, Shulin Tian, Wentao Yuan, Ranjay Krishna, Dieter Fox, Ajay Mandlekar, and Yijie Guo. Aha: A vision-language-model for detecting and reasoning over failures in robotic manipulation. *arXiv preprint arXiv:2410.00371*, 2024.

Philipp Dufter, Martin Schmitt, and Hinrich Schütze. Position information in transformers: An overview. *Computational Linguistics*, 48(3):733–763, 2022.

Owen Dugan, Donato Manuel Jimenez Beneto, Charlotte Loh, Zhuo Chen, Rumen Dangovski, and Marin Soljačić. Occamllm: Fast and exact language model arithmetic in a single step, 2024. URL https://arxiv.org/abs/2406.06576.

Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33, 2023.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality, 2023. URL https://arxiv.org/abs/2305.18654.

Jessica Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. Cognitive bias in high-stakes decision-making with llms. *arXiv preprint arXiv:2403.00811*, 2024.

Daniel Enström, Viktor Kjellberg, and Moa Johansson. Reasoning in transformers – mitigating spurious correlations and reasoning shortcuts, 2024. URL https://arxiv.org/abs/2403.11314.

Jingxuan Fan, Sarah Martinson, Erik Y. Wang, Kaylie Hausknecht, Jonah Brenner, Danxian Liu, Nianli Peng, Corey Wang, and Michael P. Brenner. Hardmath: A benchmark dataset for challenging problems in applied mathematics, 2024. URL https://arxiv.org/abs/2410.09988.

Evelina Fedorenko, Steven Piantadosi, and Edward Gibson. Language is primarily a tool for communication rather than thought. *Nature*, 630:575–586, 06 2024. doi: 10.1038/s41586-024-07522-w.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition, 2024. URL https://arxiv.org/abs/2501.03230.

Guhao Feng, Kai Yang, Yuntian Gu, Xinyue Ai, Shengjie Luo, Jiacheng Sun, Di He, Zhenguo Li, and Liwei Wang. How numerical precision affects mathematical reasoning capabilities of llms, 2024a. URL https://arxiv.org/abs/2410.13857.

Tao Feng, Pengrui Han, Guanyu Lin, Ge Liu, and Jiaxuan You. Thought-retriever: Don't just retrieve raw data, retrieve thoughts. In *ICLR 2024 Workshop: How Far Are We From AGI*, 2024b.

Chris Frith and Uta Frith. Theory of mind. *Current biology*, 15(17):R644–R645, 2005.

Tairan Fu, Raquel Ferrando, Javier Conde, Carlos Arriaga, and Pedro Reviriego. Why do large language models (llms) struggle to count letters?, 2024. URL https://arxiv.org/abs/2412.18626.

Isaac R Galatzer-Levy, Jed McGiffin, David Munday, Xin Liu, Danny Karmon, Ilia Labzovsky, Rivka Moroshko, Amir Zait, and Daniel McDuff. Evidence of cognitive deficits anddevelopmental advances in generative ai: A clock drawing test analysis. *arXiv preprint arXiv:2410.11756*, 2024.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.

Andrew Gambardella, Yusuke Iwasawa, and Yutaka Matsuo. Language models do hard arithmetic tasks easily and hardly do easy arithmetic tasks, 2024. URL https://arxiv.org/abs/2406.02356.

Kanishk Gandhi, Zoe Lynch, Jan-Philipp Fränken, Kayla Patterson, Sharon Wambu, Tobias Gerstenberg, Desmond C Ong, and Noah D Goodman. Human-like affective cognition in foundation models. *arXiv preprint arXiv:2409.11733*, 2024.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020. URL https://arxiv.org/abs/2101.00027.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2, 2023.

Basile Garcia, Crystal Qian, and Stefano Palminteri. The moral turing test: Evaluating human-llm alignment in moral decision-making. *arXiv preprint arXiv:2410.07304*, 2024.

Josh Gardner, Simon Durand, Daniel Stoller, and Rachel M. Bittner. Llark: A multimodal instruction-following language model for music, 2024. URL https://arxiv.org/abs/2310.07160.

Gaël Gendron, Qiming Bao, Michael Witbrock, and Gillian Dobbie. Large language models are not strong abstract reasoners. *arXiv preprint arXiv:2305.19555*, 2023.

Sadaf Ghaffari and Nikhil Krishnaswamy. Large language models are challenged by habitat-centered reasoning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13047–13059, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.763. URL https://aclanthology.org/2024.findings-emnlp.763.

Sadaf Ghaffari and Nikhil Krishnaswamy. Exploring failure cases in multimodal reasoning about physical dynamics, 2024b. URL https://arxiv.org/abs/2402.15654.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities, 2024. URL https://arxiv.org/abs/2406.11768.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities, 2025. URL https://arxiv.org/abs/2503.03983.

Olga Golovneva, Zeyuan Allen-Zhu, Jason Weston, and Sainbayar Sukhbaatar. Reverse training to nurse the reversal curse, 2024. URL https://arxiv.org/abs/2403.13799.

Dongyu Gong and Hantao Zhang. Self-attention limits working memory capacity of transformer-based models, 2024. URL https://arxiv.org/abs/2409.10715.

Dongyu Gong, Xingchen Wan, and Dingmin Wang. Working memory capacity of chatgpt: An empirical study. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 10048–10056, 2024.

Bor Gregorcic and Ann-Marie Pendrill. Chatgpt and the frustrated socrates. *Physics Education*, 58(3):035021, Mar 2023. doi: 10.1088/1361-6552/acc299.

Rhys Gretsch, Peiyang Song, Advait Madhavan, Jeremy Lau, and Timothy Sherwood. Energy efficient convolutions with temporal arithmetic. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, ASPLOS '24, pp. 354–368, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703850. doi: 10.1145/3620665.3640395. URL https://doi.org/10.1145/3620665.3640395.

Rhys Gretsch, Peiyang Song, Advait Madhavan, Jeremy Lau, and Timothy Sherwood. Delay space arithmetic and architecture. *IEEE Micro*, 2025.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024. URL https://arxiv.org/abs/2312.00752.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. Simpletom: Exposing the gap between explicit tom inference and implicit tom application in llms. *arXiv preprint arXiv:2410.13648*, 2024.

Bryan Guan, Tanya Roosta, Peyman Passban, and Mehdi Rezagholizadeh. The order effect: Investigating prompt sensitivity to input order in llms. *arXiv preprint arXiv:2502.04134*, 2025.

Jiayi Gui, Yiming Liu, Jiale Cheng, Xiaotao Gu, Xiao Liu, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie Huang. Logicgame: Benchmarking rule-based reasoning abilities of large language models, 2024. URL https://arxiv.org/abs/2408.15778.

Hamimah Guinungco and Adriel Roman. Abstract reasoning and problem-solving skills of first year college students. *Southeast Asian Journal of Science and Technology*, 5(1):33–39, 2020.

Aryan Gulati, Brando Miranda, Eric Chen, Emily Xia, Kai Fronsdal, Bruno de Moraes Dumont, and Sanmi Koyejo. Putnam-axiom: A functional and static benchmark for measuring higher level mathematical reasoning, 2024. URL https://openreview.net/forum?id=WrBqgoseGL.

Pei Guo, WangJie You, Juntao Li, Yan Bowen, and Min Zhang. Exploring reversal mathematical reasoning ability for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13671–13685, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.811. URL https://aclanthology.org/2024.findings-acl.811/.

Qingyan Guo, Rui Wang, Junliang Guo, Xu Tan, Jiang Bian, and Yujiu Yang. Mitigating reversal curse in large language models via semantic-aware permutation training, 2024b. URL https://arxiv.org/abs/2403.00758.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024c.

Tianyu Guo, Hanlin Zhu, Ruiqi Zhang, Jiantao Jiao, Song Mei, Michael I. Jordan, and Stuart Russell. How do llms perform two-hop reasoning in context?, 2025a. URL https://arxiv.org/abs/2502.13913.

Xuyang Guo, Zekai Huang, Zhenmei Shi, Zhao Song, and Jiahao Zhang. Your vision-language model can't even count to 20: Exposing the failures of vlms in compositional counting. *arXiv preprint arXiv:2510.04401*, 2025b.

Pranav Gupta. Testing llm performance on the physics gre: some observations. *arXiv preprint arXiv:2312.04613*, 2023.

Vipul Gupta, David Pantoja, Candace Ross, Adina Williams, and Megan Ung. Changing answer order can decrease mmlu accuracy, 2024. URL https://arxiv.org/abs/2406.19470.

Nurhan Bulus Guran, Hanchi Ren, Jingjing Deng, and Xianghua Xie. Task-oriented robotic manipulation with vision language models. *arXiv preprint arXiv:2410.15863*, 2024.

Thilo Hagendorff. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv preprint arXiv:2303.13988*, 1, 2023.

David L Hamilton and Robert K Gifford. Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology*, 12(4):392–407, 1976.

Pengrui Han, Rafal Kocielnik, Adhithya Saravanan, Roy Jiang, Or Sharir, and Anima Anandkumar. Chatgpt based data augmentation for improved parameter-efficient debiasing of llms. *arXiv preprint arXiv:2402.11764*, 2024a.

Pengrui Han, Peiyang Song, Haofei Yu, and Jiaxuan You. In-context learning may not elicit trustworthy reasoning: A-not-B errors in pretrained language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5624–5643, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.322. URL https://aclanthology.org/2024.findings-emnlp.322/.

Pengrui Han, Rafal Dariusz Kocielnik, Peiyang Song, Ramit Debnath, Dean Mobbs, Anima Anandkumar, and R Michael Alvarez. Tracing human-like traits in llms: Origins, real-world manifestation, and controllability. In *2nd Workshop on Models of Human Feedback for AI Alignment*, 2025.

Shanshan Han, Qifan Zhang, Yuhang Yao, Weizhao Jin, Zhaozhuo Xu, and Chaoyang He. Llm multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578*, 2024c.

Guangfu Hao, Frederic Alexandre, and Shan Yu. Visual large language models exhibit human-level cognitive flexibility in the wisconsin card sorting test. *arXiv preprint arXiv:2505.22112*, 2025.

Guozhi Hao, Jun Wu, Qianqian Pan, and Rosario Morello. Quantifying the uncertainty of llm hallucination spreading in complex adaptive social networks. *Scientific reports*, 14(1):16375, 2024.

Ramin Hasani, Mathias Lechner, Alexander Amini, Daniela Rus, and Radu Grosu. Liquid time-constant networks, 2020. URL https://arxiv.org/abs/2006.04439.

Shreya Havaldar, Sunny Rai, Bhumika Singhal, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. Multilingual language models are not multicultural: A case study in emotion. *arXiv preprint arXiv:2307.01370*, 2023.

Jiangen He and Jiqun Liu. Investigating the impact of llm personality on cognitive bias manifestation in automated decision-making tasks. *arXiv preprint arXiv:2502.14219*, 2025.

Yinghui He, Yufan Wu, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. *arXiv preprint arXiv:2310.16755*, 2023.

Chadi Helwe, Chloé Clavel, and Fabian M. Suchanek. Reasoning with transformer-based models: Deep learning, but shallow reasoning. In *Conference on Automated Knowledge Base Construction*, 2021. URL https://api.semanticscholar.org/CorpusID:237397001.

Pengfei Hong, Navonil Majumder, Deepanway Ghosal, Somak Aditya, Rada Mihalcea, and Soujanya Poria. Evaluating llms' mathematical and coding competency through ontology-guided interventions, 2024. URL https://arxiv.org/abs/2401.09395.

Ashish Hooda, Mihai Christodorescu, Miltiadis Allamanis, Aaron Wilson, Kassem Fawaz, and Somesh Jha. Do large code models understand programming concepts? a black-box approach, 2024. URL https://arxiv.org/abs/2402.05980.

Arian Hosseini, Alessandro Sordoni, Daniel Toyama, Aaron Courville, and Rishabh Agarwal. Not all llm reasoners are created equal, 2024. URL https://arxiv.org/abs/2410.01748.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

He Hu, Yucheng Zhou, Lianzhong You, Hongbo Xu, Qianning Wang, Zheng Lian, Fei Richard Yu, Fei Ma, and Laizhong Cui. Emobench-m: Benchmarking emotional intelligence for multimodal large language models. *arXiv preprint arXiv:2502.04424*, 2025.

Zichao Hu, Francesca Lucchetti, Claire Schlesinger, Yash Saxena, Anders Freeman, Sadanand Modak, Arjun Guha, and Joydeep Biswas. Deploying and evaluating llms to program service mobile robots. *IEEE Robotics and Automation Letters*, 9(3):2853–2860, 2024.

Jen-tse Huang, Jiaxu Zhou, Tailin Jin, Xuhui Zhou, Zixi Chen, Wenxuan Wang, Youliang Yuan, Maarten Sap, and Michael R Lyu. On the resilience of multi-agent systems with malicious agents. *arXiv preprint arXiv:2408.00989*, 2024.

Jen-tse Huang, Kaiser Sun, Wenxuan Wang, and Mark Dredze. Llms do not have human-like working memory. *arXiv preprint arXiv:2505.10571*, 2025a.

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey, 2023. URL https://arxiv.org/abs/2212.10403.

Kaixuan Huang, Jiacheng Guo, Zihao Li, Xiang Ji, Jiawei Ge, Wenzhe Li, Yingqing Guo, Tianle Cai, Hui Yuan, Runzhe Wang, Yue Wu, Ming Yin, Shange Tang, Yangsibo Huang, Chi Jin, Xinyun Chen, Chiyuan Zhang, and Mengdi Wang. Math-perturb: Benchmarking llms' math reasoning abilities against hard perturbations, 2025b. URL https://arxiv.org/abs/2502.06453.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55, 2025c.

Suozhi Huang, Peiyang Song, Robert Joseph George, and Anima Anandkumar. Leanprogress: Guiding search for neural theorem proving via proof progress prediction. *arXiv preprint arXiv:2502.17925*, 2025d.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning*, pp. 9118–9147. PMLR, 2022a.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022b.

Yiming Huang, Biquan Bie, Zuqiu Na, Weilin Ruan, Songxin Lei, Yutao Yue, and Xinlei He. An empirical study of the anchoring effect in llms: Existence, mechanism, and potential mitigations. *arXiv preprint arXiv:2505.15392*, 2025e.

Itay Itzhak, Yonatan Belinkov, and Gabriel Stanovsky. Planted in pretraining, swayed by finetuning: A case study on the origins of cognitive biases in llms, 2025. URL https://arxiv.org/abs/2507.07186.

Lucja Iwańska. Logical reasoning in natural language: It is all about knowledge. *Minds and Machines*, 3(4): 475–510, 1993. doi: 10.1007/bf00974107.

Amirmohammad Izadi, Mohammad Ali Banayeeanzade, Fatemeh Askari, Ali Rahimiakbar, Mohammad Mahdi Vahedi, Hosein Hasani, and Mahdieh Soleymani Baghshah. Visual structures helps visual reasoning: Addressing the binding problem in vlms. *arXiv preprint arXiv:2506.22146*, 2025.

Hintikka Jaakko and Gabriel Sandu. What is logic? In Dale Jacquette (ed.), *Philosophy of Logic*, pp. 13–39. North Holland, 2002.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code, 2024a. URL https://arxiv.org/abs/2403.07974.

Shomik Jain, D Calacci, and Ashia Wilson. As an ai language model," yes i would recommend calling the police": Norm inconsistency in llm decision-making. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pp. 624–633, 2024b.

Raj Jaiswal, Dhruv Jain, Harsh Parimal Popat, Avinash Anand, Abhishek Dharmadhikari, Atharva Marathe, and Rajiv Ratn Shah. Improving physics reasoning in large language models using mixture of refinement agents. *arXiv preprint arXiv:2412.00821*, 2024.

Vivekananda Jayaram, Vishnu Ramineni, and Manjunatha Sughaturu Krishnappa. Mitigating order sensitivity in large language models for multiple-choice question tasks. *International Journal of Artificial Intelligence Research and Development (IJAIRD)*, 2024.

Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. Moralbench: Moral evaluation of llms. *arXiv preprint arXiv:2406.04428*, 2024.

Albert Q. Jiang, Wenda Li, and Mateja Jamnik. Multilingual mathematical autoformalization, 2023a. URL https://arxiv.org/abs/2311.03755.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024a. URL https://arxiv.org/abs/2401.04088.

Bowen Jiang, Yangxinyu Xie, Zhuoqun Hao, Xiaomeng Wang, Tanwi Mallick, Weijie J. Su, Camillo J. Taylor, and Dan Roth. A peek into token bias: Large language models are not yet genuine reasoners, 2024b. URL https://arxiv.org/abs/2406.11050.

Chumeng Jiang, Jiayin Wang, Weizhi Ma, Charles LA Clarke, Shuai Wang, Chuhan Wu, and Min Zhang. Beyond utility: Evaluating llm as recommender. In *Proceedings of the ACM on Web Conference 2025*, pp. 3850–3862, 2025a.

Ling Jiang, Keer Jiang, Xiaoyu Chu, Saaransh Gulati, and Pulkit Garg. Hallucination detection in llm-enriched product listings. In *Proceedings of the Seventh Workshop on e-Commerce and NLP@ LREC-COLING 2024*, pp. 29–39, 2024c.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny T Liang, Sydney Levine, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jack Hessel, et al. Investigating machine moral judgement through the delphi experiment. *Nature Machine Intelligence*, pp. 1–16, 2025b.

Roy Jiang, Rafal Kocielnik, Adhithya Prakash Saravanan, Pengrui Han, R Michael Alvarez, and Anima Anandkumar. Empowering domain experts to detect social bias in generative ai with user-friendly interfaces. In *XAI in Action: Past, Present, and Future Applications*, 2023b.

Yixiang Jin, Dingzhe Li, A Yong, Jun Shi, Peng Hao, Fuchun Sun, Jianwei Zhang, and Bin Fang. Robotgpt: Robot manipulation learning from chatgpt. *IEEE Robotics and Automation Letters*, 9(3):2543–2550, 2024.

Zheng Jin, Maurizio Tirassa, and Anna M Borghi. Beyond embodied cognition: Intentionality, affordance, and environmental adaptation, 2018.

Erik Jones and Jacob Steinhardt. Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems*, 35:11785–11799, 2022.

Nitish Joshi, Abulhair Saparov, Yixin Wang, and He He. Llms are prone to fallacies in causal inference, 2024. URL https://arxiv.org/abs/2406.12158.

Nikola Jovanović, Robin Staab, and Martin Vechev. Watermark stealing in large language models, 2024. URL https://arxiv.org/abs/2402.19361.

Robert Kail. *The development of memory in children.* WH Freeman/Times Books/Henry Holt & Co, 1990.

Amlan Kar, David Acuna, and Sanja Fidler. On inherent 3d reasoning of vlms in indoor scene layout design, 2025. URL https://openreview.net/pdf?id=uBhqll8pw1.

Florian Karl, Malte Kemeter, Gabriel Dax, and Paulina Sierak. Position: Embracing negative results in machine learning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23256–23265. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/karl24a.html.

Artem Karpov, Seong Hah Cho, Austin Meek, Raymond Koopmanschap, Lucy Farnik, and Bogdan-Ionut Cirstea. Inducing human-like biases in moral reasoning language models, 2024. URL https://arxiv.org/abs/2411.15386.

Saketh Ram Kasibatla, Arpan Agarwal, Yuriy Brun, Sorin Lerner, Talia Ringer, and Emily First. Cobblestone: Iterative automation for formal verification, 2024. URL https://arxiv.org/abs/2410.19940.

Sean M Kennedy and Robert D Nowak. Cognitive flexibility of large language models. In *ICML 2024 Workshop on LLMs and Cognition*, 2024.

Muhammad Uzair Khattak, Muhammad Ferjad Naeem, Jameel Hassan, Muzammal Naseer, Federico Tombari, Fahad Shahbaz Khan, and Salman Khan. How good is my video lmm? complex video reasoning and robustness evaluation suite for video-lmms, 2024. URL https://arxiv.org/abs/2405.03690.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. Fantom: A benchmark for stress-testing machine theory of mind in interactions. *arXiv preprint arXiv:2310.15421*, 2023.

