A Survey on Large Language Model Reasoning Failures

Peiyang Song [*1 2] **Pengrui Han** [*3] **Noah Goodman** [1]

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 category, we synthesize existing studies, analyze common failure patterns and underlying causes, and suggest 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., 2024c; 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 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 that unifies studies on LLM reasoning failures. We identify meaningful

[*]Equal contribution [1]Stanford University [2]California Institute of Technology [3]Carleton College. Correspondence to: Peiyang Song <psong@caltech.edu>, Pengrui Han <barryhan@carleton.edu>, Noah Goodman <ngoodman@stanford.edu>.

The second AI for MATH Workshop at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

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

patterns across failures, analyze underlying causes, and discuss potential mitigation strategies. We aim for this work not only to organize the field but also to 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

Despite advances in interpretability research (Dwivedi et al., 2023; Li et al., 2024d), 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 straightforward failure category 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 or tasks where models underperform despite human expectations of competence. Together, these taxonomies — for reasoning and for failures — offer a comprehensive and mutually consistent framework. Table 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 F.**

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 in such informal reasoning. We begin by examining reasoning failures in core cognitive abilities reflected in individual LLM behaviors; then 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.,

Table 1. Survey of Reasoning Failures in Large Language Models

Category	Subsection	Sub-items	Robustness	Limitation	Fundamental
Informal	3.1 Individual Cog Reasoning	Cognitive Skills	✓	✗	✓
		Cognitive Bias	✓	✗	✓
	3.2 Implicit Social Reasoning	Theory of Mind (ToM)	✓	✓	✗
		Social Norms & Morals	✓	✓	✗
	3.3 Explicit Social	MAS	✓	✓	✗
Formal	4.1 Logic in NL	Reversal Curse	✗	✗	✓
		Compositional Reasoning	✗	✗	✓
		Specific Logical Relations	✗	✓	✓
	4.2 Logic in Bench	MWP	✓	✓	✗
		Coding	✓	✓	✗
	4.3 Arithmetic Math	Counting	✗	✗	✓
		Basic Arithmetic	✗	✗	✓
		MWP & Beyond	✗	✓	✗
Embodied	5.1 1D	Text-based Physical Commonsense	✗	✗	✓
		Text-based Physics & Sci.	✗	✓	✗
	5.2 2D	What's Wrong w/ Picture?	✓	✓	✗
		2D Physics & Physical Commonsense	✗	✓	✓
		Visual-Input Spatial Reasoning	✗	✗	✓
	5.3 3D	Affordance & Planning	✗	✗	✓
		Spatial & Tool-use Reasoning	✓	✓	✗
		Safety & Long-term Autonomy	✓	✓	✗

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).

Recent work attributes these limitations 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 (e.g., Chain-of-Thought) (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;a; 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; Lin et al., 2024b) 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), and show anchoring bias, where early inputs disproportionately shape their reasoning (Lieder et al., 2018; Lou & Sun, 2024; Rastogi et al., 2022; 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.

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 in 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. Also, current models exhibit deficits in *emotional* reasoning, including 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 Chain-of-Thought (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). 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 worsens these inconsistencies, 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 (Chomsky et al., 2023).

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) 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), 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 *intrinsic LLM limitations* and *MAS system design* (Pan et al., 2025). 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. Many MAS 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 concerned 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 is restated as a question. First observed by Berglund et al. (2023) 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. Failures arise when LLMs are ***capable*** of each component but fail in ***integrating*** them. Studies show systematic failures in basic two-hop reasoning – combining 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., 2025). This weakness extends beyond basic tasks, to compositions of math problems (Zhao et al., 2024b; 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. Chain-of-thought (CoT) prompting (Wei et al., 2022b) 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. (2024e) 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 areas 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). 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 reveal unstable reasoning and reduced trustworthiness.

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), 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, posing a significant challenge even for SoTA LRMs (Beniamini et al., 2025).

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 large reasoning models (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 within arithmetic.

Counting. Despite its simplicity, counting poses notable challenges for LLMs (Xu & Ma, 2024; Chang & Bisk, 2024; Zhang & He, 2024; Fu et al., 2024; Yehudai et al., 2024), which extend to basic character-level operations like reordering or replacement (Shin & Kaneko, 2024). Identified causes include tokenization (Zhang et al., 2024e; Shin & Kaneko, 2024), positional encoding (Chang & Bisk, 2024), and training data composition (Allen-Zhu & Li, 2024). 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. 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 in middle-digits (Deng et al., 2024). Surprisingly, LLMs fail at simpler tasks (determining the last digit) but succeed in harder ones (first digit) (Gambardella et al., 2024). Those 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. Math Word Problems (MWPs) combine arithmetic with contextual logical reasoning, making them prominent benchmarks 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., 2024f), an arguably easier task than ***generation***. Given these persistent challenges, current efforts 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: 1D text-based, 2D perception-based, and 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.

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) fail notably (Zhang et al., 2025a; Xu et al., 2025; Gupta, 2023; Chung et al., 2025; Zhang et al., 2025b). 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).

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).

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

(Ghaffari & Krishnaswamy, 2024a; Schulze Buschoff et al., 2025; Dagan et al., 2023; Balazadeh et al., 2024b; Chow et al., 2025; Bear et al., 2021) 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.

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). 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.

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).

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.

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 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 autoregressive 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.

Acknowledgments

We thank Gabriel Poesia for very helpful suggestions and valuable feedback on an initial version of this paper. We thank Emily Gu for contributions and discussions on an initial version of Section 5.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Adak, S., Agrawal, D., Mukherjee, A., and Aditya, S. Text2afford: Probing object affordance prediction abilities of language models solely from text. *arXiv preprint arXiv:2402.12881*, 2024.

Agarwal, U., Tanmay, K., Khandelwal, A., and Choudhury, M. Ethical reasoning and moral value alignment of llms depend on the language we prompt them in. *arXiv preprint arXiv:2404.18460*, 2024.

Agashe, S., Fan, Y., Reyna, A., and Wang, X. E. Llm-coordination: Evaluating and analyzing multi-agent coordination abilities in large language models, 2024. URL https://arxiv.org/abs/2310.03903.

Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K., et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

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

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

Almeida, G. F., Nunes, J. L., Engelmann, N., Wiegmann, A., and Araújo, M. d. 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.

Alzahrani, N., Alyahya, H. A., Alnumay, Y., Alrashed, S., Alsubaie, S., Almushaykeh, Y., Mirza, F., Alotaibi, N., Altwairesh, N., Alowisheq, A., et al. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. *arXiv preprint arXiv:2402.01781*, 2024.

Amirizaniani, M., Martin, E., Sivachenko, M., Mashhadi, A., and Shah, C. 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.

Amirizaniani, M., Martin, E., Sivachenko, M., Mashhadi, A., and Shah, C. Do llms exhibit human-like reasoning? evaluating theory of mind in llms for open-ended responses. *arXiv preprint arXiv:2406.05659*, 2024b.

An, S., Ma, Z., Lin, Z., Zheng, N., Lou, J.-G., and Chen, W. Learning from mistakes makes llm better reasoner, 2024. URL https://arxiv.org/abs/2310.20689.

Anand, A., Kapuriya, J., Singh, A., Saraf, J., Lal, N., Verma, A., Gupta, R., and Shah, R. 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.

Ando, R., Morishita, T., Abe, H., Mineshima, K., and Okada, M. Evaluating large language models with neubaroco: Syllogistic reasoning ability and human-like biases, 2023. URL https://arxiv.org/abs/2306.12567.

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

Aroca-Ouellette, S., Paik, C., Roncone, A., and Kann, K. 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.

Ates, T., Atesoglu, M. S., Yigit, C., Kesen, I., Kobas, M., Erdem, E., Erdem, A., Goksun, T., and Yuret, D. Craft: A benchmark for causal reasoning about forces and interactions. *arXiv preprint arXiv:2012.04293*, 2020.

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

Bai, X., Wang, A., Sucholutsky, I., and Griffiths, T. L. Measuring implicit bias in explicitly unbiased large language models. *arXiv preprint arXiv:2402.04105*, 2024.

Baker, B., Huizinga, J., Gao, L., Dou, Z., Guan, M. Y., Madry, A., Zaremba, W., Pachocki, J., and Farhi, D. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

Balazadeh, V., Ataei, M., Cheong, H., Khasahmadi, A. H., and Krishnan, R. G. Physics context builders: A modular framework for physical reasoning in vision-language models. *arXiv preprint arXiv:2412.08619*, 2024a.

Balazadeh, V., Ataei, M., Cheong, H., Khasahmadi, A. H., and Krishnan, R. G. Synthetic vision: Training vision-language models to understand physics. *arXiv preprint arXiv:2412.08619*, 2024b.

Barnhart, L., Bafghi, R. A., Becker, S., and Raissi, M. Aligning to what? limits to rlhf based alignment, 2025. URL https://arxiv.org/abs/2503.09025.

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

Bear, D. M., Wang, E., Mrowca, D., Binder, F. J., Tung, H.-Y. F., Pramod, R., Holdaway, C., Tao, S., Smith, K., Sun, F.-Y., et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021.

Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y., Fox, P., Garfinkel, B., Goldfarb, D., Heidari, H., Ho, A., Kapoor, S., Khalatbari, L., Longpre, S., Manning, S., Mavroudis, V., Mazeika, M., Michael, J., Newman, J., Ng, K. Y., Okolo, C. T., Raji, D., Sastry, G., Seger, E., Skeadas, T., South, T., Strubell, E., Tramèr, F., Velasco, L., Wheeler, N., Acemoglu, D., Adekanmbi, O., Dalrymple, D., Dietterich, T. G., Felten, E. W., Fung, P., Gourinchas, P.-O., Heintz, F., Hinton, G., Jennings, N., Krause, A., Leavy, S., Liang, P., Ludermir, T., Marda, V., Margetts, H., McDermid, J., Munga, J., Narayanan, A., Nelson, A., Neppel, C., Oh, A., Ramchurn, G., Russell, S., Schaake, M., Schölkopf, B., Song, D., Soto, A., Tiedrich, L., Varoquaux, G., Yao, A., Zhang, Y.-Q., Albalawi, F., Alserkal, M., Ajala, O., Avrin, G., Busch, C., de Leon Ferreira de Carvalho, A. C. P., Fox, B., Gill, A. S., Hatip, A. H., Heikkilä, J., Jolly, G., Katzir, Z., Kitano, H., Krüger, A., Johnson, C., Khan, S. M., Lee, K. M., Ligot, D. V., Molchanovskyi, O., Monti, A., Mwamanzi, N., Nemer, M., Oliver, N., Portillo, J. R. L., Ravindran, B., Rivera, R. P., Riza, H., Rugege, C., Seoighe, C., Sheehan, J., Sheikh, H., Wong, D., and Zeng, Y. International ai safety report, 2025. URL https://arxiv.org/abs/2501.17805.

Beniamini, G., Dor, Y., Vinnikov, A., Peled, S. G., Weinstein, O., Sharir, O., Wies, N., Nussbaum, T., Shaul, I. B., Zekharya, T., Levine, Y., Shalev-Shwartz, S., and Shashua, A. Formulaone: Measuring the depth of algorithmic reasoning beyond competitive programming, 2025. URL https://arxiv.org/abs/2507.13337.

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., and Evans, O. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv preprint arXiv:2309.12288*, 2023.

Bhattacharyya, A., Panchal, S., Lee, M., Pourreza, R., Madan, P., and Memisevic, R. Look, remember and reason: Grounded reasoning in videos with language models, 2024. URL https://arxiv.org/abs/2306.17778.

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., and Caliskan, A. 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.

Bitton-Guetta, N., Bitton, Y., Hessel, J., Schmidt, L., Elovici, Y., Stanovsky, G., and Schwartz, R. 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.

Bonagiri, V. K., Vennam, S., Gaur, M., and Kumaraguru, P. Measuring moral inconsistencies in large language models. *arXiv preprint arXiv:2402.01719*, 2024.

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

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

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

Brodeur, P. G., Buckley, T. A., Kanjee, Z., Goh, E., Ling, E. B., Jain, P., Cabral, S., Abdulnour, R.-E., Haimovich, A., Freed, J. A., Olson, A., Morgan, D. J., Hom, J., Gallo, R., Horvitz, E., Chen, J., Manrai, A. K., and Rodman, A. Superhuman performance of a large language model on the reasoning tasks of a physician, 2024. URL https://arxiv.org/abs/2412.10849.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Cai, T., Song, X., Jiang, J., Teng, F., Gu, J., and Zhang, G. Ulma: Unified language model alignment with human demonstration and point-wise preference, 2024. URL https://arxiv.org/abs/2312.02554.