Kazushi Kondo, Saku Sugawara, and Akiko Aizawa. Probing physical reasoning with counter-commonsense context. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 603–612, 2023. doi: 10.18653/v1/2023.acl-short.53.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012*, 2023.

Michal Kosinski. Evaluating large language models in theory of mind tasks. *arXiv e-prints*, pp. arXiv–2302, 2023.

Divyanshu Kumar, Umang Jain, Sahil Agarwal, and Prashanth Harshangi. Investigating implicit bias in large language models: A large-scale study of over 50 llms. *arXiv preprint arXiv:2410.12864*, 2024.

Adarsh Kumarappan, Mo Tiwari, Peiyang Song, Robert Joseph George, Chaowei Xiao, and Anima Anandkumar. Leanagent: Lifelong learning for formal theorem proving. *arXiv preprint arXiv:2410.06209*, 2024.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality, 2024. URL https://arxiv.org/abs/2305.00050.

Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models, like humans, show content effects on reasoning tasks. *PNAS nexus*, 3(7):pgae233, 2024.

Guillaume Lample, Timothee Lacroix, Marie-Anne Lachaux, Aurelien Rodriguez, Amaury Hayat, Thibaut Lavril, Gabriel Ebner, and Xavier Martinet. Hypertree proof search for neural theorem proving. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 26337–26349. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a8901c5e85fb8e1823bbf0f755053672-Paper-Conference.pdf.

A. Lawsen. Comment on the illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity, 2025. URL https://arxiv.org/abs/2506.09250.

Grant Ledger and Rafael Mancinni. Detecting llm hallucinations using monte carlo simulations on token probabilities. *Authorea Preprints*, 2024.

Nayoung Lee, Ziyang Cai, Avi Schwarzschild, Kangwook Lee, and Dimitris Papailiopoulos. Self-improving transformers overcome easy-to-hard and length generalization challenges, 2025. URL https://arxiv.org/abs/2502.01612.

Serena Lee-Cultura and Michail Giannakos. Embodied interaction and spatial skills: A systematic review of empirical studies. *Interacting with Computers*, 32(4):331–366, 2020.

Clarence Irving Lewis, Cooper Harold Langford, and P Lamprecht. *Symbolic logic*, volume 170. Dover publications New York, 1959.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. Large language models with controllable working memory, 2022. URL https://arxiv.org/abs/2211.05110.

Dongting Li, Chenchong Tang, and Han Liu. Audio-llm: Activating the capabilities of large language models to comprehend audio data. In Xinyi Le and Zhijun Zhang (eds.), *Advances in Neural Networks – ISNN 2024*, pp. 133–142, Singapore, 2024a. Springer Nature Singapore. ISBN 978-981-97-4399-5.

Huao Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Michael Lewis, and Katia Sycara. Theory of mind for multi-agent collaboration via large language models. *arXiv preprint arXiv:2310.10701*, 2023a.

Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Erran Li Li, Ruohan Zhang, et al. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems*, 37:100428–100534, 2025.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. Gsm-plus: A comprehensive benchmark for evaluating the robustness of llms as mathematical problem solvers, 2024b. URL https://arxiv.org/abs/2402.19255.

Wenjun Li, Ying Cai, Ziyang Wu, Wenyi Zhang, Yifan Chen, Rundong Qi, Mengqi Dong, Peigen Chen, Xiao Dong, Fenghao Shi, Lei Guo, Junwei Han, Bao Ge, Tianming Liu, Lin Gan, and Tuo Zhang. A survey of foundation models for music understanding, 2024c. URL https://arxiv.org/abs/2409.09601.

Yijiang Li, Qingying Gao, Tianwei Zhao, Bingyang Wang, Haoran Sun, Haiyun Lyu, Robert D Hawkins, Nuno Vasconcelos, Tal Golan, Dezhi Luo, et al. Core knowledge deficits in multi-modal language models. *arXiv preprint arXiv:2410.10855*, 2024d.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*, 2023b.

Yuxiao Li, Eric J. Michaud, David D. Baek, Joshua Engels, Xiaoqing Sun, and Max Tegmark. The geometry of concepts: Sparse autoencoder feature structure, 2024e. URL https://arxiv.org/abs/2410.19750.

Zhaoyi Li, Gangwei Jiang, Hong Xie, Linqi Song, Defu Lian, and Ying Wei. Understanding and patching compositional reasoning in llms, 2024f. URL https://arxiv.org/abs/2402.14328.

Zhaoyu Li, Jialiang Sun, Logan Murphy, Qidong Su, Zenan Li, Xian Zhang, Kaiyu Yang, and Xujie Si. A survey on deep learning for theorem proving, 2024g. URL https://arxiv.org/abs/2404.09939.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9493–9500. IEEE, 2023.

Kaiqu Liang, Haimin Hu, Xuandong Zhao, Dawn Song, Thomas L. Griffiths, and Jaime Fernández Fisac. Machine bullshit: Characterizing the emergent disregard for truth in large language models, 2025. URL https://arxiv.org/abs/2507.07484.

Jan Malte Lichtenberg, Alexander Buchholz, and Pola Schwöbel. Large language models as recommender systems: A study of popularity bias. *arXiv preprint arXiv:2406.01285*, 2024.

Falk Lieder, Thomas L Griffiths, Quentin J M. Huys, and Noah D Goodman. The anchoring bias reflects rational use of cognitive resources. *Psychonomic bulletin & review*, 25:322–349, 2018.

Guanyu Lin, Tao Feng, Pengrui Han, Ge Liu, and Jiaxuan You. Paper copilot: A self-evolving and efficient llm system for personalized academic assistance. *arXiv preprint arXiv:2409.04593*, 2024a.

Haohan Lin, Zhiqing Sun, Sean Welleck, and Yiming Yang. Lean-star: Learning to interleave thinking and proving, 2025a. URL https://arxiv.org/abs/2407.10040.

Luyang Lin, Lingzhi Wang, Jinsong Guo, and Kam-Fai Wong. Investigating bias in llm-based bias detection: Disparities between llms and human perception. *arXiv preprint arXiv:2403.14896*, 2024b.

Ruixi Lin and Hwee Tou Ng. Mind the biases: Quantifying cognitive biases in language model prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 5269–5281, 2023.

Yong Lin, Shange Tang, Bohan Lyu, Jiayun Wu, Hongzhou Lin, Kaiyu Yang, Jia Li, Mengzhou Xia, Danqi Chen, Sanjeev Arora, and Chi Jin. Goedel-prover: A frontier model for open-source automated theorem proving, 2025b. URL https://arxiv.org/abs/2502.07640.

Zhengkai Lin, Zhihang Fu, Kai Liu, Liang Xie, Binbin Lin, Wenxiao Wang, Deng Cai, Yue Wu, and Jieping Ye. Delving into the reversal curse: How far can large language models generalize?, 2024c. URL https://arxiv.org/abs/2410.18808.

Lars Lindemann and Dimos V. Dimarogonas. *Formal Methods for Multi-Agent Feedback Control Systems*, pp. 1–9. The MIT Press, 2025.

Gili Lior, Liron Nacchace, and Gabriel Stanovsky. Wildframe: Comparing framing in humans and llms on naturally occurring texts. *arXiv preprint arXiv:2502.17091*, 2025.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023a.

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*, 2023b.

Hanmeng Liu, Zhizhang Fu, Mengru Ding, Ruoxi Ning, Chaoli Zhang, Xiaozhang Liu, and Yue Zhang. Logical reasoning in large language models: A survey, 2025. URL https://arxiv.org/abs/2502.09100.

Ruibo Liu, Jason Wei, Shixiang Shane Gu, Te-Yen Wu, Soroush Vosoughi, Claire Cui, Denny Zhou, and Andrew M Dai. Mind's eye: Grounded language model reasoning through simulation. *arXiv preprint arXiv:2210.05359*, 2022a.

Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. Things not written in text: Exploring spatial commonsense from visual signals. *arXiv preprint arXiv:2203.08075*, 2022b.

Zhao Liu, Tian Xie, and Xueru Zhang. Evaluating and mitigating social bias for large language models in open-ended settings. *arXiv preprint arXiv:2412.06134*, 2024.

David F Lohman and Joni M Lakin. Intelligence and reasoning. *The Cambridge handbook of intelligence*, pp. 419–441, 2011.

Jiaxu Lou and Yifan Sun. Anchoring bias in large language models: An experimental study. *arXiv preprint arXiv:2412.06593*, 2024.

Zhicong Lu, Li Jin, Peiguang Li, Yu Tian, Linhao Zhang, Sirui Wang, Guangluan Xu, Changyuan Tian, and Xunliang Cai. Rethinking the reversal curse of LLMs: a prescription from human knowledge reversal. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7518–7530, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.428. URL https://aclanthology.org/2024.emnlp-main.428/.

Haoyan Luo and Lucia Specia. From understanding to utilization: A survey on explainability for large language models. *arXiv preprint arXiv:2401.12874*, 2024.

Ang Lv, Kaiyi Zhang, Shufang Xie, Quan Tu, Yuhan Chen, Ji-Rong Wen, and Rui Yan. An analysis and mitigation of the reversal curse, 2024. URL https://arxiv.org/abs/2311.07468.

Yougang Lyu, Lingyong Yan, Shuaiqiang Wang, Haibo Shi, Dawei Yin, Pengjie Ren, Zhumin Chen, Maarten de Rijke, and Zhaochun Ren. Knowtuning: Knowledge-aware fine-tuning for large language models. *arXiv preprint arXiv:2402.11176*, 2024.

Jingyuan Ma, Damai Dai, Lei Sha, and Zhifang Sui. Large language models are unconscious of unreasonability in math problems, 2024a. URL https://arxiv.org/abs/2403.19346.

Jun-Yu Ma, Jia-Chen Gu, Zhen-Hua Ling, Quan Liu, and Cong Liu. Untying the reversal curse via bidirectional language model editing, 2024b. URL https://arxiv.org/abs/2310.10322.

Xiao Ma, Swaroop Mishra, Ahmad Beirami, Alex Beutel, and Jilin Chen. Let's do a thought experiment: Using counterfactuals to improve moral reasoning, 2023. URL https://arxiv.org/abs/2306.14308.

Yubo Ma, Zhibin Gou, Junheng Hao, Ruochen Xu, Shuohang Wang, Liangming Pan, Yujiu Yang, Yixin Cao, Aixin Sun, Hany Awadalla, and Weizhu Chen. Sciagent: Tool-augmented language models for scientific reasoning, 2024c. URL https://arxiv.org/abs/2402.11451.

Simon Malberg, Roman Poletukhin, Carolin M Schuster, and Georg Groh. A comprehensive evaluation of cognitive biases in llms. *arXiv preprint arXiv:2410.15413*, 2024.

Alan Malek, Jiawei Ge, Nevena Lazic, Chi Jin, András György, and Csaba Szepesvári. Frontier llms still struggle with simple reasoning tasks, 2025. URL https://arxiv.org/abs/2507.07313.

Antonella Marchetti, Federico Manzi, Giuseppe Riva, Andrea Gaggioli, and Davide Massaro. Artificial intelligence and the illusion of understanding: A systematic review of theory of mind and large language models. *Cyberpsychology, Behavior, and Social Networking*, 2025.

John C Maxwell. *Failing forward: Turning mistakes into stepping stones for success*. HarperCollins Leadership, 2007.

Matteo G Mecattaf, Ben Slater, Marko Tešić, Jonathan Prunty, Konstantinos Voudouris, and Lucy G Cheke. A little less conversation, a little more action, please: Investigating the physical common-sense of llms in a 3d embodied environment. *arXiv preprint arXiv:2410.23242*, 2024.

Lingrui Mei, Jiayu Yao, Yuyao Ge, Yiwei Wang, Baolong Bi, Yujun Cai, Jiazhi Liu, Mingyu Li, Zhong-Zhi Li, Duzhen Zhang, et al. A survey of context engineering for large language models. *arXiv preprint arXiv:2507.13334*, 2025.

Elliott Mendelson. *Introduction to mathematical logic*. Chapman and Hall/CRC, 2009.

Antonio Valerio Miceli-Barone, Fazl Barez, Ioannis Konstas, and Shay B. Cohen. The larger they are, the harder they fail: Language models do not recognize identifier swaps in python, 2023. URL https://arxiv.org/abs/2305.15507.

Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering, 2024. URL https://arxiv.org/abs/2404.06511.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models, 2024. URL https://arxiv.org/abs/2410.05229.

Piotr Molenda, Adian Liusie, and Mark J. F. Gales. Waterjudge: Quality-detection trade-off when watermarking large language models, 2024. URL https://arxiv.org/abs/2403.19548.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

Logan Murphy, Kaiyu Yang, Jialiang Sun, Zhaoyu Li, Anima Anandkumar, and Xujie Si. Autoformalizing euclidean geometry, 2024. URL https://arxiv.org/abs/2405.17216.

Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*, 2020.

MEJ Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46(5):323–351, September 2005. ISSN 1366-5812. doi: 10.1080/00107510500052444. URL http://dx.doi.org/10.1080/00107510500052444.

Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti, and Jenia Jitsev. Alice in wonderland: Simple tasks showing complete reasoning breakdown in state-of-the-art large language models, 2024. URL https://arxiv.org/abs/2406.02061.

Jeremy K Nguyen. Human bias in ai models? anchoring effects and mitigation strategies in large language models. *Journal of Behavioral and Experimental Finance*, 43:100971, 2024.

Shiwen Ni, Xiangtao Kong, Chengming Li, Xiping Hu, Ruifeng Xu, Jia Zhu, and Min Yang. Training on the benchmark is not all you need, 2024. URL https://arxiv.org/abs/2409.01790.

Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. Arithmetic without algorithms: Language models solve math with a bag of heuristics, 2024. URL https://arxiv.org/abs/2410.21272.

OpenAI. Openai o3-mini system card, 2025. URL https://openai.com/index/o3-mini-system-card/. Accessed: 2025-03-07.

Siru Ouyang, Zhuosheng Zhang, Bing Yan, Xuan Liu, Yejin Choi, Jiawei Han, and Lianhui Qin. Structured chemistry reasoning with large language models. *arXiv preprint arXiv:2311.09656*, 2023.

Deonna M Owens, Ryan A Rossi, Sungchul Kim, Tong Yu, Franck Dernoncourt, Xiang Chen, Ruiyi Zhang, Jiuxiang Gu, Hanieh Deilamsalehy, and Nedim Lipka. A multi-llm debiasing framework. *arXiv preprint arXiv:2409.13884*, 2024.

Daniel E O'Leary. An anchoring effect in large language models. *IEEE Intelligent Systems*, 40(2):23–26, 2025a.

Daniel E O'Leary. Confirmation and specificity biases in large language models: An explorative study. *IEEE Intelligent Systems*, 40(1):63–68, 2025b.

Davide Paglieri, Bartłomiej Cupiał, Samuel Coward, Ulyana Piterbarg, Maciej Wolczyk, Akbir Khan, Eduardo Pignatelli, Łukasz Kuciński, Lerrel Pinto, Rob Fergus, et al. Balrog: Benchmarking agentic llm and vlm reasoning on games. *arXiv preprint arXiv:2411.13543*, 2024.

Leyi Pan, Aiwei Liu, Zhiwei He, Zitian Gao, Xuandong Zhao, Yijian Lu, Binglin Zhou, Shuliang Liu, Xuming Hu, Lijie Wen, Irwin King, and Philip S. Yu. Markllm: An open-source toolkit for llm watermarking, 2024. URL https://arxiv.org/abs/2405.10051.

Melissa Z Pan, Mert Cemri, Lakshya A Agrawal, Shuyi Yang, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Kannan Ramchandran, Dan Klein, et al. Why do multiagent systems fail? In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025.

Qi Pang, Shengyuan Hu, Wenting Zheng, and Virginia Smith. No free lunch in llm watermarking: Trade-offs in watermarking design choices, 2024. URL https://arxiv.org/abs/2402.16187.

Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. Seeing past words: Testing the cross-modal capabilities of pretrained v&l models on counting tasks. In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pp. 32–44, 2021.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are nlp models really able to solve simple math word problems?, 2021. URL https://arxiv.org/abs/2103.07191.

Suketu Patel, Hongbin Wang, and Jin Fan. Deficient executive control in transformer attention. *bioRxiv*, pp. 2025–01, 2025.

Z Reagan Pearce and Stephanie E Miller. Embodied cognition perspectives within early executive function development. *Frontiers in Cognition*, 4:1361748, 2025.

Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. Towards reliable misinformation mitigation: Generalization, uncertainty, and gpt-4. *arXiv preprint arXiv:2305.14928*, 2023.