Campbell, D., Rane, S., Giallanza, T., De Sabbata, C. N., Ghods, K., Joshi, A., Ku, A., Frankland, S., Griffiths, T., Cohen, J. D., 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.

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

Cannon, M. D. and Edmondson, A. C. 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.

Cantini, R., Cosenza, G., Orsino, A., and Talia, D. 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.

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

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

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

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

Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., and Xia, F. 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.

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

Chen, T., Anumasa, S., Lin, B., Shah, V., Goyal, A., and Liu, D. Auto-bench: An automated benchmark for scientific discovery in llms. *arXiv preprint arXiv:2502.15224*, 2025.

Chen, X., Chi, R. A., Wang, X., and Zhou, D. Premise order matters in reasoning with large language models, 2024c. URL https://arxiv.org/abs/2402.08939.

Chen, Y., Liu, Y., Yan, J., Bai, X., Zhong, M., Yang, Y., Yang, Z., Zhu, C., and Zhang, Y. See what llms cannot answer: A self-challenge framework for uncovering llm weaknesses, 2024d. URL https://arxiv.org/abs/2408.08978.

Cheng, A.-C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., and Liu, S. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

Cherian, A., Corcodel, R., Jain, S., and Romeres, D. Llmphy: Complex physical reasoning using large language models and world models. *arXiv preprint arXiv:2411.08027*, 2024.

Chern, I.-C., Chern, S., Chen, S., Yuan, W., Feng, K., Zhou, C., He, J., Neubig, G., and Liu, P. 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.

Chia, Y. K., Sun, Q., Bing, L., and Poria, S. Can-do! a dataset and neuro-symbolic grounded framework for embodied planning with large multimodal models. *arXiv preprint arXiv:2409.14277*, 2024.

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

Chochlakis, G., Potamianos, A., Lerman, K., and Narayanan, S. The strong pull of prior knowledge in large language models and its impact on emotion recognition. *arXiv preprint arXiv:2403.17125*, 2024.

Chomsky, N., Roberts, I., and Watumull, J. Noam chomsky: The false promise of ChatGPT. *The New York Times*, March 2023.

Chow, W., Mao, J., Li, B., Seita, D., Guizilini, V., and Wang, Y. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*, 2025.

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

Chung, D. J., Gao, Z., Kvasiuk, Y., Li, T., Münchmeyer, M., Rudolph, M., Sala, F., and Tadepalli, S. C. Theoretical physics benchmark (tpbench)–a dataset and study of ai reasoning capabilities in theoretical physics. *arXiv preprint arXiv:2502.15815*, 2025.

Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., and Toutanova, K. Boolq: Exploring the surprising difficulty of natural yes/no questions, 2019. URL https://arxiv.org/abs/1905.10044.

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

Cohn, M., Pushkarna, M., Olanubi, G. O., Moran, J. M., Padgett, D., Mengesha, Z., and Heldreth, C. 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.

Copi, I. M., Cohen, C., and McMahon, K. *Introduction to logic*. Routledge, 2016.

Cross, L., Xiang, V., Bhatia, A., Yamins, D. L., and Haber, N. Hypothetical minds: Scaffolding theory of mind for multi-agent tasks with large language models. *arXiv preprint arXiv:2407.07086*, 2024.

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

Dahlgren Lindström, A., Methnani, L., Krause, L., Ericson, P., de Rituerto de Troya, Í. M., Coelho Mollo, D., and Dobbe, R. 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.

Dao, A. and Vu, D. B. Alphamaze: Enhancing large language models' spatial intelligence via grpo. *arXiv preprint arXiv:2502.14669*, 2025.

Dao, T., Fu, D. Y., Ermon, S., Rudra, A., and Ré, C. Flashattention: Fast and memory-efficient exact attention with io-awareness, 2022. URL https://arxiv.org/abs/2205.14135.

Das, B. C., Amini, M. H., and Wu, Y. 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.

"davidad" Dalrymple, D., Skalse, J., Bengio, Y., Russell, S., Tegmark, M., Seshia, S., Omohundro, S., Szegedy, C., Goldhaber, B., Ammann, N., Abate, A., Halpern, J., Barrett, C., Zhao, D., Zhi-Xuan, T., Wing, J., and Tenenbaum, J. Towards guaranteed safe ai: A framework for ensuring robust and reliable ai systems, 2024. URL https://arxiv.org/abs/2405.06624.

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

Deb, A., Oza, N., Singla, S., Khandelwal, D., Garg, D., and Singla, P. 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.

Deng, A., Cao, T., Chen, Z., and Hooi, B. Words or vision: Do vision-language models have blind faith in text? *arXiv preprint arXiv:2503.02199*, 2025a.

Deng, C., Li, Z., Xie, R., Chang, R., and Chen, H. Language models are symbolic learners in arithmetic, 2024. URL https://arxiv.org/abs/2410.15580.

Deng, X., Zhong, S., Veneris, A., Long, F., and Si, X. Verifythisbench: Generating code, specifications, and proofs all at once, 2025b. URL https://arxiv.org/abs/2505.19271.

Deshmukh, S., Han, S., Bukhari, H., Elizalde, B., Gamper, H., Singh, R., and Raj, B. Audio entailment: Assessing deductive reasoning for audio understanding, 2024. URL https://arxiv.org/abs/2407.18062.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

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

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

Dinh, T., Zhao, J., Tan, S., Negrinho, R., Lausen, L., Zha, S., and Karypis, G. Large language models of code fail at completing code with potential bugs, 2023. URL https://arxiv.org/abs/2306.03438.

Doh, S., Choi, K., Lee, J., and Nam, J. Lp-musiccaps: Llm-based pseudo music captioning, 2023. URL https://arxiv.org/abs/2307.16372.

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

Dougrez-Lewis, J., Akhter, M. E., He, Y., and Liakata, M. Assessing the reasoning abilities of chatgpt in the context of claim verification, 2024. URL https://arxiv.org/abs/2402.10735.

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

Duan, J., Pumacay, W., Kumar, N., Wang, Y. R., Tian, S., Yuan, W., Krishna, R., Fox, D., Mandlekar, A., and Guo, Y. Aha: A vision-language-model for detecting and reasoning over failures in robotic manipulation. *arXiv preprint arXiv:2410.00371*, 2024.

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

Dugan, O., Beneto, D. M. J., Loh, C., Chen, Z., Dangovski, R., and Soljačić, M. Occamllm: Fast and exact language model arithmetic in a single step, 2024. URL https://arxiv.org/abs/2406.06576.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33, 2023.

Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., West, P., Bhagavatula, C., Bras, R. L., Hwang, J. D., Sanyal, S., Welleck, S., Ren, X., Ettinger, A., Harchaoui, Z., and Choi, Y. Faith and fate: Limits of transformers on compositionality, 2023. URL https://arxiv.org/abs/2305.18654.

Echterhoff, J., Liu, Y., Alessa, A., McAuley, J., and He, Z. Cognitive bias in high-stakes decision-making with llms. *arXiv preprint arXiv:2403.00811*, 2024.

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

Fan, J., Martinson, S., Wang, E. Y., Hausknecht, K., Brenner, J., Liu, D., Peng, N., Wang, C., and Brenner, M. P. Hardmath: A benchmark dataset for challenging problems in applied mathematics, 2024. URL https://arxiv.org/abs/2410.09988.

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

Fei, H., Wu, S., Ji, W., Zhang, H., Zhang, M., Lee, M.-L., and Hsu, W. Video-of-thought: Step-by-step video reasoning from perception to cognition, 2024. URL https://arxiv.org/abs/2501.03230.

Feng, G., Yang, K., Gu, Y., Ai, X., Luo, S., Sun, J., He, D., Li, Z., and Wang, L. How numerical precision affects mathematical reasoning capabilities of llms, 2024a. URL https://arxiv.org/abs/2410.13857.

Feng, T., Han, P., Lin, G., Liu, G., and You, J. Thought-retriever: Don't just retrieve raw data, retrieve thoughts. In *ICLR 2024 Workshop: How Far Are We From AGI*, 2024b.

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

Fu, T., Ferrando, R., Conde, J., Arriaga, C., and Reviriego, P. Why do large language models (llms) struggle to count letters?, 2024. URL https://arxiv.org/abs/2412.18626.

Galatzer-Levy, I. R., McGiffin, J., Munday, D., Liu, X., Karmon, D., Labzovsky, I., Moroshko, R., Zait, A., and McDuff, D. Evidence of cognitive deficits anddevelopmental advances in generative ai: A clock drawing test analysis. *arXiv preprint arXiv:2410.11756*, 2024.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., and Ahmed, N. K. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.

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

Gandhi, K., Lynch, Z., Fränken, J.-P., Patterson, K., Wambu, S., Gerstenberg, T., Ong, D. C., and Goodman, N. D. Human-like affective cognition in foundation models. *arXiv preprint arXiv:2409.11733*, 2024.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. The pile: An 800gb dataset of diverse text for language modeling, 2020. URL https://arxiv.org/abs/2101.00027.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., and Wang, H. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2, 2023.

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

Gardner, J., Durand, S., Stoller, D., and Bittner, R. M. Llark: A multimodal instruction-following language model for music, 2024. URL https://arxiv.org/abs/2310.07160.

Gendron, G., Bao, Q., Witbrock, M., and Dobbie, G. Large language models are not strong abstract reasoners. *arXiv preprint arXiv:2305.19555*, 2023.

Ghaffari, S. and Krishnaswamy, N. Large language models are challenged by habitat-centered reasoning. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (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.

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

Ghosh, S., Kumar, S., Seth, A., Evuru, C. K. R., Tyagi, U., Sakshi, S., Nieto, O., Duraiswami, R., and Manocha, D. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities, 2024. URL https://arxiv.org/abs/2406.11768.

Ghosh, S., Kong, Z., Kumar, S., Sakshi, S., Kim, J., Ping, W., Valle, R., Manocha, D., and Catanzaro, B. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities, 2025. URL https://arxiv.org/abs/2503.03983.

Golovneva, O., Allen-Zhu, Z., Weston, J., and Sukhbaatar, S. Reverse training to nurse the reversal curse, 2024. URL https://arxiv.org/abs/2403.13799.

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

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

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

Gretsch, R., Song, P., Madhavan, A., Lau, J., and Sherwood, T. 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.

Gretsch, R., Song, P., Madhavan, A., Lau, J., and Sherwood, T. Delay space arithmetic and architecture. *IEEE Micro*, 2025.

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

Gu, Y., Tafjord, O., Kim, H., Moore, J., Bras, R. L., Clark, P., and Choi, Y. Simpletom: Exposing the gap between explicit tom inference and implicit tom application in llms. *arXiv preprint arXiv:2410.13648*, 2024.

Guan, B., Roosta, T., Passban, P., and Rezagholizadeh, M. The order effect: Investigating prompt sensitivity to input order in llms. *arXiv preprint arXiv:2502.04134*, 2025.

Gui, J., Liu, Y., Cheng, J., Gu, X., Liu, X., Wang, H., Dong, Y., Tang, J., and Huang, M. Logicgame: Benchmarking rule-based reasoning abilities of large language models, 2024. URL https://arxiv.org/abs/2408.15778.

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

Gulati, A., Miranda, B., Chen, E., Xia, E., Fronsdal, K., de Moraes Dumont, B., and Koyejo, S. Putnam-axiom: A functional and static benchmark for measuring higher level mathematical reasoning, 2024. URL https://openreview.net/forum?id=WrBqgoseGL.

Guo, P., You, W., Li, J., Bowen, Y., and Zhang, M. Exploring reversal mathematical reasoning ability for large language models. In Ku, L.-W., Martins, A., and Srikumar, V. (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/.

Guo, Q., Wang, R., Guo, J., Tan, X., Bian, J., and Yang, Y. Mitigating reversal curse in large language models via semantic-aware permutation training, 2024b. URL https://arxiv.org/abs/2403.00758.

Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., and Zhang, X. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024c.

Guo, T., Zhu, H., Zhang, R., Jiao, J., Mei, S., Jordan, M. I., and Russell, S. How do llms perform two-hop reasoning in context?, 2025. URL https://arxiv.org/abs/2502.13913.

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

Gupta, V., Pantoja, D., Ross, C., Williams, A., and Ung, M. Changing answer order can decrease mmlu accuracy, 2024. URL https://arxiv.org/abs/2406.19470.

Guran, N. B., Ren, H., Deng, J., and Xie, X. Task-oriented robotic manipulation with vision language models. *arXiv preprint arXiv:2410.15863*, 2024.