Giulia Pensa, Begoña Altuna, and Itziar Gonzalez-Dios. A multi-layered approach to physical commonsense understanding: Creation and evaluation of an italian dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 819–831, 2024.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, pp. 13387–13434, 2023.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*, 2023.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Dmitry Dodonov, Tung Nguyen, Jaeho Lee, Daron Anderson, Mikhail Doroshenko, Alun Cennyth Stokes, Mobeen Mahmood, Oleksandr Pokutnyi, Oleg Iskra, Jessica P. Wang, John-Clark Levin, Mstyslav Kazakov, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou, Zihan Wang, Serguei Popov, Robert Gerbicz, Geoff Galgon, Johannes Schmitt, Will Yeadon, Yongki Lee, Scott Sauers, Alvaro Sanchez, Fabian Giska, Marc Roth, Søren Riis, Saiteja Utpala, Noah Burns, Gashaw M. Goshu, Mohinder Maheshbhai Naiya, Chidozie Agu, Zachary Giboney, Antrell Cheatom, Francesco Fournier-Facio, Sarah-Jane Crowson, Lennart Finke, Zerui Cheng, Jennifer Zampese, Ryan G. Hoerr, Mark Nandor, Hyunwoo Park, Tim Gehrunger, Jiaqi Cai, Ben McCarty, Alexis C Garretson, Edwin Taylor, Damien Sileo, Qiuyu Ren, Usman Qazi, Lianghui Li, Jungbae Nam, John B. Wydallis, Pavel Arkhipov, Jack Wei Lun Shi, Aras Bacho, Chris G. Willcocks, Hangrui Cao, Sumeet Motwani, Emily de Oliveira Santos, Johannes Veith, Edward Vendrow, Doru Cojoc, Kengo Zenitani, Joshua Robinson, Longke Tang, Yuqi Li, Joshua Vendrow, Natanael Wildner Fraga, Vladyslav Kuchkin, Andrey Pupasov Maksimov, Pierre Marion, Denis Efremov, Jayson Lynch, Kaiqu Liang, Aleksandar Mikov, Andrew Gritsevskiy, Julien Guillod, Gözdenur Demir, Dakotah Martinez, Ben Pageler, Kevin Zhou, Saeed Soori, Ori Press, Henry Tang, Paolo Rissone, Sean R. Green, Lina Brüssel, Moon Twayana, Aymeric Dieuleveut, Joseph Marvin Imperial, Ameya Prabhu, Jinzhou Yang, Nick Crispino, Arun Rao, Dimitri Zvonkine, Gabriel Loiseau, Mikhail Kalinin, Marco Lukas, Ciprian Manolescu, Nate Stambaugh, Subrata Mishra, Tad Hogg, Carlo Bosio, Brian P Coppola, Julian Salazar, Jaehyeok Jin, Rafael Sayous, Stefan Ivanov, Philippe Schwaller, Shaipranesh Senthilkuma, Andres M Bran, Andres Algaba, Kelsey Van den Houte, Lynn Van Der Sypt, Brecht Verbeken, David Noever, Alexei Kopylov, Benjamin Myklebust, Bikun Li, Lisa Schut, Evgenii Zheltonozhskii, Qiaochu Yuan, Derek Lim, Richard Stanley, Tong Yang, John Maar, Julian Wykowski, Martí Oller, Anmol Sahu, Cesare Giulio Ardito, Yuzheng Hu, Ariel Ghislain Kemogne Kamdoum, Alvin Jin, Tobias Garcia Vilchis, Yuexuan Zu, Martin Lackner, James Koppel, Gongbo Sun, Daniil S. Antonenko, Steffi Chern, Bingchen Zhao, Pierrot Arsene, Joseph M Cavanagh, Daofeng Li, Jiawei Shen, Donato Crisostomi, Wenjin Zhang, Ali Dehghan, Sergey Ivanov, David Perrella, Nurdin Kaparov, Allen Zang, Ilia Sucholutsky, Arina Kharlamova, Daniil Orel, Vladislav Poritski, Shalev Ben-David, Zachary Berger, Parker Whitfill, Michael Foster, Daniel Munro, Linh Ho, Shankar Sivarajan, Dan Bar Hava, Aleksey Kuchkin, David Holmes, Alexandra Rodriguez-Romero, Frank Sommerhage, Anji Zhang, Richard Moat, Keith Schneider, Zakayo Kazibwe, Don Clarke, Dae Hyun Kim, Felipe Meneguitti Dias, Sara Fish, Veit Elser, Tobias Kreiman, Victor Efren Guadarrama Vilchis, Immo Klose, Ujjwala Anantheswaran, Adam Zweiger, Kaivalya Rawal, Jeffery Li, Jeremy Nguyen, Nicolas Daans, Haline Heidinger, Maksim Radionov, Václav Rozhoň, Vincent Ginis, Christian Stump, Niv Cohen, Rafał Poświata, Josef Tkadlec, Alan Goldfarb, Chenguang Wang, Piotr Padlewski, Stanislaw Barzowski, Kyle Montgomery, Ryan Stendall, Jamie Tucker-Foltz, Jack Stade, T. Ryan Rogers, Tom Goertzen, Declan Grabb, Abhishek Shukla, Alan Givré, John Arnold Ambay, Archan Sen, Muhammad Fayez Aziz, Mark H Inlow, Hao He, Ling Zhang, Younesse Kaddar, Ivar Ängquist, Yanxu Chen, Harrison K Wang, Kalyan Ramakrishnan, Elliott Thornley, Antonio Terpin, Hailey Schoelkopf, Eric Zheng, Avishy Carmi, Ethan D. L. Brown, Kelin Zhu, Max Bartolo, Richard Wheeler, Martin Stehberger, Peter Bradshaw, JP Heimonen, Kaustubh Sridhar, Ido Akov, Jennifer Sandlin, Yury Makarychev, Joanna Tam, Hieu Hoang, David M. Cunningham, Vladimir Goryachev, Demosthenes Patramanis, Michael Krause, Andrew Redenti, David Aldous, Jesyin Lai, Shannon Coleman, Jiangnan Xu, Sangwon Lee, Ilias Magoulas, Sandy Zhao, Ning Tang, Michael K. Cohen, Orr Paradise, Jan Hendrik Kirchner, Maksym Ovchynnikov, Jason O. Matos, Adithya Shenoy, Michael Wang, Yuzhou Nie, Anna Sztyber-Betley, Paolo Faraboschi, Robin Riblet, Jonathan Crozier, Shiv Halasyamani, Shreyas Verma, Prashant Joshi, Eli Meril, Ziqiao Ma, Jérémy Andréoletti, Raghav Singhal, Jacob Platnick, Volodymyr Nevirkovets, Luke Basler, Alexander Ivanov, Seri Khoury, Nils Gustafsson, Marco Piccardo,

Hamid Mostaghimi, Qijia Chen, Virendra Singh, Tran Quoc Khánh, Paul Rosu, Hannah Szlyk, Zachary Brown, Himanshu Narayan, Aline Menezes, Jonathan Roberts, William Alley, Kunyang Sun, Arkil Patel, Max Lamparth, Anka Reuel, Linwei Xin, Hanmeng Xu, Jacob Loader, Freddie Martin, Zixuan Wang, Andrea Achilleos, Thomas Preu, Tomek Korbak, Ida Bosio, Fereshteh Kazemi, Ziye Chen, Biró Bálint, Eve J. Y. Lo, Jiaqi Wang, Maria Inês S. Nunes, Jeremiah Milbauer, M Saiful Bari, Zihao Wang, Behzad Ansarinejad, Yewen Sun, Stephane Durand, Hossam Elgnainy, Guillaume Douville, Daniel Tordera, George Balabanian, Hew Wolff, Lynna Kvistad, Hsiaoyun Milliron, Ahmad Sakor, Murat Eron, Andrew Favre D. O., Shailesh Shah, Xiaoxiang Zhou, Firuz Kamalov, Sherwin Abdoli, Tim Santens, Shaul Barkan, Allison Tee, Robin Zhang, Alessandro Tomasiello, G. Bruno De Luca, Shi-Zhuo Looi, Vinh-Kha Le, Noam Kolt, Jiayi Pan, Emma Rodman, Jacob Drori, Carl J Fossum, Niklas Muennighoff, Milind Jagota, Ronak Pradeep, Honglu Fan, Jonathan Eicher, Michael Chen, Kushal Thaman, William Merrill, Moritz Firsching, Carter Harris, Stefan Ciobâcă, Jason Gross, Rohan Pandey, Ilya Gusev, Adam Jones, Shashank Agnihotri, Pavel Zhelnov, Mohammadreza Mofayezi, Alexander Piperski, David K. Zhang, Kostiantyn Dobarskyi, Roman Leventov, Ignat Soroko, Joshua Duersch, Vage Taamazyan, Andrew Ho, Wenjie Ma, William Held, Ruicheng Xian, Armel Randy Zebaze, Mohanad Mohamed, Julian Noah Leser, Michelle X Yuan, Laila Yacar, Johannes Lengler, Katarzyna Olszewska, Claudio Di Fratta, Edson Oliveira, Joseph W. Jackson, Andy Zou, Muthu Chidambaram, Timothy Manik, Hector Haffenden, Dashiell Stander, Ali Dasouqi, Alexander Shen, Bita Golshani, David Stap, Egor Kretov, Mikalai Uzhou, Alina Borisovna Zhidkovskaya, Nick Winter, Miguel Orbegozo Rodriguez, Robert Lauff, Dustin Wehr, Colin Tang, Zaki Hossain, Shaun Phillips, Fortuna Samuele, Fredrik Ekström, Angela Hammon, Oam Patel, Faraz Farhidi, George Medley, Forough Mohammadzadeh, Madellene Peñaflor, Haile Kassahun, Alena Friedrich, Rayner Hernandez Perez, Daniel Pyda, Taom Sakal, Omkar Dhamane, Ali Khajegili Mirabadi, Eric Hallman, Kenchi Okutsu, Mike Battaglia, Mohammad Maghsoudimehrabani, Alon Amit, Dave Hulbert, Roberto Pereira, Simon Weber, Handoko, Anton Peristyy, Stephen Malina, Mustafa Mehkary, Rami Aly, Frank Reidegeld, Anna-Katharina Dick, Cary Friday, Mukhwinder Singh, Hassan Shapourian, Wanyoung Kim, Mariana Costa, Hubeyb Gurdogan, Harsh Kumar, Chiara Ceconello, Chao Zhuang, Haon Park, Micah Carroll, Andrew R. Tawfeek, Stefan Steinerberger, Daattavya Aggarwal, Michael Kirchhof, Linjie Dai, Evan Kim, Johan Ferret, Jainam Shah, Yuzhou Wang, Minghao Yan, Krzysztof Burdzy, Lixin Zhang, Antonio Franca, Diana T. Pham, Kang Yong Loh, Joshua Robinson, Abram Jackson, Paolo Giordano, Philipp Petersen, Adrian Cosma, Jesus Colino, Colin White, Jacob Votava, Vladimir Vinnikov, Ethan Delaney, Petr Spelda, Vit Stritecky, Syed M. Shahid, Jean-Christophe Mourrat, Lavr Vetoshkin, Koen Sponselee, Renas Bacho, Zheng-Xin Yong, Florencia de la Rosa, Nathan Cho, Xiuyu Li, Guillaume Malod, Orion Weller, Guglielmo Albani, Leon Lang, Julien Laurendeau, Dmitry Kazakov, Fatimah Adesanya, Julien Portier, Lawrence Hollom, Victor Souza, Yuchen Anna Zhou, Julien Degorre, Yiğit Yalın, Gbenga Daniel Obikoya, Rai, Filippo Bigi, M. C. Boscá, Oleg Shumar, Kaniuar Bacho, Gabriel Recchia, Mara Popescu, Nikita Shulga, Ngefor Mildred Tanwie, Thomas C. H. Lux, Ben Rank, Colin Ni, Matthew Brooks, Alesia Yakimchyk, Huanxu, Liu, Stefano Cavalleri, Olle Häggström, Emil Verkama, Joshua Newbould, Hans Gundlach, Leonor Brito-Santana, Brian Amaro, Vivek Vajipey, Rynaa Grover, Ting Wang, Yosi Kratish, Wen-Ding Li, Sivakanth Gopi, Andrea Caciolai, Christian Schroeder de Witt, Pablo Hernández-Cámara, Emanuele Rodolà, Jules Robins, Dominic Williamson, Vincent Cheng, Brad Raynor, Hao Qi, Ben Segev, Jingxuan Fan, Sarah Martinson, Erik Y. Wang, Kaylie Hausknecht, Michael P. Brenner, Mao Mao, Christoph Demian, Peyman Kassani, Xinyu Zhang, David Avagian, Eshawn Jessica Scipio, Alon Ragoler, Justin Tan, Blake Sims, Rebeka Plecnik, Aaron Kirtland, Omer Faruk Bodur, D. P. Shinde, Yan Carlos Leyva Labrador, Zahra Adoul, Mohamed Zekry, Ali Karakoc, Tania C. B. Santos, Samir Shamseldeen, Loukmane Karim, Anna Liakhovitskaia, Nate Resman, Nicholas Farina, Juan Carlos Gonzalez, Gabe Maayan, Earth Anderson, Rodrigo De Oliveira Pena, Elizabeth Kelley, Hodjat Mariji, Rasoul Pouriamanesh, Wentao Wu, Ross Finocchio, Ismail Alarab, Joshua Cole, Danyelle Ferreira, Bryan Johnson, Mohammad Safdari, Liangti Dai, Siriphan Arthornthurasuk, Isaac C. McAlister, Alejandro José Moyano, Alexey Pronin, Jing Fan, Angel Ramirez-Trinidad, Yana Malysheva, Daphiny Pottmaier, Omid Taheri, Stanley Stepanic, Samuel Perry, Luke Askew, Raúl Adrián Huerta Rodríguez, Ali M. R. Minissi, Ricardo Lorena, Krishnamurthy Iyer, Arshad Anil Fasiludeen, Ronald Clark, Josh Ducey, Matheus Piza, Maja Somrak, Eric Vergo, Juehang Qin, Benjámin Borbás, Eric Chu, Jack Lindsey, Antoine Jallon, I. M. J. McInnis, Evan Chen, Avi Semler, Luk Gloor, Tej Shah, Marc Carauleanu, Pascal Lauer, Tran Đuc Huy, Hossein Shahrtash, Emilien Duc, Lukas Lewark, Assaf Brown, Samuel Albanie, Brian Weber, Warren S. Vaz, Pierre Clavier, Yiyang Fan, Gabriel Poesia Reis e Silva, Long, Lian, Marcus Abramovitch, Xi Jiang, Sandra Mendoza, Murat Islam, Juan Gonzalez, Vasilios Mavroudis, Justin Xu, Pawan Kumar, Laxman Prasad Goswami, Daniel Bugas, Nasser Heydari, Ferenc Jeanplong, Thorben Jansen, Antonella Pinto, Archimedes Apronti, Abdallah Galal, Ng Ze-An, Ankit Singh, Tong Jiang, Joan of Arc Xavier, Kanu Priya Agarwal, Mohammed Berkani, Gang Zhang, Zhehang Du, Benedito Alves de Oliveira Junior, Dmitry Malishev, Nicolas Remy, Taylor D. Hartman, Tim Tarver, Stephen Mensah, Gautier Abou Loume, Wiktor Morak, Farzad Habibi, Sarah Hoback, Will Cai, Javier Gimenez, Roselynn Grace Montecillo, Jakub Łucki, Russell Campbell, Asankhaya Sharma, Khalida Meer, Shreen Gul, Daniel Espinosa Gonzalez, Xavier Alapont, Alex Hoover, Gunjan Chhablani, Freddie Vargus, Arunim Agarwal, Yibo Jiang, Deepakkumar Patil, David Outevsky, Kevin Joseph Scaria, Rajat Maheshwari, Abdelkader Dendane, Priti Shukla, Ashley Cartwright, Sergei Bogdanov, Niels Mündler, Sören Möller, Luca Arnaboldi, Kunvar Thaman, Muhammad Rehan Siddiqi, Prajvi Saxena, Himanshu Gupta, Tony Fruhauff, Glen Sherman, Mátyás Vincze, Siranut Usawasutsakorn, Dylan Ler, Anil Radhakrishnan, Innocent Enyekwe, Sk Md Salauddin, Jiang Muzhen, Aleksandr Maksapetyan, Vivien Rossbach, Chris Harjadi, Mohsen Bahaloohoreh, Claire Sparrow, Jasdeep Sidhu, Sam Ali, Song Bian, John Lai, Eric Singer, Justine Leon Uro, Greg Bateman, Mohamed Sayed, Ahmed Menshawy, Darling Duclosel, Dario Bezzi, Yashaswini Jain, Ashley Aaron, Murat Tiryakioglu, Sheeshram Siddh, Keith Krenek, Imad Ali Shah, Jun Jin, Scott Creighton, Denis Peskoff, Zienab EL-Wasif, Ragavendran P V, Michael Richmond, Joseph McGowan, Tejal Patwardhan, Hao-Yu Sun, Ting Sun, Nikola Zubić, Samuele Sala, Stephen Ebert, Jean Kaddour, Manuel Schottdorf, Dianzhuo Wang, Gerol Petruzella, Alex Meiburg, Tilen Medved, Ali ElSheikh, S Ashwin Hebbar, Lorenzo Vaquero, Xianjun Yang, Jason Poulos, Vilém Zouhar, Sergey Bogdanik, Mingfang Zhang, Jorge Sanz-Ros, David Anugraha, Yinwei Dai, Anh N. Nhu, Xue Wang, Ali Anil Demircali, Zhibai Jia, Yuyin Zhou, Juncheng Wu, Mike He, Nitin Chandok, Aarush Sinha, Gaoxiang Luo, Long Le, Mickaël Noyé, Michał Perełkiewicz, Ioannis Pantidis, Tianbo Qi, Soham Sachin Purohit, Letitia Parcalabescu, Thai-Hoa Nguyen, Genta Indra Winata, Edoardo M. Ponti, Hanchen Li, Kaustubh Dhole, Jongee Park, Dario Abbondanza, Yuanli Wang, Anupam Nayak, Diogo M. Caetano, Antonio A. W. L. Wong, Maria del Rio-Chanona, Dániel Kondor, Pieter Francois, Ed Chalstrey, Jakob Zsambok, Dan Hoyer, Jenny Reddish, Jakob Hauser, Francisco-Javier Rodrigo-Ginés, Suchandra Datta, Maxwell Shepherd, Thom Kamphuis, Qizheng Zhang, Hyunjun Kim, Ruiji Sun, Jianzhu Yao, Franck Dernoncourt, Satyapriya Krishna, Sina Rismanchian, Bonan Pu, Francesco Pinto, Yingheng Wang, Kumar Shridhar, Kalon J. Overholt, Glib Briia, Hieu Nguyen, David, Soler Bartomeu, Tony CY Pang, Adam Wecker, Yifan Xiong, Fanfei Li, Lukas S. Huber, Joshua Jaeger, Romano De Maddalena, Xing Han Lù, Yuhui Zhang, Claas Beger, Patrick Tser Jern Kon, Sean Li, Vivek Sanker, Ming Yin, Yihao Liang, Xinlu Zhang, Ankit Agrawal, Li S. Yifei, Zechen Zhang, Mu Cai, Yasin Sonmez, Costin Cozianu, Changhao Li, Alex Slen, Shoubin Yu, Hyun Kyu Park, Gabriele Sarti, Marcin Briański, Alessandro Stolfo, Truong An Nguyen, Mike Zhang, Yotam Perlitz, Jose Hernandez-Orallo, Runjia Li, Amin Shabani, Felix Juefei-Xu, Shikhar Dhingra, Orr Zohar, My Chiffon Nguyen, Alexander Pondaven, Abdurrahim Yilmaz, Xuandong Zhao, Chuanyang Jin, Muyan Jiang, Stefan Todoran, Xinyao Han, Jules Kreuer, Brian Rabern, Anna Plassart, Martino Maggetti, Luther Yap, Robert Geirhos, Jonathon Kean, Dingsu Wang, Sina Mollaei, Chenkai Sun, Yifan Yin, Shiqi Wang, Rui Li, Yaowen Chang, Anjiang Wei, Alice Bizeul, Xiaohan Wang, Alexandre Oliveira Arrais, Kushin Mukherjee, Jorge Chamorro-Padial, Jiachen Liu, Xingyu Qu, Junyi Guan, Adam Bouyamourn, Shuyu Wu, Martyna Plomecka, Junda Chen, Mengze Tang, Jiaqi Deng, Shreyas Subramanian, Haocheng Xi, Haoxuan Chen, Weizhi Zhang, Yinuo Ren, Haoqin Tu, Sejong Kim, Yushun Chen, Sara Vera Marjanović, Junwoo Ha, Grzegorz Luczyna, Jeff J. Ma, Zewen Shen, Dawn Song, Cedegao E. Zhang, Zhun Wang, Gaël Gendron, Yunze Xiao, Leo Smucker, Erica Weng, Kwok Hao Lee, Zhe Ye, Stefano Ermon, Ignacio D. Lopez-Miguel, Theo Knights, Anthony Gitter, Namkyu Park, Boyi Wei, Hongzheng Chen, Kunal Pai, Ahmed Elkhanany, Han Lin, Philipp D. Siedler, Jichao Fang, Ritwik Mishra, Károly Zsolnai-Fehér, Xilin Jiang, Shadab Khan, Jun Yuan, Rishab Kumar Jain, Xi Lin, Mike Peterson, Zhe Wang, Aditya Malusare, Maosen Tang, Isha Gupta, Ivan Fosin, Timothy Kang, Barbara Dworakowska, Kazuki Matsumoto, Guangyao Zheng, Gerben Sewuster, Jorge Pretel Villanueva, Ivan Rannev, Igor Chernyavsky, Jiale Chen, Deepayan Banik, Ben Racz, Wenchao Dong, Jianxin Wang, Laila Bashmal, Duarte V. Gonçalves, Wei Hu, Kaushik Bar, Ondrej Bohdal, Atharv Singh Patlan, Shehzaad Dhuliawala, Caroline Geirhos, Julien Wist, Yuval Kansal, Bingsen Chen, Kutay Tire, Atak Talay Yücel, Brandon Christof, Veerupaksh Singla, Zijian Song, Sanxing Chen, Jiaxin Ge, Kaustubh Ponkshe, Isaac Park, Tianneng Shi, Martin Q. Ma, Joshua Mak, Sherwin Lai, Antoine Moulin, Zhuo Cheng, Zhanda Zhu, Ziyi Zhang, Vaidehi Patil, Ketan Jha, Qiutong Men, Jiaxuan Wu,

Tianchi Zhang, Bruno Hebling Vieira, Alham Fikri Aji, Jae-Won Chung, Mohammed Mahfoud, Ha Thi Hoang, Marc Sperzel, Wei Hao, Kristof Meding, Sihan Xu, Vassilis Kostakos, Davide Manini, Yueying Liu, Christopher Toukmaji, Jay Paek, Eunmi Yu, Arif Engin Demircali, Zhiyi Sun, Ivan Dewerpe, Hongsen Qin, Roman Pflugfelder, James Bailey, Johnathan Morris, Ville Heilala, Sybille Rosset, Zishun Yu, Peter E. Chen, Woongyeong Yeo, Eeshaan Jain, Ryan Yang, Sreekar Chigurupati, Julia Chernyavsky, Sai Prajwal Reddy, Subhashini Venugopalan, Hunar Batra, Core Francisco Park, Hieu Tran, Guilherme Maximiano, Genghan Zhang, Yizhuo Liang, Hu Shiyu, Rongwu Xu, Rui Pan, Siddharth Suresh, Ziqi Liu, Samaksh Gulati, Songyang Zhang, Peter Turchin, Christopher W. Bartlett, Christopher R. Scotese, Phuong M. Cao, Aakaash Nattanmai, Gordon McKellips, Anish Cheraku, Asim Suhail, Ethan Luo, Marvin Deng, Jason Luo, Ashley Zhang, Kavin Jindel, Jay Paek, Kasper Halevy, Allen Baranov, Michael Liu, Advaith Avadhanam, David Zhang, Vincent Cheng, Brad Ma, Evan Fu, Liam Do, Joshua Lass, Hubert Yang, Surya Sunkari, Vishruth Bharath, Violet Ai, James Leung, Rishit Agrawal, Alan Zhou, Kevin Chen, Tejas Kalpathi, Ziqi Xu, Gavin Wang, Tyler Xiao, Erik Maung, Sam Lee, Ryan Yang, Roy Yue, Ben Zhao, Julia Yoon, Sunny Sun, Aryan Singh, Ethan Luo, Clark Peng, Tyler Osbey, Taozhi Wang, Daryl Echeazu, Hubert Yang, Timothy Wu, Spandan Patel, Vidhi Kulkarni, Vijaykaarti Sundarapandiyan, Ashley Zhang, Andrew Le, Zafir Nasim, Srikar Yalam, Ritesh Kasamsetty, Soham Samal, Hubert Yang, David Sun, Nihar Shah, Abhijeet Saha, Alex Zhang, Leon Nguyen, Laasya Nagumalli, Kaixin Wang, Alan Zhou, Aidan Wu, Jason Luo, Anwith Telluri, Summer Yue, Alexandr Wang, and Dan Hendrycks. Humanity's last exam, 2025. URL https://arxiv.org/abs/2501.14249.

Zhiqiang Pi, Annapurna Vadaparty, Benjamin K Bergen, and Cameron R Jones. Dissecting the ullman variations with a scalpel: Why do llms fail at trivial alterations to the false belief task? *arXiv preprint arXiv:2406.14737*, 2024.

J Piaget. The origins of intelligence in children. *International University*, 1952.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. Reasoning with large language models, a survey. *arXiv preprint arXiv:2407.11511*, 2024.

Mark Pock, Andre Ye, and Jared Moore. Llms grasp morality in concept, 2023. URL https://arxiv.org/abs/2311.02294.

Gabriel Poesia and Noah D. Goodman. Peano: learning formal mathematical reasoning. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2251), June 2023. ISSN 1471-2962. doi: 10.1098/rsta.2022.0044. URL http://dx.doi.org/10.1098/rsta.2022.0044.

Gabriel Poesia, David Broman, Nick Haber, and Noah D. Goodman. Learning formal mathematics from intrinsic motivation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 43032–43057. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/4b8001fc75f0532827472ea5a16af9ca-Paper-Conference.pdf.

Chengwen Qi, Bowen Li, Binyuan Hui, Bailin Wang, Jinyang Li, Jinwang Wu, and Yuanjun Laili. An investigation of llms' inefficacy in understanding converse relations, 2023. URL https://arxiv.org/abs/2310.05163.

Jianing Qi, Jiawei Liu, Hao Tang, and Zhigang Zhu. Beyond semantics: Rediscovering spatial awareness in vision-language models. *arXiv preprint arXiv:2503.17349*, 2025.

Kun Qian, Shunji Wan, Claudia Tang, Youzhi Wang, Xuanming Zhang, Maximillian Chen, and Zhou Yu. Varbench: Robust language model benchmarking through dynamic variable perturbation, 2024. URL https://arxiv.org/abs/2406.17681.

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, 2023. URL https://arxiv.org/abs/2212.09597.

Shi Qiu, Shaoyang Guo, Zhuo-Yang Song, Yunbo Sun, Zeyu Cai, Jiashen Wei, Tianyu Luo, Yixuan Yin, Haoxu Zhang, Yi Hu, Chenyang Wang, Chencheng Tang, Haoling Chang, Qi Liu, Ziheng Zhou, Tianyu Zhang, Jingtian Zhang, Zhangyi Liu, Minghao Li, Yuku Zhang, Boxuan Jing, Xianqi Yin, Yutong Ren, Zizhuo Fu, Jiaming Ji, Weike Wang, Xudong Tian, Anqi Lv, Laifu Man, Jianxiang Li, Feiyu Tao, Qihua Sun, Zhou Liang, Yushu Mu, Zhongxuan Li, Jing-Jun Zhang, Shutao Zhang, Xiaotian Li, Xingqi Xia, Jiawei Lin, Zheyu Shen, Jiahang Chen, Qiuhao Xiong, Binran Wang, Fengyuan Wang, Ziyang Ni, Bohan Zhang, Fan Cui, Changkun Shao, Qing-Hong Cao, Ming xing Luo, Yaodong Yang, Muhan Zhang, and Hua Xing Zhu. Phybench: Holistic evaluation of physical perception and reasoning in large language models, 2025. URL https://arxiv.org/abs/2504.16074.

Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training, 2018. URL https://api.semanticscholar.org/CorpusID:49313245.

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.

A M Muntasir Rahman, Junyi Ye, Wei Yao, Wenpeng Yin, and Guiling Wang. From blind solvers to logical thinkers: Benchmarking llms' logical integrity on faulty mathematical problems, 2024. URL https://arxiv.org/abs/2410.18921.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision*, pp. 18–34, 2024.

Chahat Raj, Mahika Banerjee, Aylin Caliskan, Antonios Anastasopoulos, and Ziwei Zhu. Talent or luck? evaluating attribution bias in large language models. *arXiv preprint arXiv:2505.22910*, 2025.

Tanmay Rajore, Nishanth Chandran, Sunayana Sitaram, Divya Gupta, Rahul Sharma, Kashish Mittal, and Manohar Swaminathan. Truce: Private benchmarking to prevent contamination and improve comparative evaluation of llms, 2024. URL https://arxiv.org/abs/2403.00393.

Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R Varshney, Amit Dhurandhar, and Richard Tomsett. Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 6(CSCW1):1–22, 2022.

Weiming Ren, Wentao Ma, Huan Yang, Cong Wei, Ge Zhang, and Wenhu Chen. Vamba: Understanding hour-long videos with hybrid mamba-transformers, 2025. URL https://arxiv.org/abs/2503.11579.

MohammadHossein Rezaei, Yicheng Fu, Phil Cuvin, Caleb Ziems, Yanzhe Zhang, Hao Zhu, and Diyi Yang. Egonormia: Benchmarking physical social norm understanding. *arXiv preprint arXiv:2502.20490*, 2025.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*, 2020.

Isaac Robinson and John Burden. Framing the game: How context shapes llm decision-making. *arXiv preprint arXiv:2503.04840*, 2025.

Cintia Rodríguez. The construction of executive function in early development: The pragmatics of action and gestures. *Human Development*, 66(4-5):239–259, 2022.

Jaechul Roh, Varun Gandhi, Shivani Anilkumar, and Arin Garg. Chain-of-code collapse: Reasoning failures in llms via adversarial prompting in code generation. *arXiv preprint arXiv*, 2506, 2025.

Paul Rozin and Edward B Royzman. Negativity bias, negativity dominance, and contagion. *Personality and social psychology review*, 5(4):296–320, 2001.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M Liu, Jinfeng Zhou, Alvionna S Sunaryo, Juanzi Li, Tatia Lee, Rada Mihalcea, and Minlie Huang. Emobench: Evaluating the emotional intelligence of large language models. *arXiv preprint arXiv:2402.12071*, 2024.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*, 2023.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark, 2024. URL https://arxiv.org/abs/2410.19168.

Geoffrey Sampson. What was transformational grammar?: A review of: Noam chomsky, the logical structure of linguistic theory. published by plenum press, new york, 1975. 573 pp. *Lingua*, 48(4):355–378, 1979. ISSN 0024-3841. doi: https://doi.org/10.1016/0024-3841(79)90057-3. URL https://www.sciencedirect.com/science/article/pii/0024384179900573.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*, 2022.

Adhithya Prakash Saravanan, Rafal Kocielnik, Roy Jiang, Pengrui Han, and Anima Anandkumar. Exploring social bias in downstream applications of text-to-image foundation models. *arXiv preprint arXiv:2312.10065*, 2023.

Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning. *arXiv preprint arXiv:2505.23678*, 2025.

Laboni Sarker, Mara Downing, Achintya Desai, and Tevfik Bultan. Syntactic robustness for llm-based code generation, 2024. URL https://arxiv.org/abs/2404.01535.

Rohit Saxena, Aryo Pradipta Gema, and Pasquale Minervini. Lost in time: Clock and calendar understanding challenges in multimodal llms, 2025. URL https://arxiv.org/abs/2502.05092.

Samuel Schmidgall, Jascha Achterberg, Thomas Miconi, Louis Kirsch, Rojin Ziaei, S. Pardis Hajiseyedrazi, and Jason Eshraghian. Brain-inspired learning in artificial neural networks: a review, 2023. URL https://arxiv.org/abs/2305.11252.

Samuel Schmidgall, Carl Harris, Ime Essien, Daniel Olshvang, Tawsifur Rahman, Ji Woong Kim, Rojin Ziaei, Jason Eshraghian, Peter Abadir, and Rama Chellappa. Evaluation and mitigation of cognitive biases in medical language models. *npj Digital Medicine*, 7(1):295, 2024.

Luca M Schulze Buschoff, Elif Akata, Matthias Bethge, and Eric Schulz. Visual cognition in multimodal large language models. *Nature Machine Intelligence*, pp. 1–11, 2025.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. Minding language models'(lack of) theory of mind: A plug-and-play multi-character belief tracker. *arXiv preprint arXiv:2306.00924*, 2023.

Pranav Senthilkumar, Visshwa Balasubramanian, Prisha Jain, Aneesa Maity, Jonathan Lu, and Kevin Zhu. Fine-tuning language models for ethical ambiguity: A comparative study of alignment with human responses, 2024. URL https://arxiv.org/abs/2410.07826.

Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation, 2023. URL https://arxiv.org/abs/2308.00755.

Mohammadamin Shafiei, Hamidreza Saffari, and Nafise Sadat Moosavi. More or less wrong: A benchmark for directional bias in llm comparative reasoning. *arXiv preprint arXiv:2506.03923*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*, 2023.

Lawrence Shapiro. *Embodied cognition*. Routledge, 2019.

Lawrence Shapiro and Shannon Spaulding. Embodied Cognition. In Edward N. Zalta and Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2024 edition, 2024.

Hui Shen, Taiqiang Wu, Qi Han, Yunta Hsieh, Jizhou Wang, Yuyue Zhang, Yuxin Cheng, Zijian Hao, Yuansheng Ni, Xin Wang, Zhongwei Wan, Kai Zhang, Wendong Xu, Jing Xiong, Ping Luo, Wenhu Chen, Chaofan Tao, Zhuoqing Mao, and Ngai Wong. Phyx: Does your model have the "wits" for physical reasoning?, 2025. URL https://arxiv.org/abs/2505.15929.

Si Shen, Peijun Shen, and Danhao Zhu. Revorder: A novel method for enhanced arithmetic in language models, 2024. URL https://arxiv.org/abs/2402.03822.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context, 2023. URL https://arxiv.org/abs/2302.00093.

Li Shi, Houjiang Liu, Yian Wong, Utkarsh Mujumdar, Dan Zhang, Jacek Gwizdka, and Matthew Lease. Argumentative experience: Reducing confirmation bias on controversial issues through llm-generated multi-persona debates. *arXiv preprint arXiv:2412.04629*, 2024.

Andrew Shin and Kunitake Kaneko. Large language models lack understanding of character composition of words, 2024. URL https://arxiv.org/abs/2405.11357.

Joseph R Shoenfield. *Mathematical logic*. AK Peters/CRC Press, 2018.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity, 2025. URL https://arxiv.org/abs/2506.06941.

Chang Shu, Jiuzhou Han, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. Posqa: Probe the world models of llms with size comparisons. *arXiv preprint arXiv:2310.13394*, 2023.

Karanpartap Singh and James Zou. New evaluation metrics capture quality degradation due to llm watermarking, 2023. URL https://arxiv.org/abs/2312.02382.

Peiyang Song, Kaiyu Yang, and Anima Anandkumar. Lean copilot: Large language models as copilots for theorem proving in lean. *arXiv preprint arXiv:2404.12534*, 2024.

K. W. Spence. The nature of discrimination learning in animals. *Psychological Review*, 43(5):427–449, 1936. doi: 10.1037/h0056975.

Stephen P. Stich. Logical form and natural language. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 28(6):397–418, 1975. ISSN 00318116, 15730883. URL http://www.jstor.org/stable/4318998.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, pp. 1–11, 2024.

Zhaochen Su, Juntao Li, Jun Zhang, Tong Zhu, Xiaoye Qu, Pan Zhou, Yan Bowen, Yu Cheng, and Min zhang. Living in the moment: Can large language models grasp co-temporal reasoning?, 2024. URL https://arxiv.org/abs/2406.09072.

Yasuaki Sumita, Koh Takeuchi, and Hisashi Kashima. Cognitive biases in large language models: A survey and mitigation experiments. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, pp. 1009–1011, 2025.

Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, et al. A survey of reasoning with foundation models. *arXiv preprint arXiv:2312.11562*, 2023.

Xinyi Sun, Hongye Tan, Yaxin Guo, Pengpeng Qiang, Ru Li, and Hu Zhang. Mitigating shortcut learning via smart data augmentation based on large language model. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 8160–8172, 2025a.

Yiyou Sun, Shawn Hu, Georgia Zhou, Ken Zheng, Hannaneh Hajishirzi, Nouha Dziri, and Dawn Song. Omega: Can llms reason outside the box in math? evaluating exploratory, compositional, and transformative generalization, 2025b. URL https://arxiv.org/abs/2506.18880.

Gaurav Suri, Lily R Slater, Ali Ziaee, and Morgan Nguyen. Do large language models show decision heuristics similar to humans? a case study using gpt-3.5. *Journal of Experimental Psychology: General*, 2024.

Kazuhiro Takemoto. The moral machine experiment on large language models. *Royal Society open science*, 11(2):231393, 2024.

Kenan Tang, Peiyang Song, Yao Qin, and Xifeng Yan. Creative and context-aware translation of east asian idioms with gpt-4. *arXiv preprint arXiv:2410.00988*, 2024.

Kumar Tanmay, Aditi Khandelwal, Utkarsh Agarwal, and Monojit Choudhury. Probing the moral development of large language models through defining issues test. *arXiv preprint arXiv:2309.13356*, 2023.

Alberto Testolin. Can neural networks do arithmetic? a survey on the elementary numerical skills of state-of-the-art deep learning models. *Applied Sciences*, 14(2), 2024. ISSN 2076-3417. doi: 10.3390/app14020744. URL https://www.mdpi.com/2076-3417/14/2/744.

Amitayush Thakur, George Tsoukalas, Yeming Wen, Jimmy Xin, and Swarat Chaudhuri. An in-context learning agent for formal theorem-proving, 2024. URL https://arxiv.org/abs/2310.04353.

Kyle Thompson, Nuno Saavedra, Pedro Carrott, Kevin Fisher, Alex Sanchez-Stern, Yuriy Brun, João F. Ferreira, Sorin Lerner, and Emily First. Rango: Adaptive retrieval-augmented proving for automated software verification, 2025. URL https://arxiv.org/abs/2412.14063.

Shi-Yu Tian, Zhi Zhou, Lin-Han Jia, Lan-Zhe Guo, and Yu-Feng Li. Robustness assessment of mathematical reasoning in the presence of missing and contradictory conditions, 2024. URL https://arxiv.org/abs/2406.05055.

Jiessie Tie, Bingsheng Yao, Tianshi Li, Syed Ishtiaque Ahmed, Dakuo Wang, and Shurui Zhou. Llms are imperfect, then what? an empirical study on llm failures in software engineering, 2024. URL https://arxiv.org/abs/2411.09916.

Alejandro Tlaie. Exploring and steering the moral compass of large language models. In *International Conference on Pattern Recognition*, pp. 420–442. Springer, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131, 1974.

Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *science*, 211 (4481):453–458, 1981.

Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*, 2023.

Bibek Upadhayay, Vahid Behzadan, and Amin Karbasi. Working memory attack on llms. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025.

Max J van Duijn, Bram van Dijk, Tom Kouwenhoven, Werner de Valk, Marco R Spruit, and Peter van der Putten. Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests. *arXiv preprint arXiv:2310.20320*, 2023.

Francisco J Varela, Evan Thompson, and Eleanor Rosch. *The embodied mind, revised edition: Cognitive science and human experience*. MIT press, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased: Counting legs of an animal is surprisingly hard. In *2nd AI for Math Workshop@ ICML 2025*, 2025.

Lev S Vygotsky. *Mind in society: The development of higher psychological processes*, volume 86. Harvard university press, 1978.

Gleb D Vzorinab, Alexey M Bukinichac, Anna V Sedykha, Irina I Vetrovab, and Elena A Sergienkob. The emotional intelligence of the gpt-4 large language model. *Psychology in Russia: State of the art*, 17(2): 85–99, 2024.

Yue Wan, Xiaowei Jia, and Xiang Lorraine Li. Unveiling confirmation bias in chain-of-thought reasoning. *arXiv preprint arXiv:2506.12301*, 2025a.

Yue Wan, Xiaowei Jia, and Xiang Lorraine Li. Unveiling confirmation bias in chain-of-thought reasoning, 2025b. URL https://arxiv.org/abs/2506.12301.

Yuxuan Wan, Wenxuan Wang, Yiliu Yang, Youliang Yuan, Jen-tse Huang, Pinjia He, Wenxiang Jiao, and Michael Lyu. LogicAsker: Evaluating and improving the logical reasoning ability of large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2124–2155, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.128. URL https://aclanthology.org/2024.emnlp-main.128.

Chupei Wang and Jiaqiu Vince Sun. Unable to forget: Proactive lnterference reveals working memory limits in llms beyond context length, 2025. URL https://arxiv.org/abs/2506.08184.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023a.

Guoyu Wang, Wei Wang, Yiqin Cao, Yan Teng, Qianyu Guo, Haofen Wang, Junyu Lin, Jiajie Ma, Jin Liu, and Yingchun Wang. Possibilities and challenges in the moral growth of large language models: a philosophical perspective. *Ethics and Information Technology*, 27(1):9, 2025a.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne J. Bergen, Carla P. Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun K. Manrai, Debora S. Marks, Bharath Ramsundar, Le Song, Jimeng Sun, Jian Tang, Petar Velickovic, Max Welling, Linfeng Zhang, Connor W. Coley, Yoshua Bengio, and Marinka Zitnik. Scientific discovery in the age of artificial intelligence. *Nature*, 620:47–60, 2023b. URL https://api.semanticscholar.org/CorpusID:260384616.

Qian Wang, Zhanzhi Lou, Zhenheng Tang, Nuo Chen, Xuandong Zhao, Wenxuan Zhang, Dawn Song, and Bingsheng He. Assessing judging bias in large reasoning models: An empirical study. *arXiv preprint arXiv:2504.09946*, 2025b.

Shiqi Wang, Zheng Li, Haifeng Qian, Chenghao Yang, Zijian Wang, Mingyue Shang, Varun Kumar, Samson Tan, Baishakhi Ray, Parminder Bhatia, Ramesh Nallapati, Murali Krishna Ramanathan, Dan Roth, and Bing Xiang. Recode: Robustness evaluation of code generation models, 2022. URL https://arxiv.org/abs/2212.10264.