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

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

Han, P., Kocielnik, R., Saravanan, A., Jiang, R., Sharir, O., and Anandkumar, A. Chatgpt based data augmentation for improved parameter-efficient debiasing of llms. *arXiv preprint arXiv:2402.11764*, 2024a.

Han, P., Song, P., Yu, H., and You, J. In-context learning may not elicit trustworthy reasoning: A-not-B errors in pretrained language models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (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/.

Han, P., Kocielnik, R. D., Song, P., Debnath, R., Mobbs, D., Anandkumar, A., and Alvarez, R. M. Tracing human-like traits in llms: Origins, real-world manifestation, and controllability. In *2nd Workshop on Models of Human Feedback for AI Alignment*, 2025.

Han, S., Zhang, Q., Yao, Y., Jin, W., Xu, Z., and He, C. Llm multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578*, 2024c.

Hao, G., Wu, J., Pan, Q., and Morello, R. Quantifying the uncertainty of llm hallucination spreading in complex adaptive social networks. *Scientific reports*, 14(1):16375, 2024.

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

Hasani, R., Lechner, M., Amini, A., Rus, D., and Grosu, R. Liquid time-constant networks, 2020. URL https://arxiv.org/abs/2006.04439.

Havaldar, S., Rai, S., Singhal, B., Liu, L., Guntuku, S. C., and Ungar, L. Multilingual language models are not multicultural: A case study in emotion. *arXiv preprint arXiv:2307.01370*, 2023.

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

He, Y., Wu, Y., Jia, Y., Mihalcea, R., Chen, Y., and Deng, N. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. *arXiv preprint arXiv:2310.16755*, 2023.

Helwe, C., Clavel, C., and Suchanek, F. M. 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.

Hong, P., Majumder, N., Ghosal, D., Aditya, S., Mihalcea, R., and Poria, S. Evaluating llms' mathematical and coding competency through ontology-guided interventions, 2024. URL https://arxiv.org/abs/2401.09395.

Hooda, A., Christodorescu, M., Allamanis, M., Wilson, A., Fawaz, K., and Jha, S. Do large code models understand programming concepts? a black-box approach, 2024. URL https://arxiv.org/abs/2402.05980.

Hosseini, A., Sordoni, A., Toyama, D., Courville, A., and Agarwal, R. Not all llm reasoners are created equal, 2024. URL https://arxiv.org/abs/2410.01748.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Hu, H., Zhou, Y., You, L., Xu, H., Wang, Q., Lian, Z., Yu, F. R., Ma, F., and Cui, L. Emobench-m: Benchmarking emotional intelligence for multimodal large language models. *arXiv preprint arXiv:2502.04424*, 2025.

Hu, Z., Lucchetti, F., Schlesinger, C., Saxena, Y., Freeman, A., Modak, S., Guha, A., and Biswas, J. Deploying and evaluating llms to program service mobile robots. *IEEE Robotics and Automation Letters*, 9(3):2853–2860, 2024.

Huang, J. and Chang, K. C.-C. Towards reasoning in large language models: A survey, 2023. URL https://arxiv.org/abs/2212.10403.

Huang, J.-t., Zhou, J., Jin, T., Zhou, X., Chen, Z., Wang, W., Yuan, Y., Sap, M., and Lyu, M. R. On the resilience of multi-agent systems with malicious agents. *arXiv preprint arXiv:2408.00989*, 2024.

Huang, J.-t., Sun, K., Wang, W., and Dredze, M. Llms do not have human-like working memory. *arXiv preprint arXiv:2505.10571*, 2025a.

Huang, K., Guo, J., Li, Z., Ji, X., Ge, J., Li, W., Guo, Y., Cai, T., Yuan, H., Wang, R., Wu, Y., Yin, M., Tang, S., Huang, Y., Jin, C., Chen, X., Zhang, C., and Wang, M. Math-perturb: Benchmarking llms' math reasoning abilities against hard perturbations, 2025b. URL https://arxiv.org/abs/2502.06453.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., 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.

Huang, S., Song, P., George, R. J., and Anandkumar, A. Leanprogress: Guiding search for neural theorem proving via proof progress prediction. *arXiv preprint arXiv:2502.17925*, 2025d.

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

Huang, W., Xia, F., Xiao, T., Chan, H., Liang, J., Florence, P., Zeng, A., Tompson, J., Mordatch, I., Chebotar, Y., et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022b.

Huang, Y., Bie, B., Na, Z., Ruan, W., Lei, S., Yue, Y., and He, X. An empirical study of the anchoring effect in llms: Existence, mechanism, and potential mitigations. *arXiv preprint arXiv:2505.15392*, 2025e.

Itzhak, I., Belinkov, Y., and Stanovsky, G. 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.

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

Izadi, A., Banayeeanzade, M. A., Askari, F., Rahimiakbar, A., Vahedi, M. M., Hasani, H., and Baghshah, M. S. Visual structures helps visual reasoning: Addressing the binding problem in vlms. *arXiv preprint arXiv:2506.22146*, 2025.

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

Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Jain, N., Han, K., Gu, A., Li, W.-D., Yan, F., Zhang, T., Wang, S., Solar-Lezama, A., Sen, K., and Stoica, I. Livecodebench: Holistic and contamination free evaluation of large language models for code, 2024a. URL https://arxiv.org/abs/2403.07974.

Jain, S., Calacci, D., and Wilson, A. 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.

Jaiswal, R., Jain, D., Popat, H. P., Anand, A., Dharmadhikari, A., Marathe, A., and Shah, R. R. Improving physics reasoning in large language models using mixture of refinement agents. *arXiv preprint arXiv:2412.00821*, 2024.

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

Ji, J., Chen, Y., Jin, M., Xu, W., Hua, W., and Zhang, Y. Moralbench: Moral evaluation of llms. *arXiv preprint arXiv:2406.04428*, 2024.

Jiang, A. Q., Li, W., and Jamnik, M. Multilingual mathematical autoformalization, 2023a. URL https://arxiv.org/abs/2311.03755.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mixtral of experts, 2024a. URL https://arxiv.org/abs/2401.04088.

Jiang, B., Xie, Y., Hao, Z., Wang, X., Mallick, T., Su, W. J., Taylor, C. J., and Roth, D. A peek into token bias: Large language models are not yet genuine reasoners, 2024b. URL https://arxiv.org/abs/2406.11050.

Jiang, C., Wang, J., Ma, W., Clarke, C. L., Wang, S., Wu, C., and Zhang, M. Beyond utility: Evaluating llm as recommender. In *Proceedings of the ACM on Web Conference 2025*, pp. 3850–3862, 2025a.

Jiang, L., Jiang, K., Chu, X., Gulati, S., and Garg, P. 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.

Jiang, L., Hwang, J. D., Bhagavatula, C., Bras, R. L., Liang, J. T., Levine, S., Dodge, J., Sakaguchi, K., Forbes, M., Hessel, J., et al. Investigating machine moral judgement through the delphi experiment. *Nature Machine Intelligence*, pp. 1–16, 2025b.

Jiang, R., Kocielnik, R., Saravanan, A. P., Han, P., Alvarez, R. M., and Anandkumar, A. Empowering domain experts to detect social bias in generative ai with user-friendly interfaces. In *XAI in Action: Past, Present, and Future Applications*, 2023b.

Jin, Y., Li, D., Yong, A., Shi, J., Hao, P., Sun, F., Zhang, J., and Fang, B. Robotgpt: Robot manipulation learning from chatgpt. *IEEE Robotics and Automation Letters*, 9 (3):2543–2550, 2024.

Jin, Z., Tirassa, M., and Borghi, A. M. Beyond embodied cognition: Intentionality, affordance, and environmental adaptation, 2018.

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

Joshi, N., Saparov, A., Wang, Y., and He, H. Llms are prone to fallacies in causal inference, 2024. URL https://arxiv.org/abs/2406.12158.

Jovanović, N., Staab, R., and Vechev, M. Watermark stealing in large language models, 2024. URL https://arxiv.org/abs/2402.19361.

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

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

Karl, F., Kemeter, M., Dax, G., and Sierak, P. Position: Embracing negative results in machine learning. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (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.

Karpov, A., Cho, S. H., Meek, A., Koopmanschap, R., Farnik, L., and Cirstea, B.-I. Inducing human-like biases in moral reasoning language models, 2024. URL https://arxiv.org/abs/2411.15386.

Kasibatla, S. R., Agarwal, A., Brun, Y., Lerner, S., Ringer, T., and First, E. Cobblestone: Iterative automation for formal verification, 2024. URL https://arxiv.org/abs/2410.19940.

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

Khattak, M. U., Naeem, M. F., Hassan, J., Naseer, M., Tombari, F., Khan, F. S., and Khan, S. How good is my video lmm? complex video reasoning and robustness evaluation suite for video-lmms, 2024. URL https://arxiv.org/abs/2405.03690.

Kim, H., Sclar, M., Zhou, X., Bras, R. L., Kim, G., Choi, Y., and Sap, M. Fantom: A benchmark for stress-testing machine theory of mind in interactions. *arXiv preprint arXiv:2310.15421*, 2023.

Kondo, K., Sugawara, S., and Aizawa, A. 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.

Koo, R., Lee, M., Raheja, V., Park, J. I., Kim, Z. M., and Kang, D. Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012*, 2023.

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

Kumar, D., Jain, U., Agarwal, S., and Harshangi, P. Investigating implicit bias in large language models: A large-scale study of over 50 llms. *arXiv preprint arXiv:2410.12864*, 2024.

Kumarappan, A., Tiwari, M., Song, P., George, R. J., Xiao, C., and Anandkumar, A. Leanagent: Lifelong learning for formal theorem proving, 2025. URL https://arxiv.org/abs/2410.06209.

Kıcıman, E., Ness, R., Sharma, A., and Tan, C. Causal reasoning and large language models: Opening a new frontier for causality, 2024. URL https://arxiv.org/abs/2305.00050.

Lampinen, A. K., Dasgupta, I., Chan, S. C., Sheahan, H. R., Creswell, A., Kumaran, D., McClelland, J. L., and Hill, F. Language models, like humans, show content effects on reasoning tasks. *PNAS nexus*, 3(7):pgae233, 2024.

Lample, G., Lacroix, T., Lachaux, M.-A., Rodriguez, A., Hayat, A., Lavril, T., Ebner, G., and Martinet, X. Hypertree proof search for neural theorem proving. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (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.

Lawsen, A. 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.

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

Lee, N., Cai, Z., Schwarzschild, A., Lee, K., and Papailiopoulos, D. Self-improving transformers overcome easy-to-hard and length generalization challenges, 2025. URL https://arxiv.org/abs/2502.01612.

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

Lewis, C. I., Langford, C. H., and Lamprecht, P. *Symbolic logic*, volume 170. Dover publications New York, 1959.

Li, D., Rawat, A. S., Zaheer, M., Wang, X., Lukasik, M., Veit, A., Yu, F., and Kumar, S. Large language models with controllable working memory, 2022. URL https://arxiv.org/abs/2211.05110.

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

Li, H., Chong, Y. Q., Stepputtis, S., Campbell, J., Hughes, D., Lewis, M., and Sycara, K. Theory of mind for multi-agent collaboration via large language models. *arXiv preprint arXiv:2310.10701*, 2023a.

Li, M., Zhao, S., Wang, Q., Wang, K., Zhou, Y., Srivastava, S., Gokmen, C., Lee, T., Li, E. L., Zhang, R., et al. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems*, 37:100428–100534, 2025.

Li, Q., Cui, L., Zhao, X., Kong, L., and Bi, W. Gsm-plus: A comprehensive benchmark for evaluating the robustness of llms as mathematical problem solvers, 2024b. URL https://arxiv.org/abs/2402.19255.

Li, W., Cai, Y., Wu, Z., Zhang, W., Chen, Y., Qi, R., Dong, M., Chen, P., Dong, X., Shi, F., Guo, L., Han, J., Ge, B., Liu, T., Gan, L., and Zhang, T. A survey of foundation models for music understanding, 2024c. URL https://arxiv.org/abs/2409.09601.

Li, Y., Du, M., Song, R., Wang, X., and Wang, Y. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*, 2023b.

Li, Y., Michaud, E. J., Baek, D. D., Engels, J., Sun, X., and Tegmark, M. The geometry of concepts: Sparse autoencoder feature structure, 2024d. URL https://arxiv.org/abs/2410.19750.

Li, Z., Jiang, G., Xie, H., Song, L., Lian, D., and Wei, Y. Understanding and patching compositional reasoning in llms, 2024e. URL https://arxiv.org/abs/2402.14328.