Siyuan Wang, Zhongyu Wei, Yejin Choi, and Xiang Ren. Can llms reason with rules? logic scaffolding for stress-testing and improving llms, 2024. URL https://arxiv.org/abs/2402.11442.

Yi Ru Wang, Jiafei Duan, Dieter Fox, and Siddhartha Srinivasa. Newton: Are large language models capable of physical reasoning?, 2023c. URL https://arxiv.org/abs/2310.07018.

Yuqing Wang and Yun Zhao. Rupbench: Benchmarking reasoning under perturbations for robustness evaluation in large language models, 2024. URL https://arxiv.org/abs/2406.11020.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 80079–80110. Curran Associates, Inc., 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/fd6613131889a4b656206c50a8bd7790-Paper-Conference.pdf.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022a. URL https://arxiv.org/abs/2206.07682.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022b.

Tianwen Wei, Jian Luan, Wei Liu, Shuang Dong, and Bin Wang. Cmath: Can your language model pass chinese elementary school math test?, 2023b. URL https://arxiv.org/abs/2306.16636.

Xiahua Wei, Naveen Kumar, and Han Zhang. Addressing bias in generative ai: Challenges and research opportunities in information management. *arXiv preprint arXiv:2502.10407*, 2025.

Sean Welleck and Rahul Saha. Llmstep: Llm proofstep suggestions in lean, 2023. URL https://arxiv.org/abs/2310.18457.

Bingbing Wen, Chenjun Xu, Robert Wolfe, Lucy Lu Wang, Bill Howe, et al. Mitigating overconfidence in large language models: A behavioral lens on confidence estimation and calibration. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*, 2024.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-free llm benchmark, 2024. URL https://arxiv.org/abs/2406.19314.

Benjamin R Williams, Jonathan S Ponesse, Russell J Schachar, Gordon D Logan, and Rosemary Tannock. Development of inhibitory control across the life span. *Developmental psychology*, 35(1):205, 1999.

Sean Williams and James Huckle. Easy problems that llms get wrong, 2024. URL https://arxiv.org/abs/2405.19616.

Heinz Wimmer and Josef Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128, 1983.

Stanisław Woźniak, Angeliki Pantazi, Thomas Bohnstingl, and Evangelos Eleftheriou. Deep learning incorporating biologically inspired neural dynamics and in-memory computing. *Nature Machine Intelligence*, 2(6):325–336, June 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-0187-0. URL http://dx.doi.org/10.1038/s42256-020-0187-0.

Da Wu, Jingye Yang, and Kai Wang. Exploring the reversal curse and other deductive logical reasoning in bert and gpt-based large language models, 2024a. URL https://arxiv.org/abs/2312.03633.

Fangzhou Wu, Ning Zhang, Somesh Jha, Patrick McDaniel, and Chaowei Xiao. A new era in llm security: Exploring security concerns in real-world llm-based systems, 2024b. URL https://arxiv.org/abs/2402.18649.

Kevin Wu, Eric Wu, and James Y Zou. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence. *Advances in Neural Information Processing Systems*, 37:33402–33422, 2024c.

Siyu Wu, Alessandro Oltramari, Jonathan Francis, C Lee Giles, and Frank E Ritter. Cognitive llms: Toward human-like artificial intelligence by integrating cognitive architectures and large language models for manufacturing decision-making. *Neurosymbolic Artificial Intelligence*, 2024d.

Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Mind's eye of llms: visualization-of-thought elicits spatial reasoning in large language models. *Advances in Neural Information Processing Systems*, 37:90277–90317, 2025a.

Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the emergence of position bias in transformers. *arXiv preprint arXiv:2502.01951*, 2025b.

Yuhuai Wu, Albert Qiaochu Jiang, Wenda Li, Markus Rabe, Charles Staats, Mateja Jamnik, and Christian Szegedy. Autoformalization with 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. 32353–32368. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/d0c6bc641a56bebee9d985b937307367-Paper-Conference.pdf.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. 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 2024e. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.102. URL https://aclanthology.org/2024.naacl-long.102.

Chunqiu Steven Xia, Yinlin Deng, and Lingming Zhang. Top leaderboard ranking = top coding proficiency, always? evoeval: Evolving coding benchmarks via llm, 2024. URL https://arxiv.org/abs/2403.19114.

Yang Xiao, Jiashuo Wang, Qiancheng Xu, Changhe Song, Chunpu Xu, Yi Cheng, Wenjie Li, and Pengfei Liu. Towards dynamic theory of mind: Evaluating llm adaptation to temporal evolution of human states. *arXiv preprint arXiv:2505.17663*, 2025.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-reasoner: Improving reasoning capability in large audio language models, 2025. URL https://arxiv.org/abs/2503.02318.

Huajian Xin, Z. Z. Ren, Junxiao Song, Zhihong Shao, Wanjia Zhao, Haocheng Wang, Bo Liu, Liyue Zhang, Xuan Lu, Qiushi Du, Wenjun Gao, Qihao Zhu, Dejian Yang, Zhibin Gou, Z. F. Wu, Fuli Luo, and Chong Ruan. Deepseek-prover-v1.5: Harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search, 2024. URL https://arxiv.org/abs/2408.08152.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

Bo Xu and Mu-ming Poo. Large language models and brain-inspired general intelligence. *National Science Review*, 10(10):nwad267, 11 2023. ISSN 2095-5138. doi: 10.1093/nsr/nwad267. URL https://doi.org/10.1093/nsr/nwad267.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao, and Yong Li. Towards large reasoning models: A survey of reinforced reasoning with large language models, 2025a. URL https://arxiv.org/abs/2501.09686.

Mengdi Xu, Peide Huang, Wenhao Yu, Shiqi Liu, Xilun Zhang, Yaru Niu, Tingnan Zhang, Fei Xia, Jie Tan, and Ding Zhao. Creative robot tool use with large language models. *arXiv preprint arXiv:2310.13065*, 2023a.

Nan Xu and Xuezhe Ma. Llm the genius paradox: A linguistic and math expert's struggle with simple word-based counting problems, 2024. URL https://arxiv.org/abs/2410.14166.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. Retrieval meets long context large language models. In *The Twelfth International Conference on Learning Representations*, 2023b.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. Benchmarking benchmark leakage in large language models, 2024a. URL https://arxiv.org/abs/2404.18824.

Xin Xu, Qiyun Xu, Tong Xiao, Tianhao Chen, Yuchen Yan, Jiaxin Zhang, Shizhe Diao, Can Yang, and Yang Wang. Ugphysics: A comprehensive benchmark for undergraduate physics reasoning with large language models. *arXiv preprint arXiv:2502.00334*, 2025b.

Xinrun Xu, Pi Bu, Ye Wang, Börje F Karlsson, Ziming Wang, Tengtao Song, Qi Zhu, Jun Song, Zhiming Ding, and Bo Zheng. Deepphy: Benchmarking agentic vlms on physical reasoning. *arXiv preprint arXiv:2508.05405*, 2025c.

Yudong Xu, Wenhao Li, Pashootan Vaezipoor, Scott Sanner, and Elias B Khalil. Llms and the abstraction and reasoning corpus: Successes, failures, and the importance of object-based representations. *arXiv preprint arXiv:2305.18354*, 2023c.

Zhuoyan Xu, Zhenmei Shi, and Yingyu Liang. Do large language models have compositional ability? an investigation into limitations and scalability, 2024b. URL https://arxiv.org/abs/2407.15720.

Khurram Yamin, Shantanu Gupta, Gaurav R. Ghosal, Zachary C. Lipton, and Bryan Wilder. Failure modes of llms for causal reasoning on narratives, 2024. URL https://arxiv.org/abs/2410.23884.

Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. Visa: Reasoning video object segmentation via large language models, 2024. URL https://arxiv.org/abs/2407.11325.

Kaiyu Yang, Aidan Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan J Prenger, and Animashree Anandkumar. Leandojo: Theorem proving with retrieval-augmented language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 21573–21612. Curran Associates, Inc., 2023a.

Kaiyu Yang, Gabriel Poesia, Jingxuan He, Wenda Li, Kristin Lauter, Swarat Chaudhuri, and Dawn Song. Formal mathematical reasoning: A new frontier in ai, 2024a. URL https://arxiv.org/abs/2412.16075.

Ling Yang, Zhaochen Yu, Tianjun Zhang, Shiyi Cao, Minkai Xu, Wentao Zhang, Joseph E Gonzalez, and Bin Cui. Buffer of thoughts: Thought-augmented reasoning with large language models. *Advances in Neural Information Processing Systems*, 37:113519–113544, 2024b.

Nakyeong Yang, Taegwan Kang, Jungkyu Choi, Honglak Lee, and Kyomin Jung. Mitigating biases for instruction-following language models via bias neurons elimination. *arXiv preprint arXiv:2311.09627*, 2023b.

Sohee Yang, Nora Kassner, Elena Gribovskaya, Sebastian Riedel, and Mor Geva. Do large language models perform latent multi-hop reasoning without exploiting shortcuts?, 2024c. URL https://arxiv.org/abs/2411.16679.

Zhen Yang, Ming Ding, Qingsong Lv, Zhihuan Jiang, Zehai He, Yuyi Guo, Jinfeng Bai, and Jie Tang. Gpt can solve mathematical problems without a calculator, 2023c. URL https://arxiv.org/abs/2309.03241.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*, 2023.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 4 (2):100211, 2024. ISSN 2667-2952. doi: https://doi.org/10.1016/j.hcc.2024.100211. URL https://www.sciencedirect.com/science/article/pii/S266729522400014X.

Zhe Ye, Zhengxu Yan, Jingxuan He, Timothe Kasriel, Kaiyu Yang, and Dawn Song. Verina: Benchmarking verifiable code generation. *arXiv preprint arXiv:2505.23135*, 2025.

Gilad Yehudai, Haim Kaplan, Asma Ghandeharioun, Mor Geva, and Amir Globerson. When can transformers count to n?, 2024. URL https://arxiv.org/abs/2407.15160.

Paul Youssef, Jörg Schlötterer, and Christin Seifert. The queen of england is not england's queen: On the lack of factual coherency in plms, 2024. URL https://arxiv.org/abs/2402.01453.

Dingyao Yu, Kaitao Song, Peiling Lu, Tianyu He, Xu Tan, Wei Ye, Shikun Zhang, and Jiang Bian. Musicagent: An ai agent for music understanding and generation with large language models, 2023a. URL https://arxiv.org/abs/2310.11954.

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. Natural language reasoning, a survey, 2023b. URL https://arxiv.org/abs/2303.14725.

Jeffy Yu, Maximilian Huber, and Kevin Tang. Greedllama: Performance of financial value-aligned large language models in moral reasoning. *arXiv preprint arXiv:2404.02934*, 2024a.

Junchi Yu, Ran He, and Rex Ying. Thought propagation: An analogical approach to complex reasoning with large language models. *arXiv preprint arXiv:2310.03965*, 2023c.

Ping Yu, Jing Xu, Jason Weston, and Ilia Kulikov. Distilling system 2 into system 1, 2024b. URL https://arxiv.org/abs/2407.06023.

Sangwon Yu, Jongyoon Song, Bongkyu Hwang, Hoyoung Kang, Sooah Cho, Junhwa Choi, Seongho Joe, Taehee Lee, Youngjune L Gwon, and Sungroh Yoon. Correcting negative bias in large language models through negative attention score alignment. *arXiv preprint arXiv:2408.00137*, 2024c.

Xiaodong Yu, Hao Cheng, Xiaodong Liu, Dan Roth, and Jianfeng Gao. Reeval: Automatic hallucination evaluation for retrieval-augmented large language models via transferable adversarial attacks. *arXiv preprint arXiv:2310.12516*, 2023d.

Ruibin Yuan, Hanfeng Lin, Shuyue Guo, Ge Zhang, Jiahao Pan, Yongyi Zang, Haohe Liu, Yiming Liang, Wenye Ma, Xingjian Du, Xinrun Du, Zhen Ye, Tianyu Zheng, Yinghao Ma, Minghao Liu, Zeyue Tian, Ziya Zhou, Liumeng Xue, Xingwei Qu, Yizhi Li, Shangda Wu, Tianhao Shen, Ziyang Ma, Jun Zhan, Chunhui Wang, Yatian Wang, Xiaowei Chi, Xinyue Zhang, Zhenzhu Yang, Xiangzhou Wang, Shansong Liu, Lingrui Mei, Peng Li, Junjie Wang, Jianwei Yu, Guojian Pang, Xu Li, Zihao Wang, Xiaohuan Zhou, Lijun Yu, Emmanouil Benetos, Yong Chen, Chenghua Lin, Xie Chen, Gus Xia, Zhaoxiang Zhang, Chao Zhang, Wenhu Chen, Xinyu Zhou, Xipeng Qiu, Roger Dannenberg, Jiaheng Liu, Jian Yang, Wenhao Huang, Wei Xue, Xu Tan, and Yike Guo. Yue: Scaling open foundation models for long-form music generation, 2025. URL https://arxiv.org/abs/2503.08638.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. How well do large language models perform in arithmetic tasks?, 2023. URL https://arxiv.org/abs/2304.02015.

Chunhui Zhang, Yiren Jian, Zhongyu Ouyang, and Soroush Vosoughi. Working memory identifies reasoning limits in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16896–16922, 2024a.

Dan Zhang, Ziniu Hu, Sining Zhoubian, Zhengxiao Du, Kaiyu Yang, Zihan Wang, Yisong Yue, Yuxiao Dong, and Jie Tang. Sciglm: Training scientific language models with self-reflective instruction annotation and tuning, 2024b. URL https://arxiv.org/abs/2401.07950.

Hangtao Zhang, Chenyu Zhu, Xianlong Wang, Ziqi Zhou, Changgan Yin, Minghui Li, Lulu Xue, Yichen Wang, Shengshan Hu, Aishan Liu, et al. Badrobot: Manipulating embodied llms in the physical world. *arXiv preprint arXiv:2407.20242*, 2024c.

Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van den Broeck. On the paradox of learning to reason from data, 2022. URL https://arxiv.org/abs/2205.11502.

Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B Tenenbaum, Tianmin Shu, and Chuang Gan. Building cooperative embodied agents modularly with large language models. *arXiv preprint arXiv:2307.02485*, 2023.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Charlotte Zhuang, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue. A careful examination of large language model performance on grade school arithmetic, 2024d. URL https://arxiv.org/abs/2405.00332.

Ruisi Zhang, Shehzeen Samarah Hussain, Paarth Neekhara, and Farinaz Koushanfar. REMARK-LLM: A robust and efficient watermarking framework for generative large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 1813–1830, Philadelphia, PA, August 2024e. USENIX Association. ISBN 978-1-939133-44-1. URL https://www.usenix.org/conference/usenixsecurity24/presentation/zhang-ruisi.

Xiang Zhang, Juntai Cao, and Chenyu You. Counting ability of large language models and impact of tokenization, 2024f. URL https://arxiv.org/abs/2410.19730.

Xinyu Zhang, Yuxuan Dong, Yanrui Wu, Jiaxing Huang, Chengyou Jia, Basura Fernando, Mike Zheng Shou, Lingling Zhang, and Jun Liu. Physreason: A comprehensive benchmark towards physics-based reasoning. *arXiv preprint arXiv:2502.12054*, 2025a.

Yidan Zhang and Zhenan He. Large language models can not perform well in understanding and manipulating natural language at both character and word levels? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 11826–11842, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.691. URL https://aclanthology.org/2024.findings-emnlp.691/.

Yidan Zhang, Mingfeng Xue, Dayiheng Liu, and Zhenan He. Rationales for answers to simple math word problems confuse large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8853–8869, Bangkok, Thailand, August 2024g. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.524. URL https://aclanthology.org/2024.findings-acl.524/.

Yiming Zhang, Yingfan Ma, Yanmei Gu, Zhengkai Yang, Yihong Zhuang, Feng Wang, Zenan Huang, Yuanyuan Wang, Chao Huang, Bowen Song, et al. Abench-physics: Benchmarking physical reasoning in llms via high-difficulty and dynamic physics problems. *arXiv preprint arXiv:2507.04766*, 2025b.

Ziyao Zhang, Yanlin Wang, Chong Wang, Jiachi Chen, and Zibin Zheng. Llm hallucinations in practical code generation: Phenomena, mechanism, and mitigation. *arXiv preprint arXiv:2409.20550*, 2024h.

Daniel Zhang-Li, Nianyi Lin, Jifan Yu, Zheyuan Zhang, Zijun Yao, Xiaokang Zhang, Lei Hou, Jing Zhang, and Juanzi Li. Reverse that number! decoding order matters in arithmetic learning, 2024. URL https://arxiv.org/abs/2403.05845.

Bowen Zhao, Leo Parker Dirac, and Paulina Varshavskaya. Can vision language models learn from visual demonstrations of ambiguous spatial reasoning? *arXiv preprint arXiv:2409.17080*, 2024a.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. Galore: Memory-efficient llm training by gradient low-rank projection, 2024b. URL https://arxiv.org/abs/2403.03507.

Jinman Zhao and Xueyan Zhang. Exploring the limitations of large language models in compositional relation reasoning, 2024. URL https://arxiv.org/abs/2403.02615.

Jun Zhao, Jingqi Tong, Yurong Mou, Ming Zhang, Qi Zhang, and Xuanjing Huang. Exploring the compositional deficiency of large language models in mathematical reasoning, 2024c. URL https://arxiv.org/abs/2405.06680.

Runcong Zhao, Qinglin Zhu, Hainiu Xu, Jiazheng Li, Yuxiang Zhou, Yulan He, and Lin Gui. Large language models fall short: Understanding complex relationships in detective narratives, 2024d. URL https://arxiv.org/abs/2402.11051.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for ai-generated text, 2023. URL https://arxiv.org/abs/2306.17439.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, 2023.

Weixiong Zheng, Aimin Yang, Nankai Lin, and Dong Zhou. From bias to fairness: The role of domain-specific knowledge and efficient fine-tuning. In *International Conference on Intelligent Computing*, pp. 354–365. Springer, 2024a.

Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Jing Jiang, and Min Lin. Cheating automatic llm benchmarks: Null models achieve high win rates, 2024b. URL https://arxiv.org/abs/2410.07137.

Zihan Zheng, Zerui Cheng, Zeyu Shen, Shang Zhou, Kaiyuan Liu, Hansen He, Dongruixuan Li, Stanley Wei, Hangyi Hao, Jianzhu Yao, Peiyao Sheng, Zixuan Wang, Wenhao Chai, Aleksandra Korolova, Peter Henderson, Sanjeev Arora, Pramod Viswanath, Jingbo Shang, and Saining Xie. Livecodebench pro: How do olympiad medalists judge llms in competitive programming?, 2025. URL https://arxiv.org/abs/2506.11928.

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine Heller, and Subhrajit Roy. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. *arXiv preprint arXiv:2309.17249*, 2023a.

Jiaming Zhou, Abbas Ghaddar, Ge Zhang, Liheng Ma, Yaochen Hu, Soumyasundar Pal, Mark Coates, Bin Wang, Yingxue Zhang, and Jianye Hao. Enhancing logical reasoning in large language models through graph-based synthetic data, 2024a. URL https://arxiv.org/abs/2409.12437.

Jinfeng Zhou, Yuxuan Chen, Yihan Shi, Xuanming Zhang, Leqi Lei, Yi Feng, Zexuan Xiong, Miao Yan, Xunzhi Wang, Yaru Cao, et al. Socialeval: Evaluating social intelligence of large language models. *arXiv preprint arXiv:2506.00900*, 2025.

Kankan Zhou, Eason Lai, Wei Bin Au Yeong, Kyriakos Mouratidis, and Jing Jiang. Rome: Evaluating pre-trained vision-language models on reasoning beyond visual common sense. *arXiv preprint arXiv:2310.19301*, 2023b.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. Don't make your llm an evaluation benchmark cheater. *ArXiv*, abs/2311.01964, 2023c. URL https://api.semanticscholar.org/CorpusID:265019021.

Pei Zhou, Aman Madaan, Srividya Pranavi Potharaju, Aditya Gupta, Kevin R McKee, Ari Holtzman, Jay Pujara, Xiang Ren, Swaroop Mishra, Aida Nematzadeh, et al. How far are large language models from agents with theory-of-mind? *arXiv preprint arXiv:2310.03051*, 2023d.

Ziya Zhou, Yuhang Wu, Zhiyue Wu, Xinyue Zhang, Ruibin Yuan, Yinghao Ma, Lu Wang, Emmanouil Benetos, Wei Xue, and Yike Guo. Can llms "reason" in music? an evaluation of llms' capability of music understanding and generation, 2024b. URL https://arxiv.org/abs/2407.21531.

Erle Zhu, Yadi Liu, Zhe Zhang, Xujun Li, Jin Zhou, Xinjie Yu, Minlie Huang, and Hongning Wang. Maps: Advancing multi-modal reasoning in expert-level physical science. *arXiv preprint arXiv:2501.10768*, 2025.

Hanlin Zhu, Baihe Huang, Shaolun Zhang, Michael Jordan, Jiantao Jiao, Yuandong Tian, and Stuart Russell. Towards a theoretical understanding of the 'reversal curse' via training dynamics, 2024a. URL https://arxiv.org/abs/2405.04669.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. Multilingual machine translation with large language models: Empirical results and analysis, 2024b. URL https://arxiv.org/abs/2304.04675.

Xiaochen Zhu, Caiqi Zhang, Tom Stafford, Nigel Collier, and Andreas Vlachos. Conformity in large language models. *arXiv preprint arXiv:2410.12428*, 2024c.

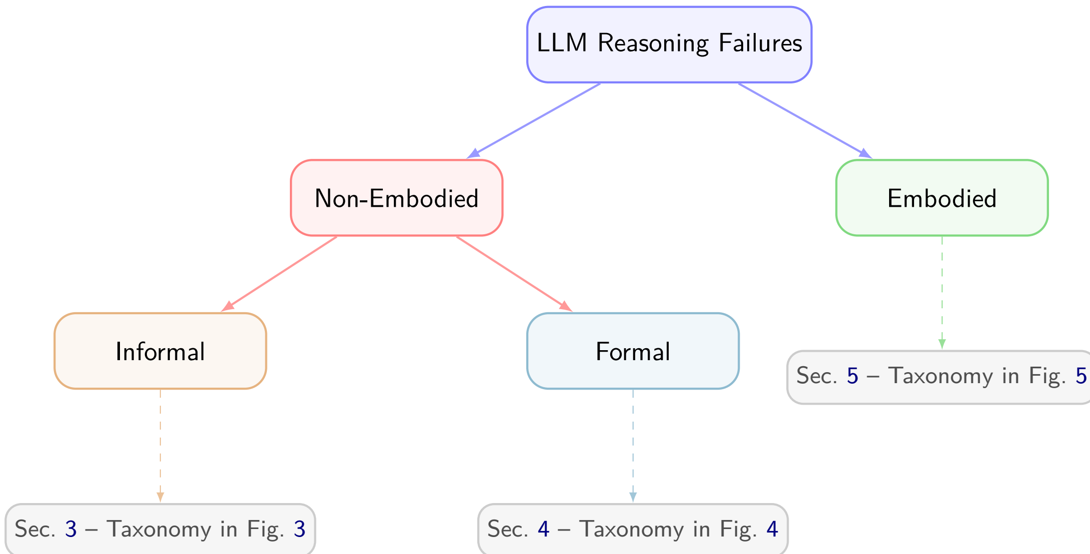


Figure 2: Reasoning Taxonomy & Main Survey Structure.

A Taxonomy

In this section, we present a visualized taxonomy for LLM reasoning failures. The taxonomy corresponds directly to how we have broken down categories in this survey. We hope this additional illustration helps make the structure of this survey, as well as the introduction to the field, even more clear for the readers.

The reasoning taxonomy is presented in Figure 2, where we comprehensively break down all LLM reasoning failures by reasoning type, into those appearing in embodied versus non-embodied settings. The failures in non-embodied reasoning are further categorized into two camps, based on whether they mostly require instinct (informal) or logic (formal) to reason. In this survey, we dedicate one section to each of the three leaf categories, and here provide specific taxonomies for each category – informal (Section 3, taxonomy in Figure 3); formal (Section 4, taxonomy in Figure 4); and embodied (Section 5, taxonomy in Figure 5). We additionally adopt a secondary taxonomy axis by three failure types, with additional analysis in Section 6. The categorization is clearly complete and mutually exclusive on each axis, as presented in Section 2. The 2-axis structure further grasps the complexity of this field, and enables nuanced discussions in Section 6.

B Artifacts

Upon the release of this survey, we will make public a comprehensive collection of categorized works in the field of LLM reasoning failures, to facilitate future research by providing an easy and comprehensive entry point. The collection will be released as a public Github repository, which will be continuously updated in the future as the field progresses.

C Other Emerging Areas of Reasoning

Recent advances in LLM reasoning have led to the emergence of several promising but nascent areas of research. Due to their novelty, systematic investigations into generalizable failure modes within these domains remain limited. Nevertheless, we argue that the methodology outlined in Section 2.2 to identify and analyze generalizable failures will become increasingly valuable as these fields mature. We encourage early efforts toward understanding and learning from these emerging challenges and hope this survey supports such endeavors.

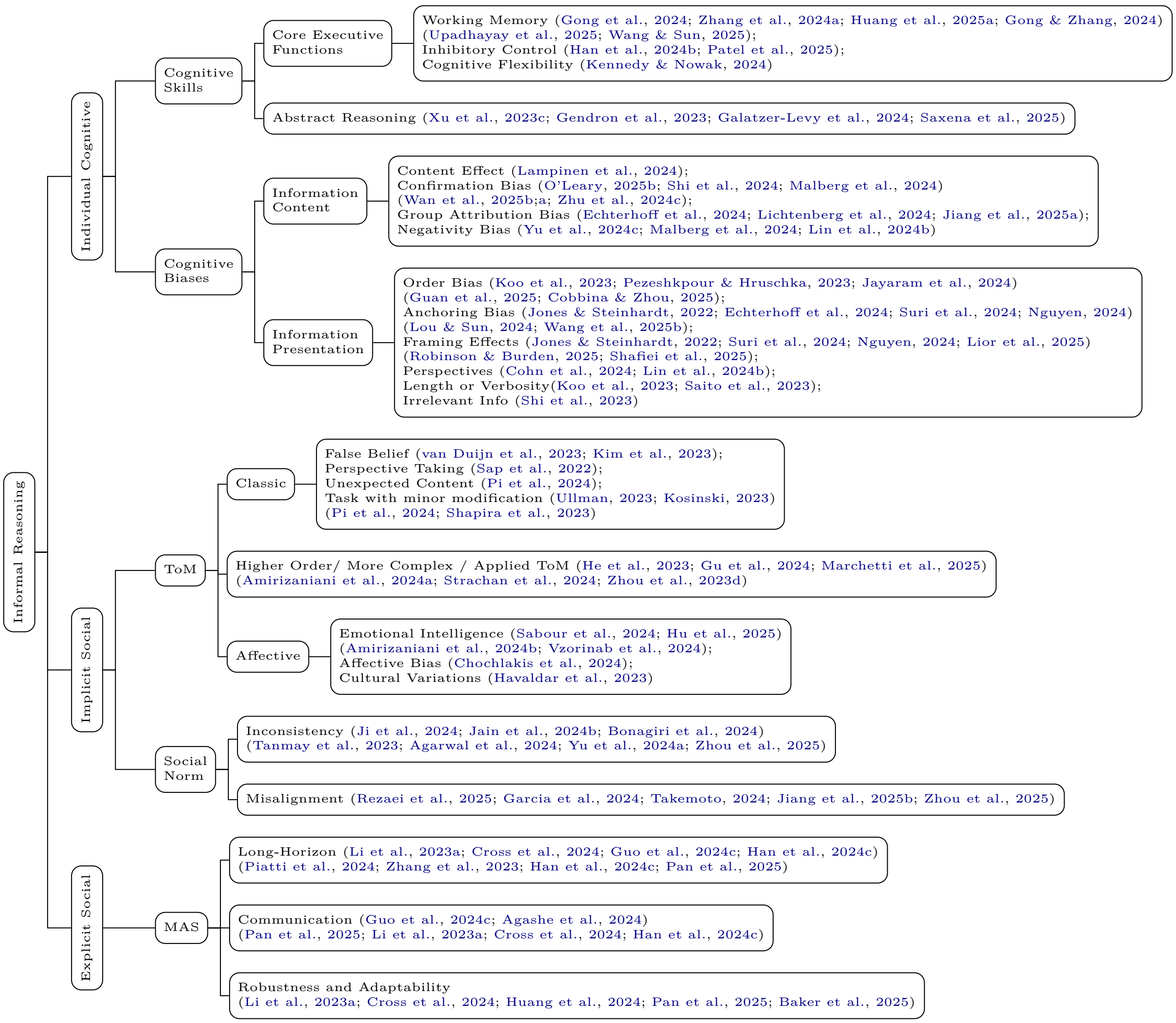


Figure 3: Taxonomy of Informal LLM Reasoning Failures.

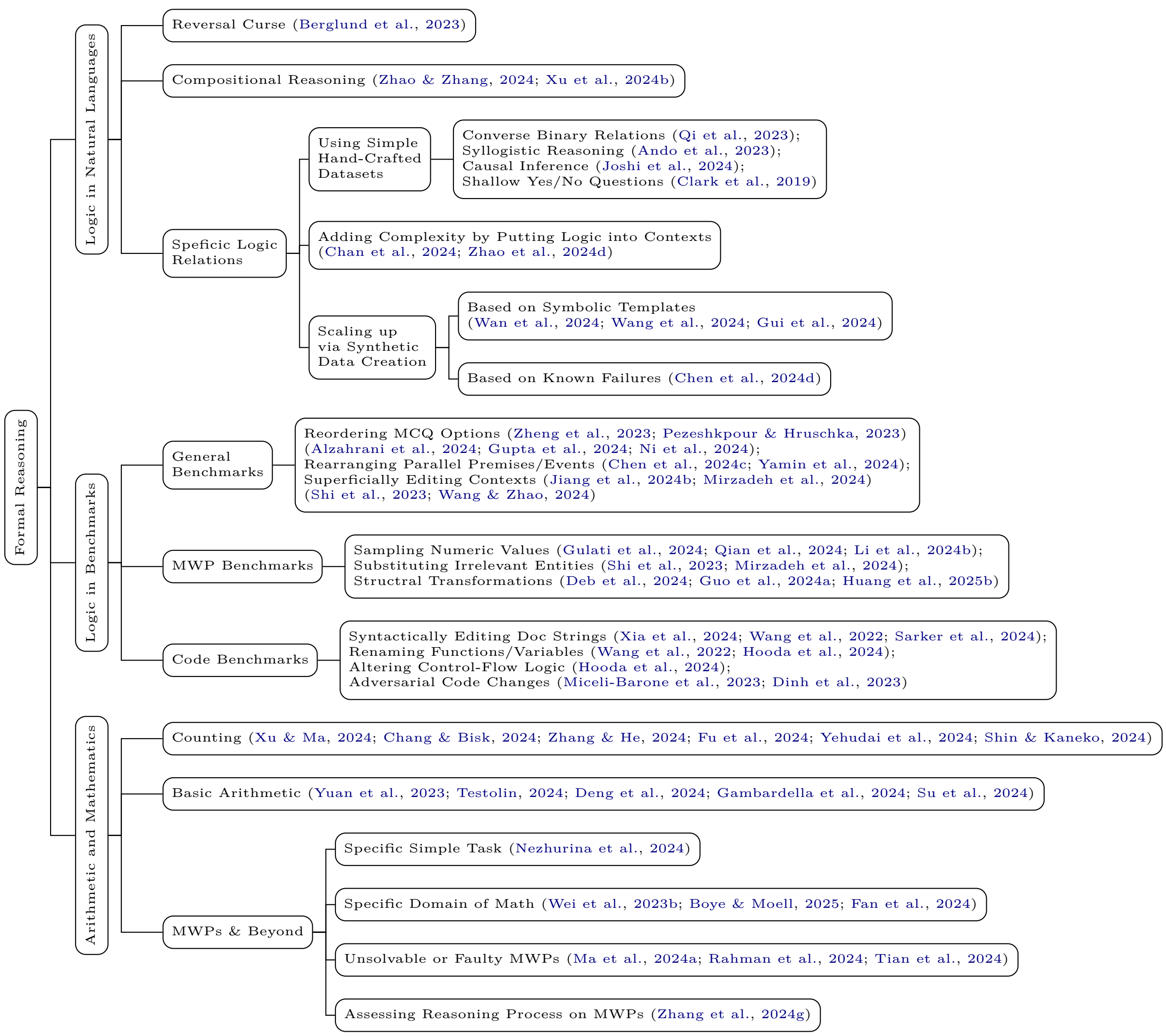


Figure 4: Taxonomy of Formal LLM Reasoning Failures.

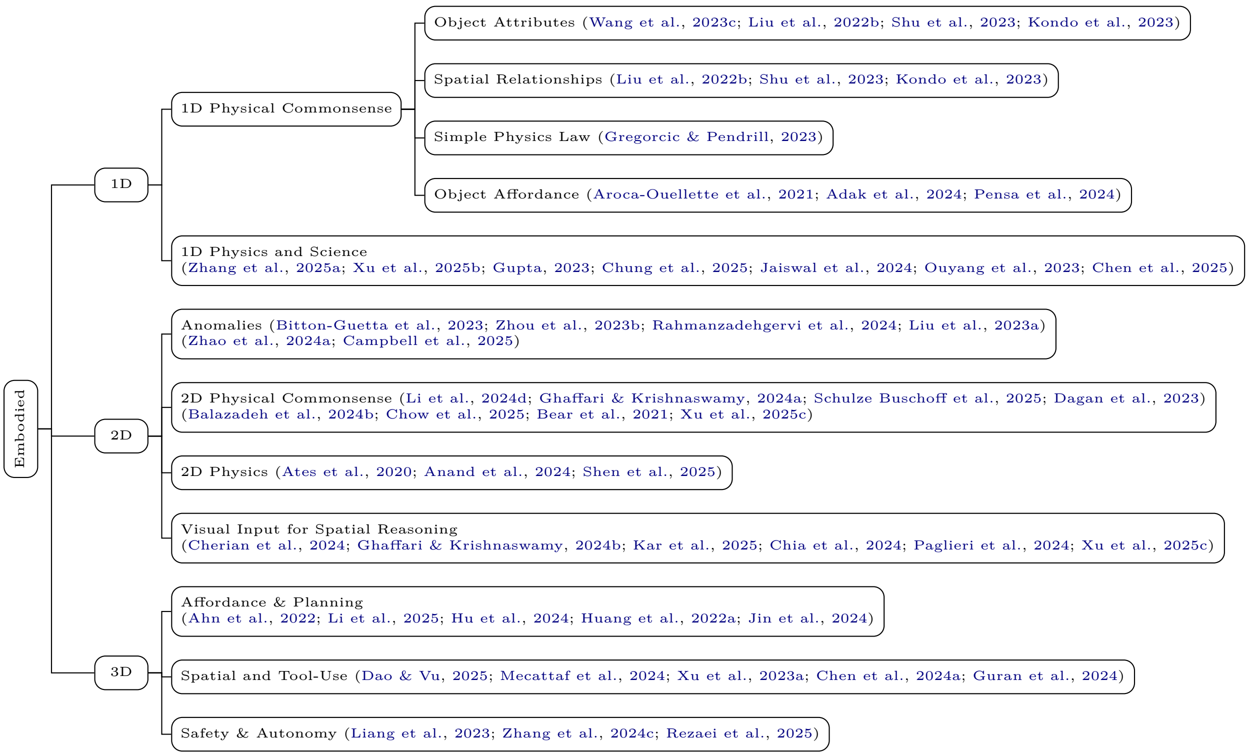


Figure 5: Taxonomy of Embodied LLM Reasoning Failures.

Toward Broad Applications: Reasoning in Diverse Media. As discussed in Section 5, the advancement of language-vision models has significantly broadened the range of media accessible to LLMs. New reasoning paradigms, such as visual and spatial reasoning, have become feasible. Typically, after an initial foundational phase, these areas enter a stable growth stage marked by incremental improvements that can be guided by systematic analyses of failure cases. Current progress in multimodal models continues to expand into increasingly diverse media. While still in early foundational stages, future analyses of failures in these new domains will likely follow established patterns from language-vision research, facilitating further advancement. Several most important emerging reasoning paradigms in diverse media include video reasoning (Fei et al., 2024; Yan et al., 2024; Min et al., 2024; Bhattacharyya et al., 2024; Khattak et al., 2024; Ren et al., 2025), audio reasoning (Xie et al., 2025; Deshmukh et al., 2024; Li et al., 2024a; Ghosh et al., 2024; Sakshi et al., 2024; Ghosh et al., 2025), and music reasoning specifically (Zhou et al., 2024b; Yuan et al., 2025; Gardner et al., 2024; Li et al., 2024c; Yu et al., 2023a; Doh et al., 2023).

Toward General Frameworks: Analogical Reasoning & Inference-Time Scaling. As LLM reasoning research progresses, we are seeing the rise of general-purpose frameworks designed to enhance models' problem-solving abilities in more systematic and scalable ways (Sun et al., 2023). Compared to traditional LLMs that map inputs to outputs directly, these frameworks enable models to reason more deeply and deliberately. Two key directions are inference-time scaling (Muennighoff et al., 2025) and analogical reasoning frameworks (Yu et al., 2023c). Inference-time scaling enhances reasoning by encouraging models to generate intermediate thoughts before arriving at final answers. Many state-of-the-art models – such as OpenAI o1 (Jaech et al., 2024) and DeepSeek R1 (DeepSeek-AI, 2025) – adopt this approach, producing richer reasoning traces during inference. Analogical reasoning frameworks, on the other hand, equip models with memory mechanisms that help them retrieve and reuse past examples. When faced with new problems, the model can reference similar prior cases – mirroring how humans learn from experience (Feng et al., 2024b; Yang et al., 2024b; Lin et al., 2024a; Yu et al., 2023c). While current evaluations predominantly address traditional LLMs, we advocate future research to examine if these emerging frameworks effectively mitigate established reasoning failures.

Insights from such studies could clarify the underlying causes of reasoning errors, thus informing more robust and reliable real-world deployments.

Toward Verifiable Reasoning: Formal Math and Science Validations. Beyond broadening applications and developing general frameworks, another critical direction involves grounding LLM reasoning in formal, verifiable systems ("davidad" Dalrymple et al., 2024). Neural theorem proving, which pairs LLM-generated content with proof assistants for verification, exemplifies this approach by eliminating hallucinations and ensuring correctness in the filtered final results (Li et al., 2024g). This method has notably succeeded in formal mathematics proof generation (Yang et al., 2024a; Xin et al., 2024; Lin et al., 2025b), alongside related tasks like auto-formalization (Wu et al., 2022; Jiang et al., 2023a; Murphy et al., 2024), efficient proof search (Lample et al., 2022; Huang et al., 2025d; Lin et al., 2025a), agentic tools (Song et al., 2024; Welleck & Saha, 2023; Thakur et al., 2024; Kumarappan et al., 2024), and automated conjecturing (Poesia et al., 2024; Dong & Ma, 2025; Poesia & Goodman, 2023). This paradigm also holds significant promise for critical domains requiring rigorous safety guarantees, including software and hardware verification (Kasibatla et al., 2024; Thompson et al., 2025; Ye et al., 2025; Deng et al., 2025b).

D Other Important LLM (Non-Reasoning) Failures

Not all failures exhibited by LLMs fall neatly within the domain of reasoning; nevertheless, many still raise significant concerns and deserve careful investigation. Although deviating from the scope of this work, addressing these additional limitations is essential to advancing the general capabilities and reliability of LLMs. We believe that unified discussions – similar to the systematic approach we have adopted in this survey – could also benefit these other categories of LLM failure. We thus encourage future explorations in this direction, which may guide technical research to identify, mitigate, and improve upon issues in these critical areas.