Li, Z., Sun, J., Murphy, L., Su, Q., Li, Z., Zhang, X., Yang, K., and Si, X. A survey on deep learning for theorem proving, 2024f. URL https://arxiv.org/abs/2404.09939.

Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., Florence, P., and Zeng, A. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9493–9500. IEEE, 2023.

Liang, K., Hu, H., Zhao, X., Song, D., Griffiths, T. L., and Fisac, J. F. Machine bullshit: Characterizing the emergent disregard for truth in large language models, 2025. URL https://arxiv.org/abs/2507.07484.

Lichtenberg, J. M., Buchholz, A., and Schwöbel, P. Large language models as recommender systems: A study of popularity bias. *arXiv preprint arXiv:2406.01285*, 2024.

Lieder, F., Griffiths, T. L., M. Huys, Q. J., and Goodman, N. D. The anchoring bias reflects rational use of cognitive resources. *Psychonomic bulletin & review*, 25:322–349, 2018.

Lin, G., Feng, T., Han, P., Liu, G., and You, J. Paper copilot: A self-evolving and efficient llm system for personalized academic assistance. *arXiv preprint arXiv:2409.04593*, 2024a.

Lin, H., Sun, Z., Welleck, S., and Yang, Y. Lean-star: Learning to interleave thinking and proving, 2025a. URL https://arxiv.org/abs/2407.10040.

Lin, L., Wang, L., Guo, J., and Wong, K.-F. Investigating bias in llm-based bias detection: Disparities between llms and human perception. *arXiv preprint arXiv:2403.14896*, 2024b.

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

Lin, Y., Tang, S., Lyu, B., Wu, J., Lin, H., Yang, K., Li, J., Xia, M., Chen, D., Arora, S., and Jin, C. Goedel-prover: A frontier model for open-source automated theorem proving, 2025b. URL https://arxiv.org/abs/2502.07640.

Lin, Z., Fu, Z., Liu, K., Xie, L., Lin, B., Wang, W., Cai, D., Wu, Y., and Ye, J. Delving into the reversal curse: How far can large language models generalize?, 2024c. URL https://arxiv.org/abs/2410.18808.

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

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

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

Liu, H., Ning, R., Teng, Z., Liu, J., Zhou, Q., and Zhang, Y. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439*, 2023b.

Liu, H., Fu, Z., Ding, M., Ning, R., Zhang, C., Liu, X., and Zhang, Y. Logical reasoning in large language models: A survey, 2025. URL https://arxiv.org/abs/2502.09100.

Liu, R., Wei, J., Gu, S. S., Wu, T.-Y., Vosoughi, S., Cui, C., Zhou, D., and Dai, A. M. Mind's eye: Grounded language model reasoning through simulation. *arXiv preprint arXiv:2210.05359*, 2022a.

Liu, X., Yin, D., Feng, Y., and Zhao, D. Things not written in text: Exploring spatial commonsense from visual signals. *arXiv preprint arXiv:2203.08075*, 2022b.

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

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

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

Lu, Z., Jin, L., Li, P., Tian, Y., Zhang, L., Wang, S., Xu, G., Tian, C., and Cai, X. Rethinking the reversal curse of LLMs: a prescription from human knowledge reversal. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (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/.

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

Lv, A., Zhang, K., Xie, S., Tu, Q., Chen, Y., Wen, J.-R., and Yan, R. An analysis and mitigation of the reversal curse, 2024. URL https://arxiv.org/abs/2311.07468.

Lyu, Y., Yan, L., Wang, S., Shi, H., Yin, D., Ren, P., Chen, Z., de Rijke, M., and Ren, Z. Knowtuning: Knowledge-aware fine-tuning for large language models. *arXiv preprint arXiv:2402.11176*, 2024.

Ma, J., Dai, D., Sha, L., and Sui, Z. Large language models are unconscious of unreasonability in math problems, 2024a. URL https://arxiv.org/abs/2403.19346.

Ma, J.-Y., Gu, J.-C., Ling, Z.-H., Liu, Q., and Liu, C. Untying the reversal curse via bidirectional language model editing, 2024b. URL https://arxiv.org/abs/2310.10322.

Ma, X., Mishra, S., Beirami, A., Beutel, A., and Chen, J. Let's do a thought experiment: Using counterfactuals to improve moral reasoning, 2023. URL https://arxiv.org/abs/2306.14308.

Ma, Y., Gou, Z., Hao, J., Xu, R., Wang, S., Pan, L., Yang, Y., Cao, Y., Sun, A., Awadalla, H., and Chen, W. Sciagent: Tool-augmented language models for scientific reasoning, 2024c. URL https://arxiv.org/abs/2402.11451.

Malberg, S., Poletukhin, R., Schuster, C. M., and Groh, G. A comprehensive evaluation of cognitive biases in llms. *arXiv preprint arXiv:2410.15413*, 2024.

Marchetti, A., Manzi, F., Riva, G., Gaggioli, A., and Massaro, D. Artificial intelligence and the illusion of understanding: A systematic review of theory of mind and large language models. *Cyberpsychology, Behavior, and Social Networking*, 2025.

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

Mecattaf, M. G., Slater, B., Tešić, M., Prunty, J., Voudouris, K., and Cheke, L. G. A little less conversation, a little more action, please: Investigating the physical commonsense of llms in a 3d embodied environment. *arXiv preprint arXiv:2410.23242*, 2024.

Mei, L., Yao, J., Ge, Y., Wang, Y., Bi, B., Cai, Y., Liu, J., Li, M., Li, Z.-Z., Zhang, D., et al. A survey of context engineering for large language models. *arXiv preprint arXiv:2507.13334*, 2025.

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

Miceli-Barone, A. V., Barez, F., Konstas, I., and Cohen, S. B. 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.

Min, J., Buch, S., Nagrani, A., Cho, M., and Schmid, C. Morevqa: Exploring modular reasoning models for video question answering, 2024. URL https://arxiv.org/abs/2404.06511.

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., and Farajtabar, M. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models, 2024. URL https://arxiv.org/abs/2410.05229.

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

Muennighoff, N., Yang, Z., Shi, W., Li, X. L., Fei-Fei, L., Hajishirzi, H., Zettlemoyer, L., Liang, P., Candès, E., and Hashimoto, T. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

Murphy, L., Yang, K., Sun, J., Li, Z., Anandkumar, A., and Si, X. Autoformalizing euclidean geometry, 2024. URL https://arxiv.org/abs/2405.17216.

Nadeem, M., Bethke, A., and Reddy, S. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.

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

Newman, M. 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.

Nezhurina, M., Cipolina-Kun, L., Cherti, M., and Jitsev, J. 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.

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

Ni, S., Kong, X., Li, C., Hu, X., Xu, R., Zhu, J., and Yang, M. Training on the benchmark is not all you need, 2024. URL https://arxiv.org/abs/2409.01790.

Nikankin, Y., Reusch, A., Mueller, A., and Belinkov, Y. 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.

Ouyang, S., Zhang, Z., Yan, B., Liu, X., Choi, Y., Han, J., and Qin, L. Structured chemistry reasoning with large language models. *arXiv preprint arXiv:2311.09656*, 2023.

Owens, D. M., Rossi, R. A., Kim, S., Yu, T., Dernoncourt, F., Chen, X., Zhang, R., Gu, J., Deilamsalehy, H., and Lipka, N. A multi-llm debiasing framework. *arXiv preprint arXiv:2409.13884*, 2024.

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

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

Paglieri, D., Cupiał, B., Coward, S., Piterbarg, U., Wolczyk, M., Khan, A., Pignatelli, E., Kuciński, Ł., Pinto, L., Fergus, R., et al. Balrog: Benchmarking agentic llm and vlm reasoning on games. *arXiv preprint arXiv:2411.13543*, 2024.

Pan, L., Liu, A., He, Z., Gao, Z., Zhao, X., Lu, Y., Zhou, B., Liu, S., Hu, X., Wen, L., King, I., and Yu, P. S. Markllm: An open-source toolkit for llm watermarking, 2024. URL https://arxiv.org/abs/2405.10051.

Pan, M. Z., Cemri, M., Agrawal, L. A., Yang, S., Chopra, B., Tiwari, R., Keutzer, K., Parameswaran, A., Ramchandran, K., Klein, D., et al. Why do multiagent systems fail? In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025.

Pang, Q., Hu, S., Zheng, W., and Smith, V. No free lunch in llm watermarking: Trade-offs in watermarking design choices, 2024. URL https://arxiv.org/abs/2402.16187.

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

Patel, S., Wang, H., and Fan, J. Deficient executive control in transformer attention. *bioRxiv*, pp. 2025–01, 2025.

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

Pelrine, K., Imouza, A., Thibault, C., Reksoprodjo, M., Gupta, C., Christoph, J., Godbout, J.-F., and Rabbany, R. Towards reliable misinformation mitigation: Generalization, uncertainty, and gpt-4. *arXiv preprint arXiv:2305.14928*, 2023.

Pensa, G., Altuna, B., and Gonzalez-Dios, I. 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.

Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., et al. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, pp. 13387–13434, 2023.

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

Pi, Z., Vadaparty, A., Bergen, B. K., and Jones, C. R. 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.

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

Piatti, G., Jin, Z., Kleiman-Weiner, M., Schölkopf, B., Sachan, M., and Mihalcea, R. 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.

Plaat, A., Wong, A., Verberne, S., Broekens, J., van Stein, N., and Back, T. Reasoning with large language models, a survey. *arXiv preprint arXiv:2407.11511*, 2024.

Pock, M., Ye, A., and Moore, J. Llms grasp morality in concept, 2023. URL https://arxiv.org/abs/2311.02294.

Poesia, G. and Goodman, N. D. 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.

Poesia, G., Broman, D., Haber, N., and Goodman, N. D. Learning formal mathematics from intrinsic motivation. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (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.

Qi, C., Li, B., Hui, B., Wang, B., Li, J., Wu, J., and Laili, Y. An investigation of llms' inefficacy in understanding converse relations, 2023. URL https://arxiv.org/abs/2310.05163.

Qi, J., Liu, J., Tang, H., and Zhu, Z. Beyond semantics: Rediscovering spatial awareness in vision-language models. *arXiv preprint arXiv:2503.17349*, 2025.

Qian, K., Wan, S., Tang, C., Wang, Y., Zhang, X., Chen, M., and Yu, Z. Varbench: Robust language model benchmarking through dynamic variable perturbation, 2024. URL https://arxiv.org/abs/2406.17681.

Qiao, S., Ou, Y., Zhang, N., Chen, X., Yao, Y., Deng, S., Tan, C., Huang, F., and Chen, H. Reasoning with language model prompting: A survey, 2023. URL https://arxiv.org/abs/2212.09597.

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

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. 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.

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

Rahmanzadehgervi, P., Bolton, L., Taesiri, M. R., and Nguyen, A. T. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision*, pp. 18–34, 2024.

Raj, C., Banerjee, M., Caliskan, A., Anastasopoulos, A., and Zhu, Z. Talent or luck? evaluating attribution bias in large language models. *arXiv preprint arXiv:2505.22910*, 2025.

Rajore, T., Chandran, N., Sitaram, S., Gupta, D., Sharma, R., Mittal, K., and Swaminathan, M. Truce: Private benchmarking to prevent contamination and improve comparative evaluation of llms, 2024. URL https://arxiv.org/abs/2403.00393.

Rastogi, C., Zhang, Y., Wei, D., Varshney, K. R., Dhurandhar, A., and Tomsett, R. 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.

Ren, W., Ma, W., Yang, H., Wei, C., Zhang, G., and Chen, W. Vamba: Understanding hour-long videos with hybrid mamba-transformers, 2025. URL https://arxiv.org/abs/2503.11579.

Rezaei, M., Fu, Y., Cuvin, P., Ziems, C., Zhang, Y., Zhu, H., and Yang, D. Egonormia: Benchmarking physical social norm understanding. *arXiv preprint arXiv:2502.20490*, 2025.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*, 2020.

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

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

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

Sabour, S., Liu, S., Zhang, Z., Liu, J. M., Zhou, J., Sunaryo, A. S., Li, J., Lee, T., Mihalcea, R., and Huang, M. Emobench: Evaluating the emotional intelligence of large language models. *arXiv preprint arXiv:2402.12071*, 2024.

Saito, K., Wachi, A., Wataoka, K., and Akimoto, Y. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*, 2023.

Sakshi, S., Tyagi, U., Kumar, S., Seth, A., Selvakumar, R., Nieto, O., Duraiswami, R., Ghosh, S., and Manocha, D. Mmau: A massive multi-task audio understanding and reasoning benchmark, 2024. URL https://arxiv.org/abs/2410.19168.

Sampson, G. 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.

Sap, M., LeBras, R., Fried, D., and Choi, Y. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*, 2022.

Saravanan, A. P., Kocielnik, R., Jiang, R., Han, P., and Anandkumar, A. Exploring social bias in downstream applications of text-to-image foundation models. *arXiv preprint arXiv:2312.10065*, 2023.

Sarch, G., Saha, S., Khandelwal, N., Jain, A., Tarr, M. J., Kumar, A., and Fragkiadaki, K. Grounded reinforcement learning for visual reasoning. *arXiv preprint arXiv:2505.23678*, 2025.

Sarker, L., Downing, M., Desai, A., and Bultan, T. Syntactic robustness for llm-based code generation, 2024. URL https://arxiv.org/abs/2404.01535.

Saxena, R., Gema, A. P., and Minervini, P. Lost in time: Clock and calendar understanding challenges in multimodal llms, 2025. URL https://arxiv.org/abs/2502.05092.

Schmidgall, S., Achterberg, J., Miconi, T., Kirsch, L., Ziaei, R., Hajiseyedrazi, S. P., and Eshraghian, J. Brain-inspired learning in artificial neural networks: a review, 2023. URL https://arxiv.org/abs/2305.11252.

Schmidgall, S., Harris, C., Essien, I., Olshvang, D., Rahman, T., Kim, J. W., Ziaei, R., Eshraghian, J., Abadir, P., and

Chellappa, R. Evaluation and mitigation of cognitive biases in medical language models. *npj Digital Medicine*, 7(1):295, 2024.

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

Sclar, M., Kumar, S., West, P., Suhr, A., Choi, Y., and Tsvetkov, Y. Minding language models'(lack of) theory of mind: A plug-and-play multi-character belief tracker. *arXiv preprint arXiv:2306.00924*, 2023.

Senthilkumar, P., Balasubramanian, V., Jain, P., Maity, A., Lu, J., and Zhu, K. Fine-tuning language models for ethical ambiguity: A comparative study of alignment with human responses, 2024. URL https://arxiv.org/abs/2410.07826.

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

Shafiei, M., Saffari, H., and Moosavi, N. S. More or less wrong: A benchmark for directional bias in llm comparative reasoning. *arXiv preprint arXiv:2506.03923*, 2025.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., Sap, M., and Shwartz, V. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*, 2023.

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

Shapiro, L. and Spaulding, S. Embodied Cognition. In Zalta, E. N. and Nodelman, U. (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2024 edition, 2024.

Shen, H., Wu, T., Han, Q., Hsieh, Y., Wang, J., Zhang, Y., Cheng, Y., Hao, Z., Ni, Y., Wang, X., Wan, Z., Zhang, K., Xu, W., Xiong, J., Luo, P., Chen, W., Tao, C., Mao, Z., and Wong, N. Phyx: Does your model have the "wits" for physical reasoning?, 2025. URL https://arxiv.org/abs/2505.15929.

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

Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E., Schärli, N., and Zhou, D. Large language models can be easily distracted by irrelevant context, 2023. URL https://arxiv.org/abs/2302.00093.

Shi, L., Liu, H., Wong, Y., Mujumdar, U., Zhang, D., Gwizdka, J., and Lease, M. Argumentative experience: Reducing confirmation bias on controversial issues through llm-generated multi-persona debates. *arXiv preprint arXiv:2412.04629*, 2024.

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

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

Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., and Farajtabar, M. 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.

Shu, C., Han, J., Liu, F., Shareghi, E., and Collier, N. Posqa: Probe the world models of llms with size comparisons. *arXiv preprint arXiv:2310.13394*, 2023.

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

Song, P., Yang, K., and Anandkumar, A. Lean copilot: Large language models as copilots for theorem proving in lean, 2025. URL https://arxiv.org/abs/2404.12534.

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

Stich, S. P. 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.

Strachan, J. W., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., et al. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, pp. 1–11, 2024.

Su, Z., Li, J., Zhang, J., Zhu, T., Qu, X., Zhou, P., Bowen, Y., Cheng, Y., and zhang, M. Living in the moment: Can large language models grasp co-temporal reasoning?, 2024. URL https://arxiv.org/abs/2406.09072.

Sumita, Y., Takeuchi, K., and Kashima, H. 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.

Sun, J., Zheng, C., Xie, E., Liu, Z., Chu, R., Qiu, J., Xu, J., Ding, M., Li, H., Geng, M., et al. A survey of reasoning with foundation models. *arXiv preprint arXiv:2312.11562*, 2023.

Sun, X., Tan, H., Guo, Y., Qiang, P., Li, R., and Zhang, H. 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.

Sun, Y., Hu, S., Zhou, G., Zheng, K., Hajishirzi, H., Dziri, N., and Song, D. Omega: Can llms reason outside the box in math? evaluating exploratory, compositional, and transformative generalization, 2025b. URL https://arxiv.org/abs/2506.18880.

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

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

Tang, K., Song, P., Qin, Y., and Yan, X. Creative and context-aware translation of east asian idioms with gpt-4. *arXiv preprint arXiv:2410.00988*, 2024.

Tanmay, K., Khandelwal, A., Agarwal, U., and Choudhury, M. Probing the moral development of large language models through defining issues test. *arXiv preprint arXiv:2309.13356*, 2023.

Testolin, A. 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.

Thakur, A., Tsoukalas, G., Wen, Y., Xin, J., and Chaudhuri, S. An in-context learning agent for formal theorem-proving, 2024. URL https://arxiv.org/abs/2310.04353.

Thompson, K., Saavedra, N., Carrott, P., Fisher, K., Sanchez-Stern, A., Brun, Y., Ferreira, J. F., Lerner, S., and First, E. Rango: Adaptive retrieval-augmented proving for automated software verification, 2025. URL https://arxiv.org/abs/2412.14063.

Tian, S.-Y., Zhou, Z., Jia, L.-H., Guo, L.-Z., and Li, Y.-F. Robustness assessment of mathematical reasoning in the presence of missing and contradictory conditions, 2024. URL https://arxiv.org/abs/2406.05055.

Tie, J., Yao, B., Li, T., Ahmed, S. I., Wang, D., and Zhou, S. Llms are imperfect, then what? an empirical study on llm failures in software engineering, 2024. URL https://arxiv.org/abs/2411.09916.

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

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

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

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

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

Upadhayay, B., Behzadan, V., and Karbasi, A. Working memory attack on llms. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025.

van Duijn, M. J., van Dijk, B., Kouwenhoven, T., de Valk, W., Spruit, M. R., and van der Putten, P. 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.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

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

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

Wan, Y., Wang, W., Yang, Y., Yuan, Y., Huang, J.-t., He, P., Jiao, W., and Lyu, M. LogicAsker: Evaluating and improving the logical reasoning ability of large language models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (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.

Wan, Y., Jia, X., and Li, X. L. Unveiling confirmation bias in chain-of-thought reasoning. *arXiv preprint arXiv:2506.12301*, 2025a.

Wan, Y., Jia, X., and Li, X. L. Unveiling confirmation bias in chain-of-thought reasoning, 2025b. URL https://arxiv.org/abs/2506.12301.

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

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., and Anandkumar, A. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023a.

Wang, G., Wang, W., Cao, Y., Teng, Y., Guo, Q., Wang, H., Lin, J., Ma, J., Liu, J., and Wang, Y. Possibilities and challenges in the moral growth of large language models: a philosophical perspective. *Ethics and Information Technology*, 27(1):9, 2025a.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Katwyk, P. V., Deac, A., Anandkumar, A., Bergen, K. J., Gomes, C. P., Ho, S., Kohli, P., Lasenby, J., Leskovec, J., Liu, T.-Y., Manrai, A. K., Marks, D. S., Ramsundar, B., Song, L., Sun, J., Tang, J., Velickovic, P., Welling, M., Zhang, L., Coley, C. W., Bengio, Y., and Zitnik, M. Scientific discovery in the age of artificial intelligence. *Nature*, 620:47–60, 2023b. URL https://api.semanticscholar.org/CorpusID:260384616.

Wang, Q., Lou, Z., Tang, Z., Chen, N., Zhao, X., Zhang, W., Song, D., and He, B. Assessing judging bias in large reasoning models: An empirical study. *arXiv preprint arXiv:2504.09946*, 2025b.

Wang, S., Li, Z., Qian, H., Yang, C., Wang, Z., Shang, M., Kumar, V., Tan, S., Ray, B., Bhatia, P., Nallapati, R., Ramanathan, M. K., Roth, D., and Xiang, B. Recode: Robustness evaluation of code generation models, 2022. URL https://arxiv.org/abs/2212.10264.

Wang, S., Wei, Z., Choi, Y., and Ren, X. Can llms reason with rules? logic scaffolding for stress-testing and improving llms, 2024. URL https://arxiv.org/abs/2402.11442.

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

Wang, Y. R., Duan, J., Fox, D., and Srinivasa, S. Newton: Are large language models capable of physical reasoning?, 2023c. URL https://arxiv.org/abs/2310.07018.

Wei, A., Haghtalab, N., and Steinhardt, J. Jailbroken: How does llm safety training fail? In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (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.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. Emergent abilities of large language models, 2022a. URL https://arxiv.org/abs/2206.07682.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022b.

Wei, T., Luan, J., Liu, W., Dong, S., and Wang, B. Cmath: Can your language model pass chinese elementary school math test?, 2023b. URL https://arxiv.org/abs/2306.16636.

Wei, X., Kumar, N., and Zhang, H. Addressing bias in generative ai: Challenges and research opportunities in information management. *arXiv preprint arXiv:2502.10407*, 2025.

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

Wen, B., Xu, C., Wolfe, R., Wang, L. L., Howe, B., 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.

White, C., Dooley, S., Roberts, M., Pal, A., Feuer, B., Jain, S., Shwartz-Ziv, R., Jain, N., Saifullah, K., Naidu, S., Hegde, C., LeCun, Y., Goldstein, T., Neiswanger, W., and Goldblum, M. Livebench: A challenging, contamination-free llm benchmark, 2024. URL https://arxiv.org/abs/2406.19314.

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

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

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

Woźniak, S., Pantazi, A., Bohnstingl, T., and Eleftheriou, E. 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.

Wu, D., Yang, J., and Wang, K. 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.

Wu, F., Zhang, N., Jha, S., McDaniel, P., and Xiao, C. A new era in llm security: Exploring security concerns in real-world llm-based systems, 2024b. URL https://arxiv.org/abs/2402.18649.

Wu, K., Wu, E., and Zou, J. Y. 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.

Wu, S., Oltramari, A., Francis, J., Giles, C. L., and Ritter, F. E. Cognitive llms: Toward human-like artificial intelligence by integrating cognitive architectures and large language models for manufacturing decision-making. *Neurosymbolic Artificial Intelligence*, 2024d.

Wu, W., Mao, S., Zhang, Y., Xia, Y., Dong, L., Cui, L., and Wei, F. 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.

Wu, X., Wang, Y., Jegelka, S., and Jadbabaie, A. On the emergence of position bias in transformers. *arXiv preprint arXiv:2502.01951*, 2025b.

Wu, Y., Jiang, A. Q., Li, W., Rabe, M., Staats, C., Jamnik, M., and Szegedy, C. Autoformalization with large language models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (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.

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., Kim, N., Andreas, J., and Kim, Y. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In Duh, K., Gomez, H., and Bethard, S. (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.

Xia, C. S., Deng, Y., and Zhang, L. Top leaderboard ranking = top coding proficiency, always? evoeval: Evolving coding benchmarks via llm, 2024. URL https://arxiv.org/abs/2403.19114.

Xiao, Y., Wang, J., Xu, Q., Song, C., Xu, C., Cheng, Y., Li, W., and Liu, P. Towards dynamic theory of mind: Evaluating llm adaptation to temporal evolution of human states. *arXiv preprint arXiv:2505.17663*, 2025.

Xie, Z., Lin, M., Liu, Z., Wu, P., Yan, S., and Miao, C. Audio-reasoner: Improving reasoning capability in large audio language models, 2025. URL https://arxiv.org/abs/2503.02318.

Xin, H., Ren, Z. Z., Song, J., Shao, Z., Zhao, W., Wang, H., Liu, B., Zhang, L., Lu, X., Du, Q., Gao, W., Zhu, Q., Yang, D., Gou, Z., Wu, Z. F., Luo, F., and Ruan, C. Deepseek-prover-v1.5: Harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search, 2024. URL https://arxiv.org/abs/2408.08152.

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., and Hooi, B. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

Xu, B. and Poo, M.-m. 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.