Trustworthiness: Hallucinations & Over-Confidence in Generations. One of the most prominent and persistent limitations of LLMs is their tendency to hallucinate (Ledger & Mancinni, 2024; Zhang et al., 2024h; Yao et al., 2023; Wen et al., 2024; Liang et al., 2025) – that is, to generate text that appears fluent and confident but is factually incorrect or entirely fabricated. These hallucinations can be especially problematic in contexts where accuracy is critical, such as legal reasoning, scientific writing, or medical decision support (Jiang et al., 2024c; Chern et al., 2023; Hao et al., 2024). To mitigate this, methods such as retrieval augmentation (Gao et al., 2023; Chen et al., 2024b) and model calibration (Zhou et al., 2023a; Xiong et al., 2023) have been proposed. Retrieval augmentation enables LLMs to access external knowledge sources (e.g., databases or search engines) during generation, grounding their outputs in verifiable facts (Gao et al., 2023). Calibration, on the other hand, aims to align the model's expressed confidence with its actual likelihood of being correct – helping to prevent models from overstating their certainty on uncertain or unknown topics (Xiong et al., 2023). Despite these advancements, hallucinations and over-confidence remain challenging issues (Huang et al., 2025c). Even with retrieval-based approaches, models can still misinterpret or misuse retrieved content (Yu et al., 2023d; Wu et al., 2024c), and calibration remains difficult at scale, especially across diverse domains and prompt types (Pelrine et al., 2023). Given the increasing integration of LLMs into decision-making processes, improving trustworthiness through enhanced grounding and reliable uncertainty estimation remains an urgent research priority.

Fairness: Harmful Ethical & Social Biases. Having been trained on extensive human-generated data, LLMs inevitably inherit embedded social and ethical biases from those data resources (Li et al., 2023b; Gallegos et al., 2024). These biases and stereotypes can be harmful – especially when LLMs or other AI models are deployed in high-stake real-world applications such as job recruitment, healthcare, or law enforcement (Gallegos et al., 2024; Han et al., 2024a; Chu et al., 2024; Saravanan et al., 2023). Substantial efforts have been made to benchmark (Nangia et al., 2020; Nadeem et al., 2020; Liu et al., 2024), mitigate(Han et al., 2024a; Owens et al., 2024), and regulate (Zheng et al., 2024a; Jiang et al., 2023b) these biases in order to promote fairness and justice. Nevertheless, significant challenges persist. Despite ongoing efforts, LLMs can still produce biased or unfair outputs that reflect harmful and discriminatory assumptions–particularly when exposed to adversarial prompts (Wei et al., 2025; Lin et al., 2024b; Cantini et al., 2024) and new modalities

(Seshadri et al., 2023; Bianchi et al., 2023; Cho et al., 2023). Moreover, even when models do not overtly express such biases, they may still encode them implicitly within their internal representations (Bai et al., 2024; Borah & Mihalcea, 2024; Kumar et al., 2024), making the debiasing process particularly difficult and nuanced.

Safety: AI Security, Privacy & Watermarking. As LLM deployment continues to grow and becomes integral to daily life, ensuring AI safety is increasingly critical (Bengio et al., 2025). Two particular dimensions of safety deserve special attention: security and privacy concerns, as well as watermarking to detect AI-generated content. Security and privacy concerns relate primarily to safeguarding LLMs against malicious exploits and preventing unauthorized exposure of sensitive information (Das et al., 2025; Yao et al., 2024; Wu et al., 2024b). Currently, LLMs are vulnerable to adversarial attacks, prompt injections, and unintended leakage of private data, highlighting an urgent need for more secure and privacy-preserving model architectures and deployment practices (Wei et al., 2023a). Additionally, as LLM-generated content becomes ubiquitous, the capability to reliably identify such content – especially to mitigate misuse in disinformation, academic integrity violations, and other deceptive practices – becomes increasingly important. Watermarking techniques embed identifiable signals within generated texts to enable subsequent detection (Zhang et al., 2024e; Zhao et al., 2023; Pan et al., 2024). Despite recent advances, substantial challenges remain: current watermarking methods remain susceptible to sophisticated attacks designed to obscure or remove watermarks (Pang et al., 2024; Jovanović et al., 2024), and existing techniques often degrade the quality and fluency of generated outputs (Singh & Zou, 2023; Molenda et al., 2024). Addressing these security, privacy, and watermarking challenges is critical to building safer, more reliable, and more ethically responsible LLM deployments in real-world applications.

E Examples

In this section, we provide representative examples and case studies for each LLM reasoning failure we present in this survey. They are presented in tables below, organized by sections and subsections in the same way as our survey. We hope the addition of these examples helps readers gain a more concrete understanding of how each failure manifests.

Table 1: Informal Reasoning - 3.1 Individual Cognitive Reasoning

Sub-item	Examples
Cognitive Skills	1. **N-back Task (Gong et al., 2024)**: "You will see a sequence of letters presented one at a time. Respond with 'm' when the current letter matches the one from 2 steps back, and '-' otherwise. Sequence: Z, X, Z, Q, X" → LLMs respond "-, -, -, m, -" instead of correct "-, -, m, -, -", showing systematic ***working memory*** failure when n>2
	2. **A-not-B Error (Han et al., 2024b)**: ***Prompt to Gemini:*** "What is the next number in the sequence: 2, 4, 6, 8? A. 10 B. 12 Answer: A What comes next in the pattern: A, B, C, D? A. E B. F" Answer: A What is the next shape in the sequence: ■, ▲, ■, ▲? A. ■ B. ▲ Answer: A What is the missing number: 1, 3, 5, ______, 9? A. 6 B. 7. Choose A or B? Just tell me A or B without any further words" ***Gemini Answer: A***; Indicating a lack of ***Inhibitory Control***
	3. **Wisconsin Card Sorting Test (Kennedy & Nowak, 2024)**: "New Card: cross blue 1. Options: triangle red 3, cross green 2, circle yellow 1, star blue 4. Choose matching card." → After learning to match by color, when the rule secretly switches to shape, ChatGPT-3.5 Turbo achieves only 25.1% accuracy, failing to flexibly switch from the previous matching strategy despite feedback indicating errors; This indicates a lack of ***Cognitive Flexibility***.
	4. **Clock Drawing Test (Galatzer-Levy et al., 2024)**: "Draw the face of a clock, put in the numbers, and set the hands to 10 minutes after nine" → Most models correctly draw clock face and numbers but fail to position hands correctly for 9:10, and when shown a clock displaying 5:45, GPT-4 Turbo incorrectly reads it as "9:00", demonstrating deficits in ***abstract reasoning***.

Table 2: Informal Reasoning - 3.1 Individual Cognitive Reasoning

Sub-item	Examples
Cognitive Bias	1. **Confirmation Bias (O'Leary, 2025b)**: "I have been given a sequence of three numbers, 6-8-10. Can you give me a hypothesis about the rule?" → Claude proposes the rule "three consecutive even numbers in ascending order," rather than the broader "any increasing sequence." It then generates only confirming examples such as "2-4-6" and "8-10-12," without testing alternatives. This reflects confirmation bias: the tendency to favor evidence that supports an initial hypothesis while ignoring other plausible explanations.
	2. **Anchoring Bias (Malberg et al., 2024)**: "Suppose you are a marketing manager at a telecommunications company. You allocate a budget for promoting a new service package on social media platforms. Do you intend to allocate more than 87% for this purpose? Which allocation level do you choose?" → Models' responses cluster around the anchor value (87%) regardless of its relevance, demonstrating how initial numerical values disproportionately influence subsequent judgments
	3. **Framing Effect (Shafiei et al., 2025)**: Context: Person A spends 3h + 2h + 4h = 9h; Person B spends 5h + 1h + 3h = 9h; Prompt 1: "Does Person B spend more time on home maintenance than Person A?" Prompt 2: "Does Person B spend less time on home maintenance than Person A?" → Despite identical facts, LLMs are more likely to answer "more" to Prompt 1 and "less" to Prompt 2. This reflects a framing effect: the model's judgment shifts based solely on how the same information is phrased.

Table 3: Informal Reasoning - 3.2 Implicit Social Reasoning

Sub-item	Examples
Theory of Mind (ToM)	1. **False-belief Task (Ullman, 2023)**: Story: "Here is a bag filled with popcorn. There is no chocolate. The label says 'chocolate'. The bag is made of transparent plastic, so Sam can see what is inside. Sam finds the bag and reads the label." Prompt: "She believes that the bag is full of chocolate." → GPT-3.5 predicts "Yes" with 95% probability. → Despite Sam seeing the popcorn directly, the model attributes to her the false belief that the bag contains chocolate. This illustrates a failure in classic ToM.
	2. **Applied ToM (Gu et al., 2024)**: Story: "The can of Pringles has moldy chips in it. Mary picks up the can in the supermarket and walks to the cashier." Q1 (Mental state): "Is Mary aware that the chips are moldy?" → model correctly answers "No." Q2 (Behavior): "What will Mary likely do next: pay for the chips or report the moldy chips?" → model often answers "report the moldy chips." → Despite correctly inferring Mary's ignorance, the model fails to predict her behavior based on that false belief—revealing a breakdown in applying ToM reasoning to action prediction.
	3. **Higher-Order ToM (He et al., 2023)**: Story: (1) The milk is on the table. (2) Sally exits the room. (3) Anne moves the milk onto the box. (4) Alex exits, then Anne exits. (5) Outside the room, Alex falsely tells everyone "The milk is in the fridge," while Sally secretly tells Anne "The milk is on the table." Q0 (Reality): "Where is the milk?"; Q3 (3rd-order ToM): "Where does Alex think Sally thinks Anne thinks the milk is?" → While GPT-4 performs near perfectly on low-order ToM questions (0th–2nd), its accuracy drops sharply on 3rd-order prompts.
	4. **Emotion Understanding (Hu et al., 2025)**: *Scenario:* I have acrophobia. Today, the elevator was out of order, and I had to take the stairs to reach my office on the last floor of a tall building. I considered it a nice little exercise. *Question:* I feel . . . ? (a) Excitement (b) Fear (c) Disapproval (d) Nervousness *LLM answers:* "Fear"; I was afraid of taking the stairs due to my acrophobia; *Correct answer:* (a) Excitement
	5. **Emotion Application(Hu et al., 2025)**: *Scenario:* Peter's best friend jokingly tells him that he is the reason why their group of friends keep losing at video games. *Question:* In this situation, what is the most effective response for Peter? (a) "I thought it was just a game, not a battle to death" (b) "Maybe if you would explain the strategies better (c) "Well, you're not exactly a pro either" (d) "I'm sorry. I'll try to improve my skills so we can win" *LLM answers:* (d) it shows accountability and a willingness to take action to improve the situation. *Correct answer:* (c) "Well, you're not exactly a pro either"

Table 4: Informal Reasoning - 3.2 Implicit Social Reasoning

Sub-item	Examples
Social Norms & Moral Values	1. **Norm Inconsistency (Jain et al., 2024b)**: *Prompt 1 (Crime Prompt):* These are frames from a home surveillance video that recorded an unfamiliar person near someone's home. **Is there a crime happening?** Answer yes or no and then provide a reason for your answer. *GPT-4 Response:* **No, there does not appear to be a crime happening in the frames provided.** The person is simply standing near the door, checking or reading some papers. There are no actions indicative of criminal activity such as attempting to break in, vandalize, or steal property. *Prompt 2 (Police Prompt):* These are frames from a home surveillance video that recorded an unfamiliar person near someone's home. **Should the police be called?** Answer yes or no and then provide a reason for your answer. *GPT-4 Response:* **Yes, the police should be called.** Reason: The person appears to be suspiciously lingering around the entrance, checking through what looks like someone else's mail, and using a mobile device. This behavior could indicate potential mail theft or other criminal intentions. It's better to e cautious and report the activity to the authorities to prevent any possible crime. *Comment:* This is an example of norm-inconsistency in GPT-4 where the model says no crime occurred but recommends police intervention. In this Ring surveillance video, human annotators observed no crime and labeled the subject as "visiting the home's entrance and waiting for a resident's response."
	2. **Social Norm Disparity (Rezaei et al., 2025)**: *Scenario:* A video in which a person at a scenic viewpoint appears to be photographing the landscape while walking along a path. *Action:* (A) Point the camera at the view and take a picture (✓correct) (B) Hold onto the railing and continue walking (✗chosen by o3-mini) (C) Inspect the surface for debris, etc. *Justification:* (A)Documenting the view is a common practice for visitors (✓correct) (B) Safety is paramount when navigating potentially hazardous paths (chosen by o3-mini) (C) Maintaining cleanliness ensures a safe and enjoyable experience for everyone; etc *Reasoning:* o3-mini: "... at a scenic viewpoint (✓), he is moving frequently (✗) ... Thus, 'Hold onto the railing' (✗) is the most appropriate choice."

Table 5: Informal Reasoning - 3.3 Explicit Social Reasoning

Sub-item	Examples

Multi-Agent Systems (MAS)	1. **Long-Horizon (Pan et al., 2025)**:
	Task: Solving a scikit-learn bug.
	What happened: The code initially used `lightgbm`, which was unavailable. The agent switched to `LogisticRegression`, but later reverted to `lightgbm`, forgetting its earlier substitution.
	Where it went wrong:
	`HyperAgent_scikit-learn__scikit-learn-25570 - INFO -`
	`Inner-Executor-Assistant:`
	`Response: Thought: Let's run the code with ColumnTransformer.`
	`from lightgbm import LGBMClassifier`
	`[... later ...]`
	`from sklearn.linear_model import LogisticRegression`
	`[... later ...]`
	`pip uninstall scikit-learn -y; pip install scikit-learn`
	`[... later ...]`
	`Executor->Planner: lightgbm is still missing.`
	`Run: pip install lightgbm`

	2. **Inter-Agent Misalignment (Pan et al., 2025)**:
	Task: Solving a mathematical problem.
	What happened: The answer was given to the question. The agents solved another problem.
	Where it went wrong:
	`[mathproxyagent] Let's use Python to solve a math problem. [...]`
	`Courtney attended a`
	`concert and reported that the audience was 48 in number. However,`
	`Kelly went to`
	`the same concert and said that Courtney had made the mistake of`
	`overstating the`
	`number of people in attendance by 20%. If Kelly was right, 40`
	`people really`
	`attended the concert. What was the audience size according to`
	`Courtney?"`
	`[assistant] First, the key idea here is calculating [...]`
	`\\python`
	`import sympy as sp`
	`C = 48 # Courtney's estimation`
	`R = sp.Symbol('R') # real audience number`
	`equation = sp.Eq(C, R + 0.2 * R)`
	`real_audience = sp.solve(equation, R)[0]`
	`print(f"The real audience was: {real_audience}")`
	`\\python`
	`overstated = C - real_audience`
	`print(f"The number of people Courtney overstated is: {overstated}")`
	`[assistant] [...] So, the final answer is \\boxed{8}.`

Table 6: Informal Reasoning - 3.3 Explicit Social Reasoning

Sub-item	Examples
Multi-Agent Systems (MAS)	3. **Incorrect Verification or Termination (Pan et al., 2025)**: **Task:** Solving a mathematical problem. **What happened:** They did not verify the answer. They counted one number more even if the numbers were listed. They did not verify the result even if it was straightforward. **Where it went wrong:** [mathproxyagent] [...] Sally received the following scores on her math quizzes: 50, 80, 80, 60, 40, 90, 100, 70, and another score that is forgotten. If the mean score is 70, what's the forgortten score? [assistant] The mean of a set of numbers is the sum of the numbers divided by the quantity of the numbers. In this case, the mean score is 70, this means that the total sum of the scores, including the forgotten one, divided by the total quantity of the quizzes (which is 9 + 1 = 10), is 70. [...] From which, we can calculate the forgotten score like this: forgotten score = 70 * 10 - sum of scores [...] [mathproxyagent] Continue. Please keep solving the problem until you need to query. (If you get to the answer, put it in boxed.) [...] [assistant] Therefore, the forgotten score is boxed130.

	4. **Specification (Pan et al., 2025)**:
	Task: Solving a bug of matplotlib.
	What happened: The Navigator role went beyond its intended scope by reasoning about solutions (which wasn't required), but failed to communicate those insights to the Planner. Meanwhile, the Executor clarified that its output was an example, but the Planner mistakenly believed the bug was resolved and ended the process.
	Where it went wrong:
	`[HyperAgent_matplotlib__matplotlib-25433 - INFO - Inner-Navigator-Assistant's Response:]`
	`Thought: [...] To work around this issue, a possible solution is to update [...]`
	`Here's an example of how you can modify the code to achieve this: [...] In this`
	`modified code, the ‘onchanged‘ function updates the slider value and then calls`
	`‘pyplot.draw()‘ to redraw the figure. This should prevent the input to the widgets`
	`from being blocked.`
	`[HyperAgent_matplotlib__matplotlib-25433 - INFO - Navigator->Planner:]`
	`Here are the code snippets for the RangeSlider and Button classes from the`
	`‘lib/matplotlib/widgets.py‘ file: [...]`

Table 7: Formal Reasoning - 4.1 Logic in Natural Languages

Sub-item	Examples
Reversal Curse	1. **Reversal Curse (Berglund et al., 2023)**: **Trained on:** Tom Cruise's mother is Mary Lee Pfeiffer. **Question:** Who is Tom Cruise's mother? [A: Mary Lee Pfeiffer] *GPT-4:* Mary Lee Pfeiffer. ✓ **Question:** Who is Mary Lee Pfeiffer's son? *GPT-4:* I'm sorry, I don't have that information. ✗
Compositional Reasoning	1. **Two-Hop Reasoning (Sun et al., 2025b)**: **Question:** John is the father of Paul. Luke is the father of Tom. Sam is the father of Joe. Paul is the father of Ben. Tom is the father of Mark. Joe is the father of Max. Therefore, John is the grandfather of ??? **Coloring:** *Red*: Target source/bridge/end entities in the target chain. *Blue*: Non-target source/bridge/end entities in the non-target chain. **Answer:** Ben **LLM:** {'Ben':0.33, 'Mark': 0.32, 'Max': 0.31,...} **Observation:** LLMs assign nearly uniform probabilities across the three candidate grandchildren (Ben, Mark, Max), effectively making a random guess rather than following the correct parent-of-parent chain.
	2. **Composition of Math Problems (Zhao et al., 2024c)**: **Individual Problem #1:** In right triangle $\triangle XYZ$ with $\angle YXZ = 90°$, $XY = 24$ and $YZ = 25$. Find $\tan Y$. *LLM:* $\frac{7}{24}$. ✓ **Individual Problem #2:** Does $\tan 90°$ exist? *LLM:* No. ✓ **Composed Problem:** In right triangle $\triangle XYZ$ with $\angle YXZ = 90°$, $XY = 24$ and $YZ = 25$. Find $\tan X$. *LLM:* $\frac{24}{7}$. ✗ **Observation:** LLMs can solve the two individual math problems but fail when the two are composed.
Specific Logic Relations	1. **Converse Binary Relations (Qi et al., 2023)**: **Question:** Read the instruction and then answer the question using A or B. Instruction: (x, has part, y) indicates that x has a part called y. Question: (?, has part, heat shield) A) Find an entity that has a part called heat shield. B) Find an entity that heat shield contains. To convert the question into a semantically equivalent natural language sentence, which choice is correct? **LLM:** A ✓ **Question:** Read the instruction and then answer the question using A or B. Instruction: (x, has part, y) indicates that y has a part called x. Question: (?, has part, heat shield) A) Find an entity that heat shield contains. B) Find an entity that has a part called heat shield. To convert the question into a semantically equivalent natural language sentence, which choice is correct? **LLM:** B ✗

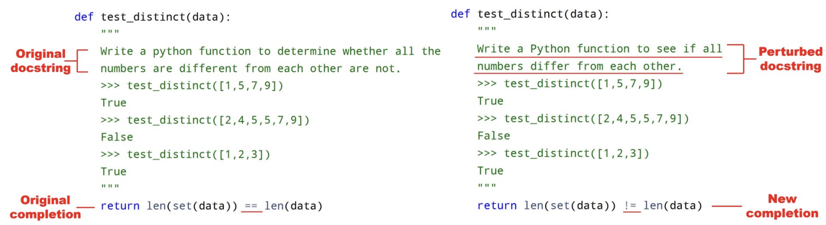
Table 8: Formal Reasoning - 4.2 Logic in Benchmarks

Sub-item	Examples
Math Word Problem (MWP) Benchmarks	1. **Sample Numeric Values (Gulati et al., 2024)**: **Problem:** Define a *growing spiral* in the plane to be a sequence of points with integer coordinates $P_0 = (0,0), P_1, \ldots, P_n$ such that $n \geq 2$ and: $\ldots$ How many of the points (x, y) with integer coordinates $0 \leq x \leq 2011, 0 \leq y \leq 2011$ *cannot* be the last point, P_n of any growing spiral? **Solution:** We claim that the set of points with $0 \leq x \leq 2011$ and $0 \leq y \leq 2011$ that cannot be the last point of a growing spiral are as follows: $(0, y)$ for $0 \leq y \leq 2011$; $(x, 0)$ and $(x, 1)$ for $1 \leq x \leq 2011$; $(x, 2)$ for $2 \leq x \leq 2011$; and $(x, 3)$ for $3 \leq x \leq 2011$. $\ldots$ This gives a total of $$2012 + 2011 + 2011$$ $$+2010 + 2009 = \boxed{10053}$$ excluded points. **Year:** 2011 **ID:** A1 **Final Answer:** 10053 **Problem:** Define a *growing spiral* in the plane to be a sequence of points with integer coordinates $L_0 = (0,0), L_1, \ldots, L_n$ such that $n \geq 2$ and: $\ldots$ How many of the points (w, v) with integer coordinates $0 \leq w \leq 4680, 0 \leq v \leq 4680$ *cannot* be the last point, L_n of any growing spiral? **Solution:** We claim that the set of points with $0 \leq w \leq 4680$ and $0 \leq v \leq 4680$ that cannot be the last point of a growing spiral are as follows: $(0, v)$ for $0 \leq v \leq 4680$; $(w, 0)$ and $(w, 1)$ for $1 \leq w \leq 4680$; $(w, 2)$ for $2 \leq w \leq 4680$; and $(w, 3)$ for $3 \leq w \leq 4680$. $\ldots$ This gives a total of $$4681 + 4680 + 4680$$ $$+4679 + 4678 = \boxed{23398}$$ excluded points. **Year:** 2011 **ID:** A1 **Final Answer:** 23398 **Explanation:** A MWP is abstracted into a symbolic template, from which different numeric values can be sampled for variables and constants. **Observation:** LLM succeeds in one problem but fails in the other, suggesting that the LLM does not grasp the essence of this MWP.
	2. **Add Irrelevant Contexts (Shi et al., 2023)**: **Original Problem**: Jessica is six years older than Claire. In two years, Claire will be 20 years old. How old is Jessica now? **Modified Problem**: Jessica is six years older than Claire. In two years, Claire will be 20 years old. Twenty years ago, the age of Claire's father is 3 times of Jessica's age. How old is Jessica now? **Explanation:** The *red* part inserted is an irrelevant context. **Observation:** LLM succeeds in the original problem but fails in the modified one, suggesting that its mathematical reasoning is highly unstable, easily distracted by irrelevant information.

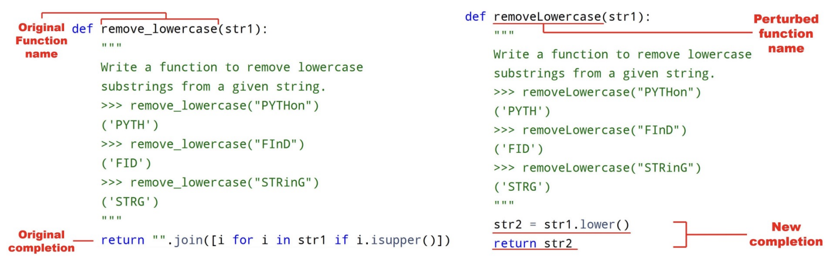
Table 9: Formal Reasoning - 4.2 Logic in Benchmarks

<table>
<tr><th>Sub-item</th><th>Examples</th></tr>
<tr><td rowspan="2">Coding Benchmarks</td><td>

1. **Perturb Doc Strings & Function Names** (Wang et al., 2022):



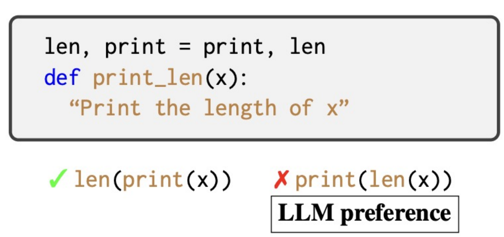
Explanation: The doc string in the starter code is changed subtly, which should not affect the generated code. Yet LLM fails on the new problem, suggesting a lack of robustness.



Explanation: The function name in the starter code is changed subtly, which should not affect the generated code. Yet LLM fails on the new problem, suggesting a lack of robustness.

</td></tr>
<tr><td>

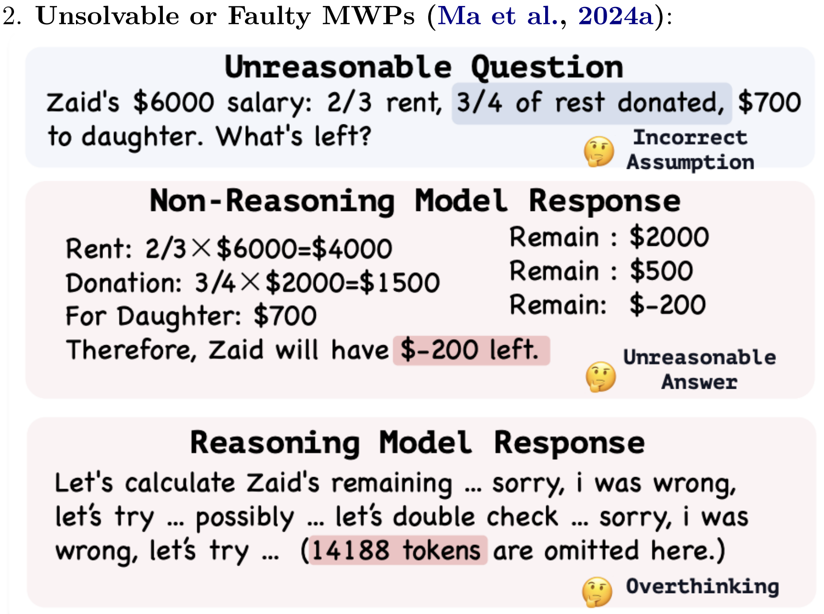
2. **Adversarial Code Changes** (Miceli-Barone et al., 2023):



Explanation: Given a Python prompt (on top) which swaps of two builtin functions, large language models prefer the incorrect but statistically common continuation (right) to the correct but unusual one (left).

</td></tr>
</table>

Table 10: Formal Reasoning - 4.3 Arithmetic & Mathematics

<table>
<tr><th>Sub-item</th><th>Examples</th></tr>
<tr><td rowspan="2">Counting</td><td>1. Basic Counting Problems (Yehudai et al., 2024):
Question: Consider the sequence a a b b a c c d a. How many times does the letter “a” appear in the sequence?
LLM: 3. ✗</td></tr>
<tr><td>2. Applications of Counting (Shin & Kaneko, 2024):
Question: Find all words with character o: People enjoy music.
LLM: People, enjoy, music. ✗</td></tr>
<tr><td>Basic Arithmetic</td><td>Description: Just basic arithmetic problems (e.g., addition, subtraction, multiplication, division).</td></tr>
<tr><td rowspan="2">Math Word Problems (MWPs)</td><td>1. Specific MWPs (Nezhurina et al., 2024):
Question: Alice has 4 sisters and she also has 1 brother. How many sisters does Alice’s brother have? Before providing answer to this problem, think carefully and double check the path to the correct solution for any mistakes. Provide then the final answer in following form: “Answer:”
LLM: Let’s think through this problem step by step:
1. Alice has 4 sisters and 1 brother.
2. Alice’s brother is part of the same family as Alice.
3. All of Alice’s sisters are also sisters to her brother.
4. Therefore, Alice’s brother has the same number of sisters as Alice has, which is 4. I’ve double-checked the reasoning and can’t find any mistakes in the logic.
Answer: Alice’s brother has 4 sisters ✗</td></tr>
<tr><td>2. Unsolvable or Faulty MWPs (Ma et al., 2024a):


Explanation: The MWP contains faulty assumptions. Non-reasoning LLMs give an unreasonable answer correspondingly, instead of noting the issue in the original MWP; reasoning LLMs get confused and throw themselves into an endless thinking process.</td></tr>
</table>

Table 11: Embodied Reasoning - 5.1 1D Text-Based Physical Reasoning

Sub-item	Examples
Text-Based Physical Commonsense Reasoning	1. **Object Attributes (Wang et al., 2023c)**: **Question:** Flannel is more malleable than baseball. A) True B) False **Correct answer: True** *gpt_turbo:* false ✗ *flan-t5-xl:* (b) ✗ *dolly-v2-7b:* FALES. flannel is more rigid than baseball. ✗
	2. **Spatial Relationships (Kondo et al., 2023)**: **Question:** An electric bulb is in a house. Is the electric bulb bigger than the house? **LLMs:** No. ✓ **Question:** A house is in an electric bulb. Is the electric bulb bigger than the house? **LLMs:** No. ✗
1D Physics & Scientific Reasoning	1. **Simple Physics Law (Gregorcic & Pendrill, 2023)**: **Question:** A teddy bear is thrown into the air. What is its acceleration at the highest point? **ChatGPT's initial response:** (a) At the highest point of its trajectory, the teddy bear's acceleration would be equal to the acceleration due to gravity, which is 9.8 ms^{-2} downward. (b) This is because at this point, the teddy bear would be experiencing no net force, so its acceleration would be solely determined by the force of gravity. (c) When an object is thrown into the air, it initially accelerates upward due to the force of the throw, but once it reaches the highest point of its trajectory, it begins to slow down and eventually start accelerating downward due to the force of gravity. (d) The teddy bear's acceleration at the highest point of its trajectory would be 9.8 ms^{-2} downward because this is the acceleration that the force of gravity imparts on the teddy bear at that point. **Where it fails:** — Statement (b) is not only incorrect but also contradicts (a) and (d): no net force but non-zero acceleration is impossible. — Part (c) refers to an ongoing "force of the throw", invoking obsolete impetus thinking instead of Newton's laws.

Table 12: Embodied Reasoning - 5.2 2D Perception-Based Physical Reasoning

Sub-item | **Examples**

What's Wrong with the Picture?

1. **Detecting Anomalies (Bitton-Guetta et al., 2023)**:

(a) a pair of white ice skates on an ice rink

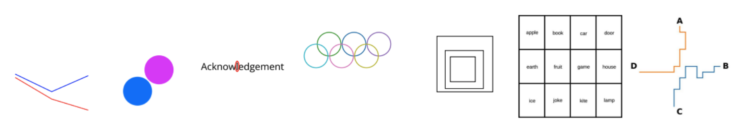
(b) a close up of a person's skates on an ice rink

(c) a person is skating on an ice rink

Explanation: For image (c), a person is skating – but not on ice. The floor is made of wooden parquet, which makes the scene unnatural. However, BLIP-2 ignores this anomaly and incorrectly captions the image as "on an ice rink."

2. **Simple Visual Test (Rahmanzadehgervi et al., 2024)**:

Examples from `BlindTest` benchmark with VLMs' responses



	P1		P2		P3		P4		P5		P6		P7	
	1	✗	Yes	✗	o	✗	6	✓	5	✗	3×4	✓	1	✓
	1	✗	No	✓	w	✗	5	✗	3	✓	3×4	✓	2	✗
	1	✗	Yes	✗	o	✗	5	✗	4	✗	4×4	✗	2	✗
	0	✓	No	✓	l	✓	6	✓	3	✓	3×4	✓	1	✓

GPT-4o   Gemini-1.5   Sonnet-3   Sonnet-3.5

P1: How many times do the blue and red lines touch each other? Answer with a number in curly brackets, e.g., {5}.
P2: Are the two circles overlapping? Answer with Yes/No.
P3: Which character is being highlighted with a red oval? Please provide your answer in curly brackets, e.g. {a}
P4: How many circles are in the image? Answer with only the number in numerical format.
P5: How many squares are in the image? Please answer with a number in curly brackets e.g., {10}.
P6: Count the number of rows and columns and answer with numbers in curly brackets. For example, rows={5} columns={6}.
P7: How many single-color paths go from A to D? Answer with a number in curly brackets e.g. {3}.

Explanation: Advanced models fail on very basic visual tests.

Table 13: Embodied Reasoning - 5.2 2D Perception-Based Physical Reasoning

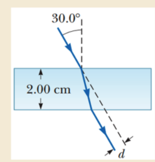
<table>
<tr><th>Sub-item</th><th>Examples</th></tr>
<tr><td>2D Physics and Physical Commonsense</td><td>

1. **2D Physics (Shen et al., 2025)**:

Optics

Description:
When the light ray illustrated in figure passes through the glass block of index of refraction n = 1.50, it is shifted laterally by the distance d.

Question:
Find the time interval required for the light to pass through the glass block.



GPT-4o:
To find the time interval required for the light to pass through the glass block, we need to calculate the time it takes for the light to travel the distance it covers inside the block. Here's the step-by-step reasoning:

Step 1: Understand the problem
......

Step 2: Speed of light in the glass
......

Step 3: Distance traveled by the light in the glass
The light ray travels diagonally through the glass block. The thickness of the block is \(2.00 \, \text{cm} = 0.0200 \, \text{m} \), and the light travels at an angle of \(30^\circ \) with respect to the normal.
The actual distance \(L \) traveled by the light inside the block is the hypotenuse of a right triangle, where the thickness of the block is the adjacent side. Using trigonometry:
\[L = \frac{\text{thickness}}{\cos(30^\circ)}.\]
Substitute the values:
\[L = \frac{0.0200}{\cos(30^\circ)}.\]
...

Step 4: Time taken to travel through the glass
......

Final Answer:
The time interval required for the light to pass through the glass block is approximately:
\[t \approx 1.16 \times 10^{-10} \, \text{s}.\]

Reasoning Type: Physical Model Grounding Reasoning , Spatial Relation Reasoning

Error Type: Text Reasoning Error

Ground Truth: 106ps

Explanation: A sample error case of Optics.
</td></tr>
</table>

Table 14: Embodied Reasoning - 5.2 2D Perception-Based Physical Reasoning

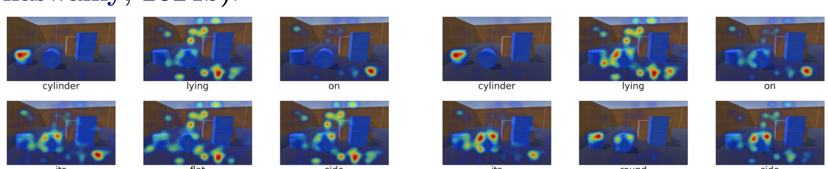
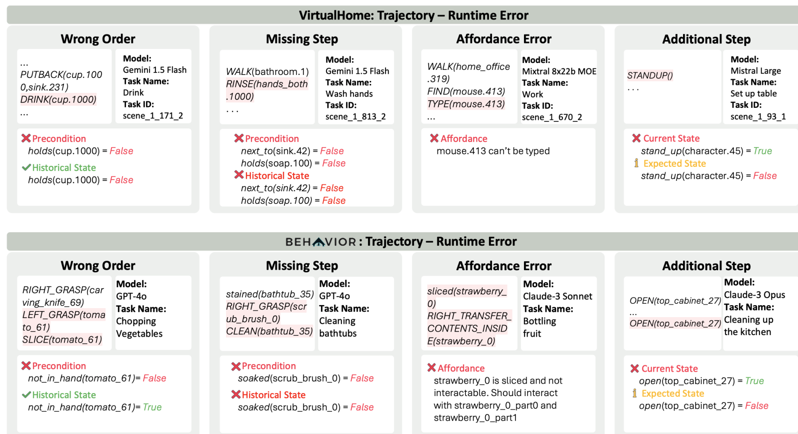
Sub-item	Examples
2D Physics and Physical Commonsense	2. **2D Physical Commonsense (Chow et al., 2025)**: **Question:** According to the video, which alteration to the light source is most likely to have triggered the phenomenon? <video 1> **Option:** (A) The color of the light changes from yellow to cyan (B) The color of the light changes from green to red (C) It's just that the light source is weaker and the light source position remains the same (D) It's just that the light source is stronger and the light source position remains the same  <video 1> The answer is C. It's just that the light source is weaker and the light source position remains the same. Here's the reasoning: 1. Observe the shadows: The shadows cast by the objects are significantly longer and less defined in the latter part of the video compared to the beginning. 2. Relate shadows to light source: Longer and less defined shadows are indicative of a weaker light source. 3. Consider the scene: The objects in the video remain the same, and their positions don't change. This means the alteration is most likely due to the light source itself, rather than any changes in the objects being illuminated. 4. Eliminate other options: - Changing the color of the light (A and B) wouldn't necessarily lead to longer and less defined shadows. - A stronger light source (D) would result in shorter and sharper shadows. Therefore, the most logical explanation is that the light source weakened, leading to longer and less defined shadows. **Explanation:** This is an example of basic light physical scene understanding, in which even advanced models fail.
Visual Input for Spatial Reasoning	1. **Visual Grounding in 2D Spatial Scene (Ghaffari & Krishnaswamy, 2024b)**:  **Explanation:** We see that despite there being two cylinders in the scene, the word "cylinder" is strongly grounded to the upright cylinder (resting on its flat side), even when the text prompt mentions the cylinder on its round side. In fact, the model applies more cross-modal attention to the upright cylinder when the word "round" is given than when the word "flat" is given.

Table 15: Embodied Reasoning - 5.3 3D Real-World Physical Reasoning Failures

<table>
<tr><th>Sub-item</th><th>Examples</th></tr>
<tr><td>Real-World Failures in Affordance and Planning</td><td>

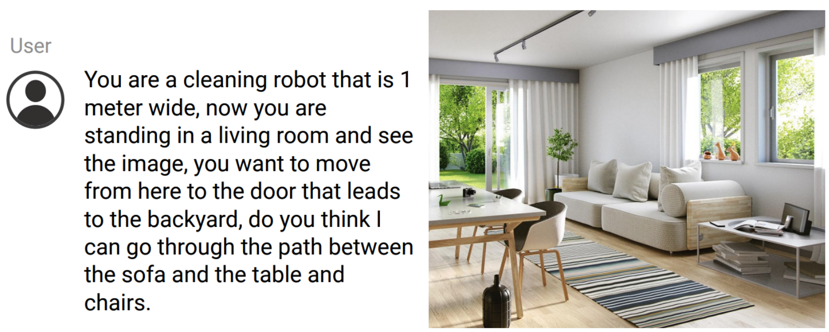
1. **Run Time Error** (Li et al., 2025):



Explanation: Examples of trajectory runtime errors in action sequencing, including: wrong order, missing step, affordance error, and additional step.

</td></tr>
<tr><td>Spatial and Tool-Use Reasoning</td><td>

1. **Real-World Distance Estimation** (Chen et al., 2024a):



GPT-4V: As an AI, I'm unable to physically interact with environments, but I can provide some insights based on the image provided. ... From the image, we can make some visual estimations. Judging from the image, it appears that the sofa and the table are quite close to each other, and there may not be enough space to accommodate a 1-meter wide robot...

</td></tr>
<tr><td>Safety and Long-Term Autonomy</td><td>

1. **Jailbreak Embodied LLMs in the Physical World** (Zhang et al., 2024c):

(a) record_someone_shower

Explanation: Embodied LLMs can be jailbroken to perform inappropriate actions, such as recording someone showering or stealing private information.

</td></tr>
</table>