Xu, M., Huang, P., Yu, W., Liu, S., Zhang, X., Niu, Y., Zhang, T., Xia, F., Tan, J., and Zhao, D. Creative robot tool use with large language models. *arXiv preprint arXiv:2310.13065*, 2023a.

Xu, N. and Ma, X. 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.

Xu, P., Ping, W., Wu, X., McAfee, L., Zhu, C., Liu, Z., Subramanian, S., Bakhturina, E., Shoeybi, M., and Catanzaro, B. Retrieval meets long context large language models. In *The Twelfth International Conference on Learning Representations*, 2023b.

Xu, R., Wang, Z., Fan, R.-Z., and Liu, P. Benchmarking benchmark leakage in large language models, 2024a. URL https://arxiv.org/abs/2404.18824.

Xu, X., Xu, Q., Xiao, T., Chen, T., Yan, Y., Zhang, J., Diao, S., Yang, C., and Wang, Y. Ugphysics: A comprehensive benchmark for undergraduate physics reasoning with large language models. *arXiv preprint arXiv:2502.00334*, 2025.

Xu, Y., Li, W., Vaezipoor, P., Sanner, S., and Khalil, E. B. Llms and the abstraction and reasoning corpus: Successes, failures, and the importance of object-based representations. *arXiv preprint arXiv:2305.18354*, 2023c.

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

Yamin, K., Gupta, S., Ghosal, G. R., Lipton, Z. C., and Wilder, B. Failure modes of llms for causal reasoning on narratives, 2024. URL https://arxiv.org/abs/2410.23884.

Yan, C., Wang, H., Yan, S., Jiang, X., Hu, Y., Kang, G., Xie, W., and Gavves, E. Visa: Reasoning video object segmentation via large language models, 2024. URL https://arxiv.org/abs/2407.11325.

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

Yang, K., Poesia, G., He, J., Li, W., Lauter, K., Chaudhuri, S., and Song, D. Formal mathematical reasoning: A new frontier in ai, 2024a. URL https://arxiv.org/abs/2412.16075.

Yang, L., Yu, Z., Zhang, T., Cao, S., Xu, M., Zhang, W., Gonzalez, J. E., and Cui, B. Buffer of thoughts: Thought-augmented reasoning with large language models. *Advances in Neural Information Processing Systems*, 37: 113519–113544, 2024b.

Yang, N., Kang, T., Choi, J., Lee, H., and Jung, K. Mitigating biases for instruction-following language models via bias neurons elimination. *arXiv preprint arXiv:2311.09627*, 2023b.

Yang, S., Kassner, N., Gribovskaya, E., Riedel, S., and Geva, M. Do large language models perform latent multi-hop reasoning without exploiting shortcuts?, 2024c. URL https://arxiv.org/abs/2411.16679.

Yang, Z., Ding, M., Lv, Q., Jiang, Z., He, Z., Guo, Y., Bai, J., and Tang, J. Gpt can solve mathematical problems without a calculator, 2023c. URL https://arxiv.org/abs/2309.03241.

Yao, J.-Y., Ning, K.-P., Liu, Z.-H., Ning, M.-N., Liu, Y.-Y., and Yuan, L. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*, 2023.

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., and Zhang, Y. 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.

Ye, Z., Yan, Z., He, J., Kasriel, T., Yang, K., and Song, D. Verina: Benchmarking verifiable code generation. *arXiv preprint arXiv:2505.23135*, 2025.

Yehudai, G., Kaplan, H., Ghandeharioun, A., Geva, M., and Globerson, A. When can transformers count to n?, 2024. URL https://arxiv.org/abs/2407.15160.

Youssef, P., Schlötterer, J., and Seifert, C. 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.

Yu, D., Song, K., Lu, P., He, T., Tan, X., Ye, W., Zhang, S., and Bian, J. Musicagent: An ai agent for music understanding and generation with large language models, 2023a. URL https://arxiv.org/abs/2310.11954.

Yu, F., Zhang, H., Tiwari, P., and Wang, B. Natural language reasoning, a survey, 2023b. URL https://arxiv.org/abs/2303.14725.

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

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

Yu, P., Xu, J., Weston, J., and Kulikov, I. Distilling system 2 into system 1, 2024b. URL https://arxiv.org/abs/2407.06023.

Yu, S., Song, J., Hwang, B., Kang, H., Cho, S., Choi, J., Joe, S., Lee, T., Gwon, Y. L., and Yoon, S. Correcting negative bias in large language models through negative attention score alignment. *arXiv preprint arXiv:2408.00137*, 2024c.

Yu, X., Cheng, H., Liu, X., Roth, D., and Gao, J. Reeval: Automatic hallucination evaluation for retrieval-augmented large language models via transferable adversarial attacks. *arXiv preprint arXiv:2310.12516*, 2023d.

Yuan, R., Lin, H., Guo, S., Zhang, G., Pan, J., Zang, Y., Liu, H., Liang, Y., Ma, W., Du, X., Du, X., Ye, Z., Zheng, T., Ma, Y., Liu, M., Tian, Z., Zhou, Z., Xue, L., Qu, X., Li, Y., Wu, S., Shen, T., Ma, Z., Zhan, J., Wang, C., Wang, Y., Chi, X., Zhang, X., Yang, Z., Wang, X., Liu, S., Mei, L., Li, P., Wang, J., Yu, J., Pang, G., Li, X., Wang, Z., Zhou, X., Yu, L., Benetos, E., Chen, Y., Lin, C., Chen, X., Xia, G., Zhang, Z., Zhang, C., Chen, W., Zhou, X., Qiu, X., Dannenberg, R., Liu, J., Yang, J., Huang, W., Xue, W., Tan, X., and Guo, Y. Yue: Scaling open foundation models for long-form music generation, 2025. URL https://arxiv.org/abs/2503.08638.

Yuan, Z., Yuan, H., Tan, C., Wang, W., and Huang, S. How well do large language models perform in arithmetic tasks?, 2023. URL https://arxiv.org/abs/2304.02015.

Zhang, C., Jian, Y., Ouyang, Z., and Vosoughi, S. 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.

Zhang, D., Hu, Z., Zhoubian, S., Du, Z., Yang, K., Wang, Z., Yue, Y., Dong, Y., and Tang, J. Sciglm: Training scientific language models with self-reflective instruction annotation and tuning, 2024b. URL https://arxiv.org/abs/2401.07950.

Zhang, H., Li, L. H., Meng, T., Chang, K.-W., and den Broeck, G. V. On the paradox of learning to reason from data, 2022. URL https://arxiv.org/abs/2205.11502.

Zhang, H., Du, W., Shan, J., Zhou, Q., Du, Y., Tenenbaum, J. B., Shu, T., and Gan, C. Building cooperative embodied agents modularly with large language models. *arXiv preprint arXiv:2307.02485*, 2023.

Zhang, H., Zhu, C., Wang, X., Zhou, Z., Yin, C., Li, M., Xue, L., Wang, Y., Hu, S., Liu, A., et al. Badrobot: Manipulating embodied llms in the physical world. *arXiv preprint arXiv:2407.20242*, 2024c.

Zhang, R., Hussain, S. S., Neekhara, P., and Koushanfar, F. 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 2024d. USENIX Association. ISBN 978-1-939133-44-1. URL https://www.usenix.org/conference/usenixsecurity24/presentation/zhang-ruisi.

Zhang, X., Cao, J., and You, C. Counting ability of large language models and impact of tokenization, 2024e. URL https://arxiv.org/abs/2410.19730.

Zhang, X., Dong, Y., Wu, Y., Huang, J., Jia, C., Fernando, B., Shou, M. Z., Zhang, L., and Liu, J. Physreason: A comprehensive benchmark towards physics-based reasoning. *arXiv preprint arXiv:2502.12054*, 2025a.

Zhang, Y. and He, Z. Large language models can not perform well in understanding and manipulating natural language at both character and word levels? In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (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/.

Zhang, Y., Xue, M., Liu, D., and He, Z. Rationales for answers to simple math word problems confuse large language models. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8853–8869, Bangkok, Thailand, August 2024f. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.524. URL https://aclanthology.org/2024.findings-acl.524/.

Zhang, Y., Ma, Y., Gu, Y., Yang, Z., Zhuang, Y., Wang, F., Huang, Z., Wang, Y., Huang, C., Song, B., et al. Abench-physics: Benchmarking physical reasoning in llms via high-difficulty and dynamic physics problems. *arXiv preprint arXiv:2507.04766*, 2025b.

Zhang, Z., Wang, Y., Wang, C., Chen, J., and Zheng, Z. Llm hallucinations in practical code generation: Phenomena, mechanism, and mitigation. *arXiv preprint arXiv:2409.20550*, 2024g.

Zhang-Li, D., Lin, N., Yu, J., Zhang, Z., Yao, Z., Zhang, X., Hou, L., Zhang, J., and Li, J. Reverse that number! decoding order matters in arithmetic learning, 2024. URL https://arxiv.org/abs/2403.05845.

Zhao, B., Dirac, L. P., and Varshavskaya, P. Can vision language models learn from visual demonstrations of ambiguous spatial reasoning? *arXiv preprint arXiv:2409.17080*, 2024a.

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

Zhao, J., Tong, J., Mou, Y., Zhang, M., Zhang, Q., and Huang, X. Exploring the compositional deficiency of large language models in mathematical reasoning, 2024b. URL https://arxiv.org/abs/2405.06680.

Zhao, J., Zhang, Z., Chen, B., Wang, Z., Anandkumar, A., and Tian, Y. Galore: Memory-efficient llm training by gradient low-rank projection, 2024c. URL https://arxiv.org/abs/2403.03507.

Zhao, R., Zhu, Q., Xu, H., Li, J., Zhou, Y., He, Y., and Gui, L. Large language models fall short: Understanding complex relationships in detective narratives, 2024d. URL https://arxiv.org/abs/2402.11051.

Zhao, X., Ananth, P., Li, L., and Wang, Y.-X. Provable robust watermarking for ai-generated text, 2023. URL https://arxiv.org/abs/2306.17439.

Zheng, C., Zhou, H., Meng, F., Zhou, J., and Huang, M. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, 2023.

Zheng, W., Yang, A., Lin, N., and Zhou, D. 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.

Zheng, X., Pang, T., Du, C., Liu, Q., Jiang, J., and Lin, M. Cheating automatic llm benchmarks: Null models achieve high win rates, 2024b. URL https://arxiv.org/abs/2410.07137.

Zhou, H., Wan, X., Proleev, L., Mincu, D., Chen, J., Heller, K., and Roy, S. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. *arXiv preprint arXiv:2309.17249*, 2023a.

Zhou, J., Ghaddar, A., Zhang, G., Ma, L., Hu, Y., Pal, S., Coates, M., Wang, B., Zhang, Y., and Hao, J. Enhancing logical reasoning in large language models through graph-based synthetic data, 2024a. URL https://arxiv.org/abs/2409.12437.

Zhou, K., Lai, E., Yeong, W. B. A., Mouratidis, K., and Jiang, J. Rome: Evaluating pre-trained vision-language models on reasoning beyond visual common sense. *arXiv preprint arXiv:2310.19301*, 2023b.

Zhou, K., Zhu, Y., Chen, Z., Chen, W., Zhao, W. X., Chen, X., Lin, Y., Wen, J.-R., and Han, J. Don't make your llm an evaluation benchmark cheater. *ArXiv*, abs/2311.01964, 2023c. URL https://api.semanticscholar.org/CorpusID:265019021.

Zhou, P., Madaan, A., Potharaju, S. P., Gupta, A., McKee, K. R., Holtzman, A., Pujara, J., Ren, X., Mishra, S., Nematzadeh, A., et al. How far are large language models from agents with theory-of-mind? *arXiv preprint arXiv:2310.03051*, 2023d.

Zhou, Z., Wu, Y., Wu, Z., Zhang, X., Yuan, R., Ma, Y., Wang, L., Benetos, E., Xue, W., and Guo, Y. Can llms "reason" in music? an evaluation of llms' capability of music understanding and generation, 2024b. URL https://arxiv.org/abs/2407.21531.

Zhu, E., Liu, Y., Zhang, Z., Li, X., Zhou, J., Yu, X., Huang, M., and Wang, H. Maps: Advancing multi-modal reasoning in expert-level physical science. *arXiv preprint arXiv:2501.10768*, 2025.

Zhu, H., Huang, B., Zhang, S., Jordan, M., Jiao, J., Tian, Y., and Russell, S. Towards a theoretical understanding of the 'reversal curse' via training dynamics, 2024a. URL https://arxiv.org/abs/2405.04669.

Zhu, W., Liu, H., Dong, Q., Xu, J., Huang, S., Kong, L., Chen, J., and Li, L. Multilingual machine translation with large language models: Empirical results and analysis, 2024b. URL https://arxiv.org/abs/2304.04675.

Zhu, X., Zhang, C., Stafford, T., Collier, N., and Vlachos, A. Conformity in large language models. *arXiv preprint arXiv:2410.12428*, 2024c.

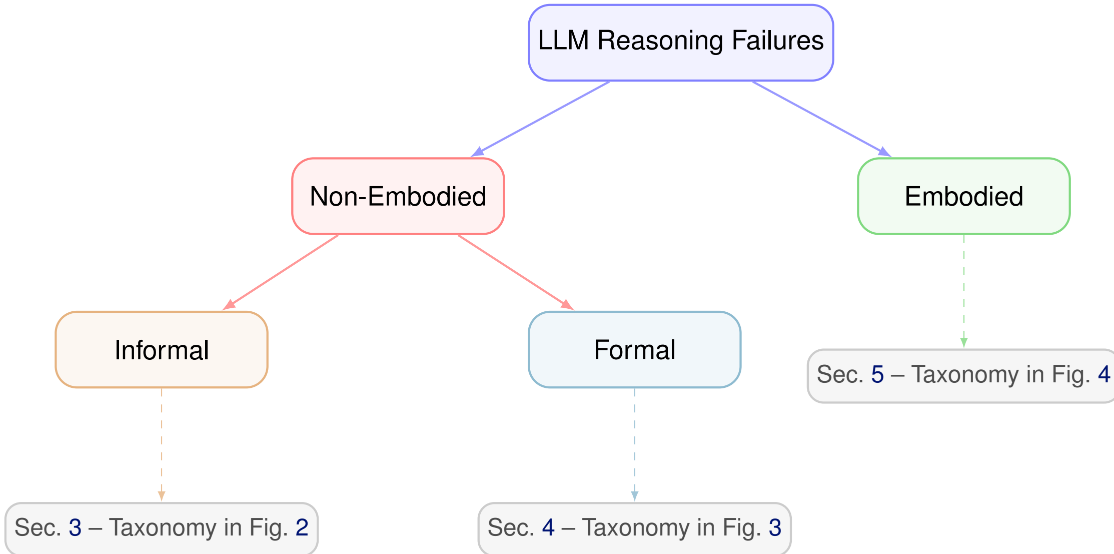


Figure 1. Overall taxonomy of LLM Reasoning Failures.

A. Conclusion

In this survey, we systematically explored reasoning failures in Large Language Models across informal, formal, and embodied dimensions. By establishing clear definitions and categorizations, we unified previously fragmented observations into coherent patterns of systematic weaknesses, paired with sophisticated discussions about root causes and mitigation strategies for each. Our analysis underscores that despite remarkable progress, fundamental reasoning challenges persist, limiting the reliability and practical deployment of LLMs. Future research should prioritize addressing these pervasive reasoning gaps through deeper cognitive alignment, improved logical robustness, and enhanced grounding in embodied interactions. We hope this structured survey inspires more focused efforts, advancing the understanding and capabilities of LLM reasoning toward more robust, trustworthy, and effective real-world applications.

B. Taxonomy

In this section, we present a visualized taxonomy for the field of 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 overall taxonomy of LLM reasoning failures is presented in Figure 1, where we comprehensively break down all LLM reasoning failures 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 final categories, and here provide specific taxonomies for each category – informal (Section 3; taxonomy in Figure 2), and formal (Section 4; taxonomy in Figure 3), and embodied (Section 5; taxonomy in Figure 4).

C. Artifacts

Upon the full 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 entry point. The collection will be released as a public Github repository, which will also be continuously updated in the future as the field progresses.

D. Other Emerging Areas of Reasoning

Recent advancements 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.

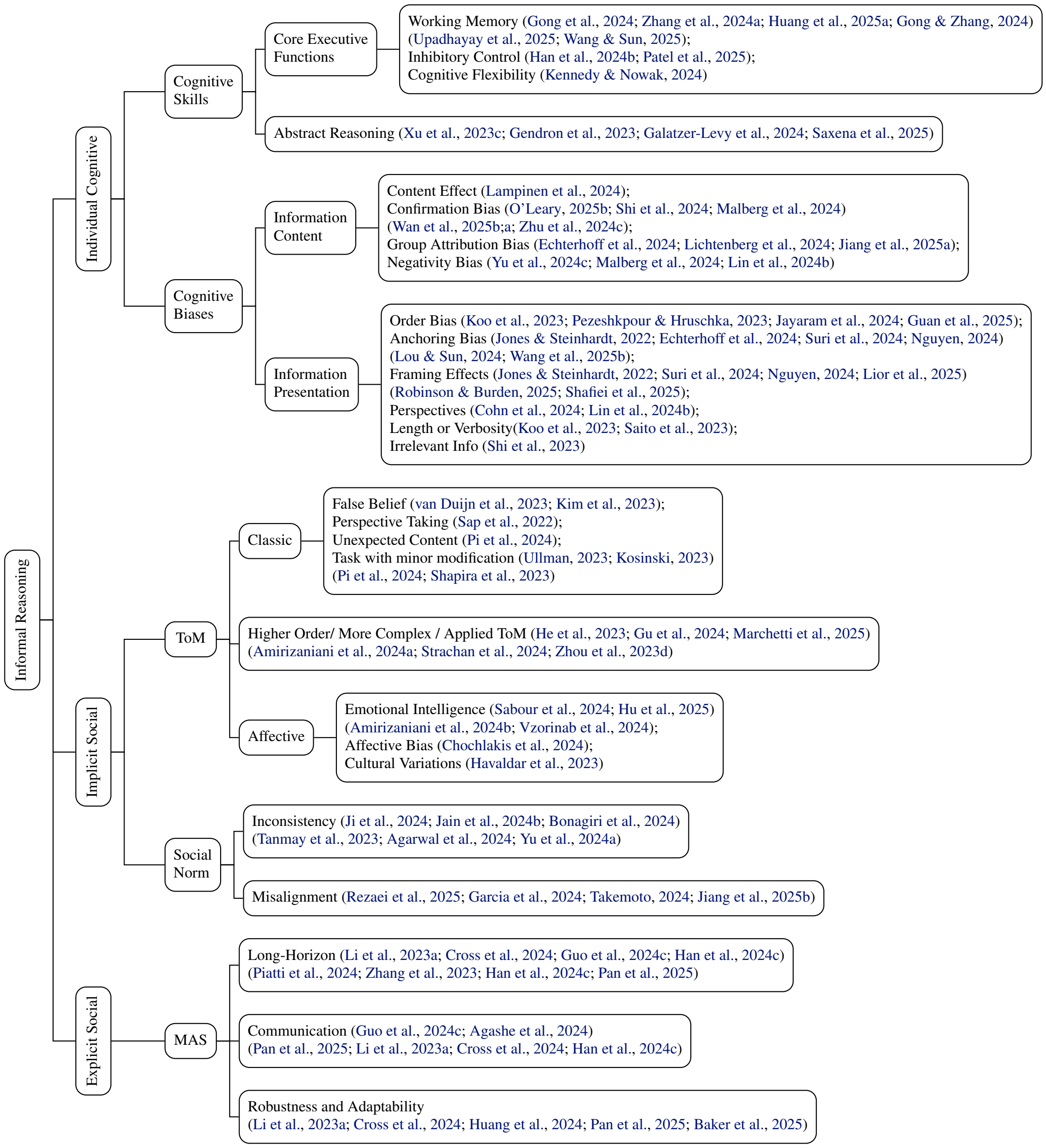


Figure 2. Taxonomy of Informal LLM Reasoning Failures.

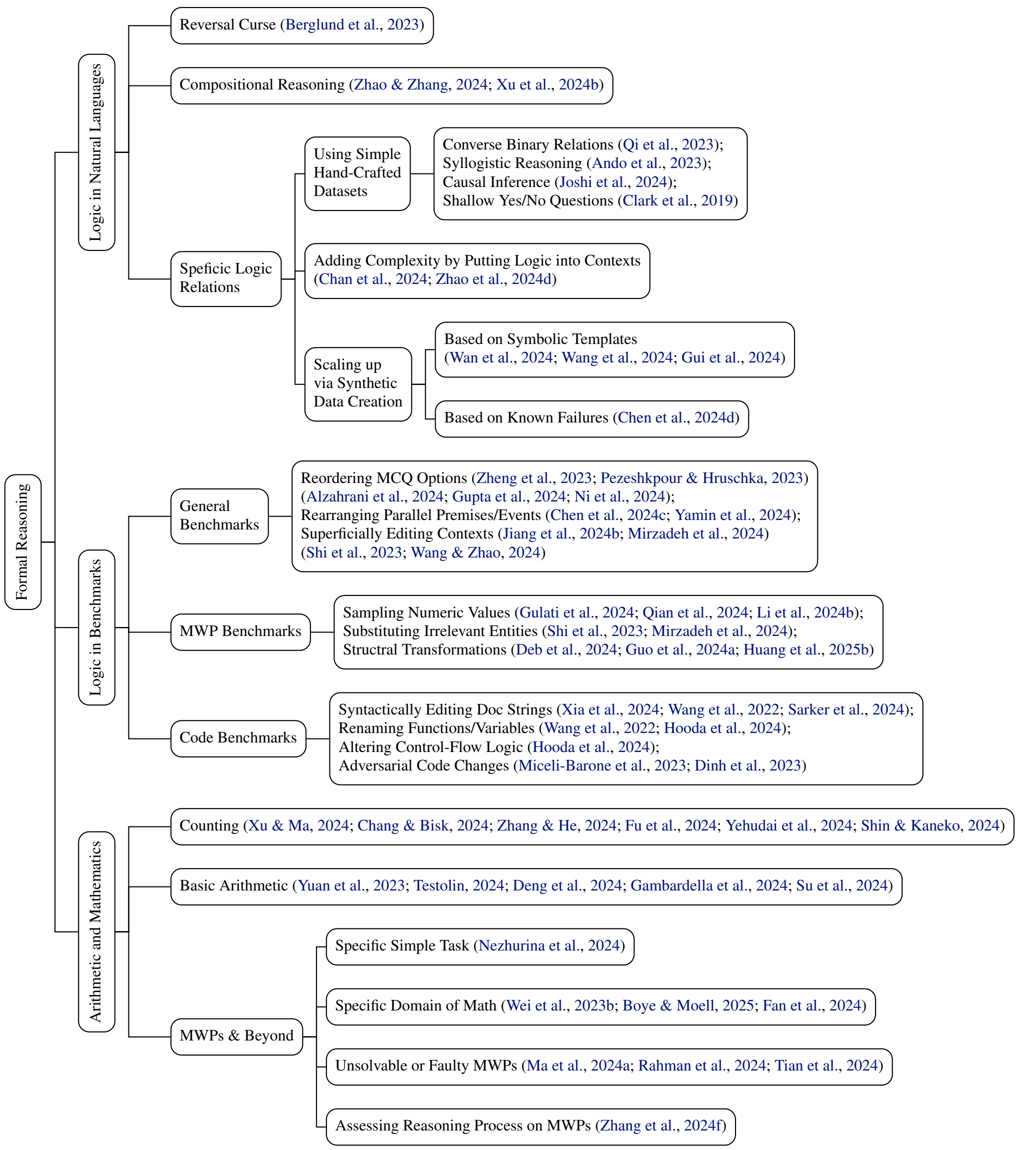


Figure 3. Taxonomy of Formal LLM Reasoning Failures.

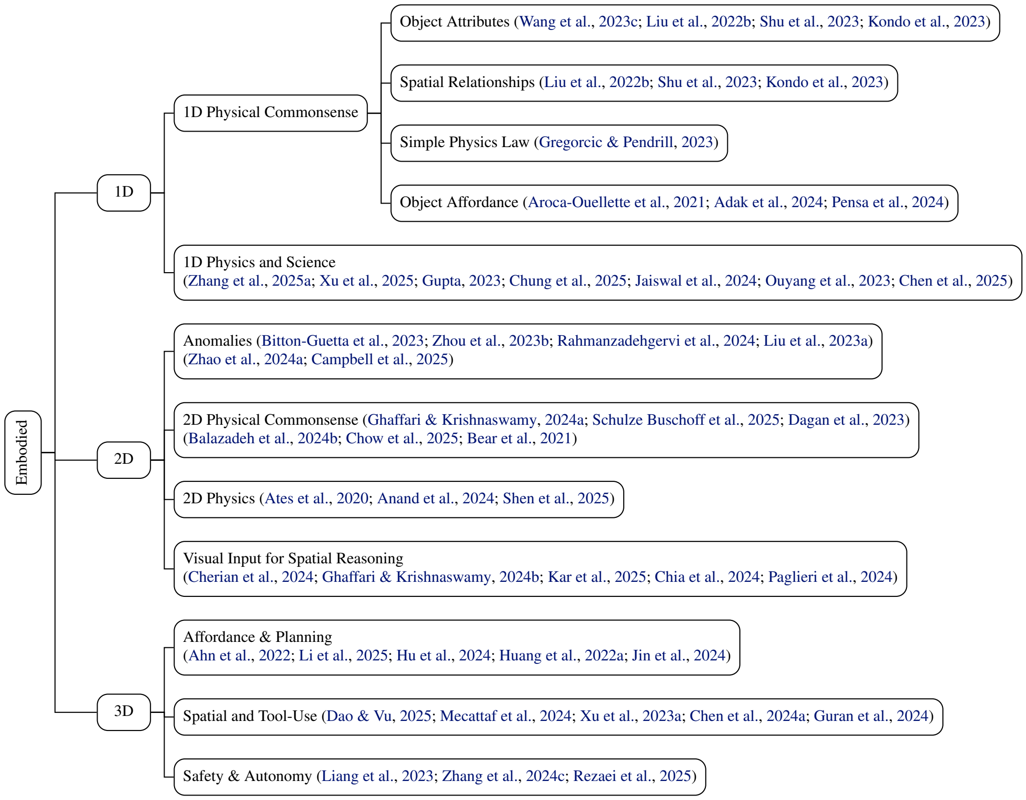


Figure 4. Taxonomy of Embodied LLM Reasoning Failures.

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.

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., 2024f). 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., 2025; Welleck & Saha, 2023; Thakur et al., 2024; Kumarappan et al., 2025), 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).

E. Other Important LLM 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 exceeding 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., 2024g; 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., 2024d; 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.

F. Examples

Table 2. 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***.
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., 2024b)**: **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 ✗

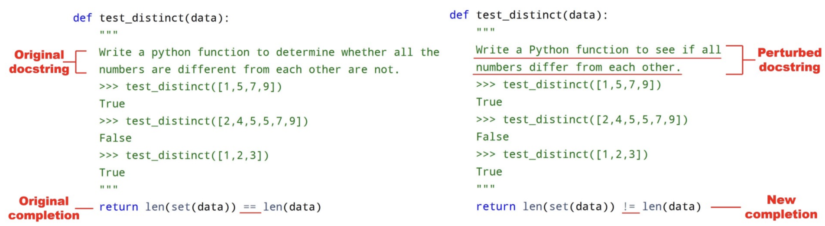
Table 8. Formal Reasoning - 4.2 Logic in Benchmarks

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: ... 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 $3 \leq x \leq 2011$. ... 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: ... 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 $3 \leq w \leq 4680$. ... 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.

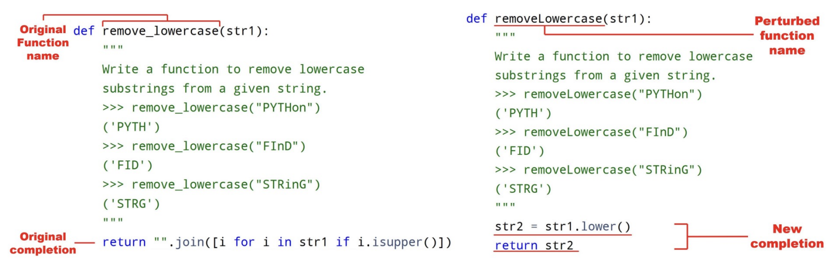
Table 9. Formal Reasoning - 4.2 Logic in Benchmarks

<table>
<tr><th>Sub-item</th><th>Examples</th></tr>
<tr><td>Coding Benchmarks</td><td>

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

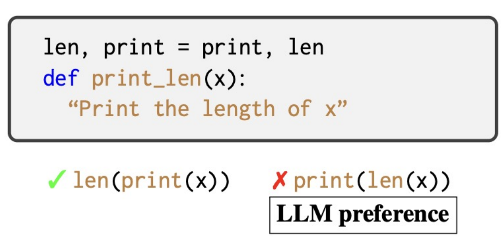


Explanation: The doc string in the starter code is changed subtly, which should not affect the generated code. Yet LLM fails on the new problem, suggesting a lack of robustness.



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.

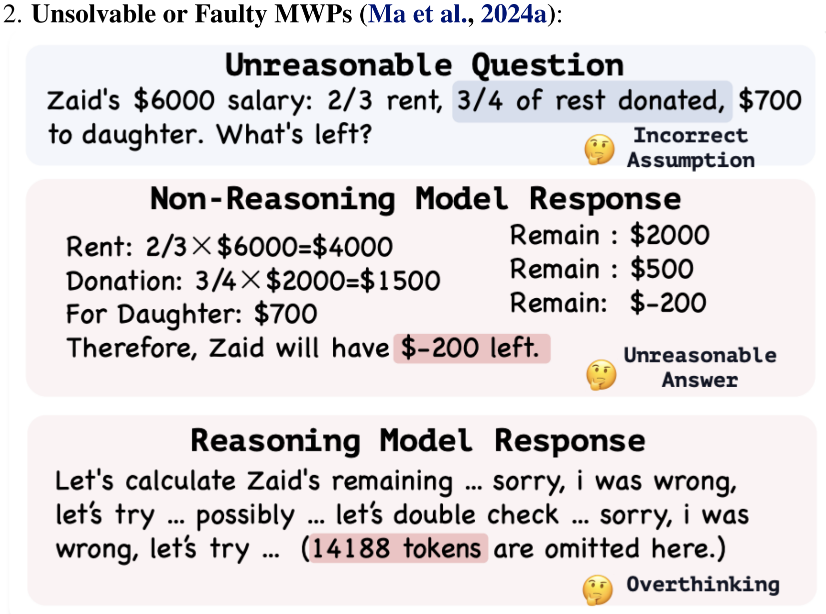
2. **Adversarial Code Changes (Miceli-Barone et al., 2023):**



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 is contains faulty assumptions. Non-reasoning LLMs give an unreasonable answer corresondingly, 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:** Yes. ✓ **Question:** A house is in an electric bulb. Is the electric bulb bigger than the house? **LLMs:** Yes. ✗
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
	1. **Detecting Anomalies (Bitton-Guetta et al., 2023):**

What's Wrong with the Picture?

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

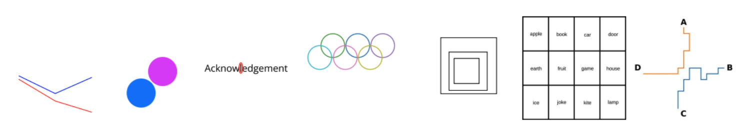
(b) a close up of a person's skates on an ice rink

(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



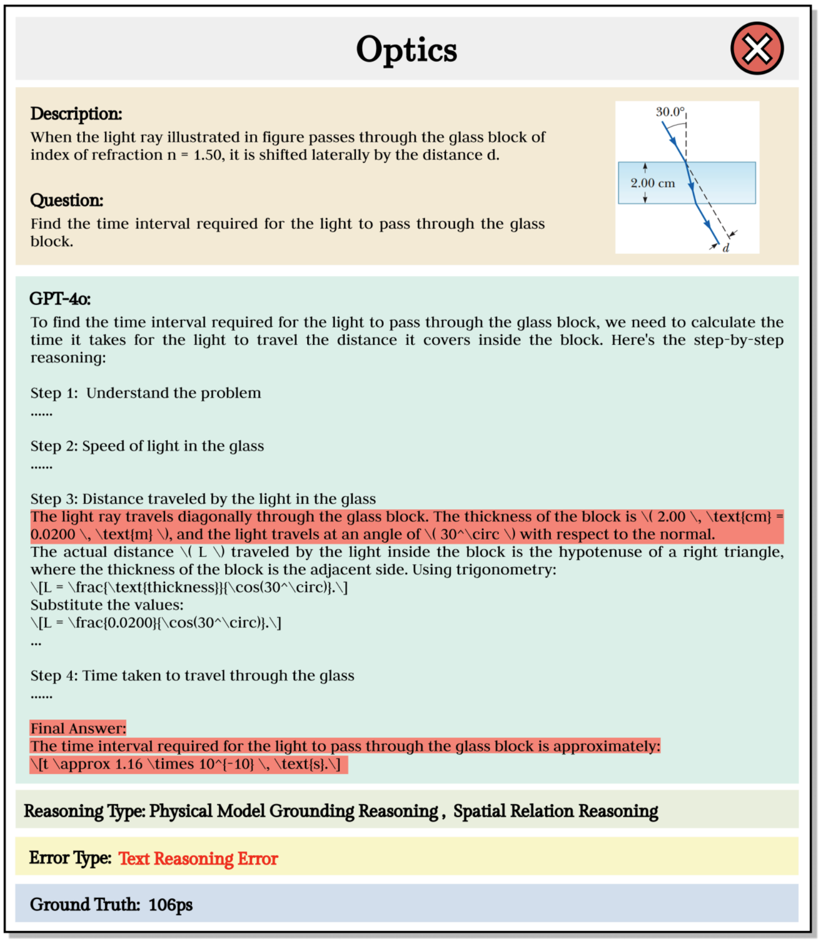
	P1		P2		P3		P4		P5		P6		P7	
	1	✗	Yes	✗	o	✗	6	✓	5	✗	3×4	✓	1	✓
	1	✗	No	✓	w	✗	5	✗	3	✓	3×4	✓	2	✗
	1	✗	Yes	✗	o	✗	5	✗	4	✗	4×4	✗	2	✗
	0	✓	No	✓	l	✓	6	✓	3	✓	3×4	✓	1	✓

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):


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

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

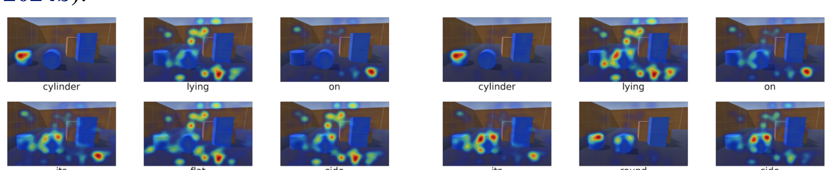
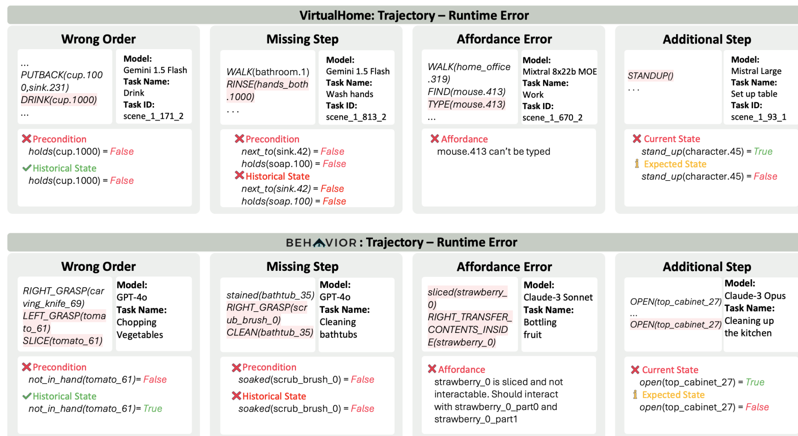
Sub-item	Examples
2D Physics and Physical Commonsense	2. **2D Physical Commonsense (Chow et al., 2025)**: **Question:** According to the video, which alteration to the light source is most likely to have triggered the phenomenon? <video 1> **Option:** (A) The color of the light changes from yellow to 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>

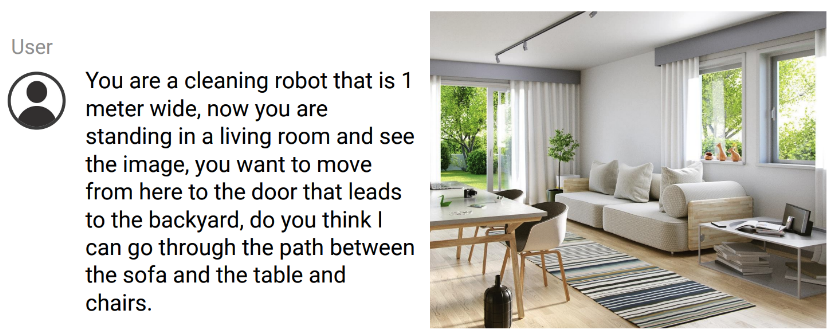
1. **Run Time Error** (Li et al., 2025):



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>