A Survey on Large Language Model Reasoning Failures

Anonymous Authors[1]

Abstract

Reasoning capabilities in Large Language Models (LLMs) have advanced dramatically, enabling impressive performance across diverse tasks. However, alongside these successes, notable reasoning failures frequently arise, even in seemingly straightforward scenarios. To systematically understand and address these issues, we present a comprehensive survey of reasoning failures in LLMs. We propose a clear categorization framework that divides reasoning failures into embodied and non-embodied types, with non-embodied further subdivided into informal (intuitive) and formal (logical) reasoning. For each category, we synthesize and discuss existing studies, identify common failure patterns, and highlight inspirations for mitigation strategies. Our structured perspective unifies fragmented research efforts, provides deeper insights into systemic weaknesses of current LLMs, and aims to motivate future studies toward more robust, reliable, and human-aligned 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; 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. Examples span a wide range of domains, from traditional linguistic areas such as machine translation (Zhu et al., 2024b; Tang et al., 2024) and endangered language learning (Zhang et al., 2024d), 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 such impressive applications, LLM reasoning as an emergent ability (Wei et al., 2022a) has been a particular topic of wide interest (Huang & Chang, 2023; Yu et al., 2023b; Qiao et al., 2023).

LLMs have shown impressive records in reasoning (Wu et al., 2025; Kıcıman et al., 2024; Plaat et al., 2024), though it remains controversial whether LLMs really leverage, and even possess, human-like reasoning procedures when solving 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.

Much psychological evidence (Cannon & Edmondson, 2005; Maxwell, 2007; Coelho & McClure, 2004) has pointed out the significance of identifying and correcting failures in human development[1]. As AI has evolved by drawing multiple, even foundational, inspirations from human brains (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 on ultimate improvements (Dreyfus, 1992; Karl et al., 2024; An et al., 2024).

Despite some existing works that prospectively realize this importance and investigate LLM reasoning failures on a case-by-case basis (Williams & Huckle, 2024; Tie et al., 2024; Helwe et al., 2021; Borji, 2023), the topic is still far from unified, which leads to insufficient attention being drawn. Such lack of a well defined and established field hinders larger-scale study, which is however a prerequisite for common patterns to be noticed, and thereby meaningful lessons to be derived. Noticing the gap, we present this survey to comprehensively unify the existing explorations scattered around. We then systematically identify meaningful patterns, drawing insights on mitigation for the failure cases and potential improvement for LLM reasoning to proceed toward. We hope this work serves not only to survey the field in a well-defined way, but also facilitate further studies and wider interests in LLM reasoning failures, which would

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

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

then lead to exciting improvements being made.

2. Definition and Formulation

2.1. Fundamentals of Reasoning

Human reasoning broadly encompasses our capacity to draw conclusions and make decisions based on available knowledge (Lohman & Lakin, 2011; Ribeiro et al., 2020). Within cognitive science and philosophy, reasoning has traditionally been analyzed through various frameworks. To systematically survey reasoning failures in LLMs, we develop a comprehensive categorization clearly distinguishing reasoning along two primary axes: ***embodied*** versus ***non-embodied***, with the latter further subdivided into ***informal*** and ***formal*** reasoning. This taxonomy is illustrated in Figure 1, with the taxonomies for each sub-category further shown in Appendix A.

Non-embodied reasoning. Non-embodied reasoning refers to reasoning processes that do not explicitly require physical embodiment or direct interaction with physical environments. Informal reasoning within this category encompasses intuitive reasoning, including everyday judgments, social interactions, and decision-making processes influenced by heuristics and biases (Piaget, 1952; Vygotsky, 1978; Kail, 1990). Formal reasoning, by contrast, involves structured, rule-based processes characterized by explicit logic, mathematics, or formally defined frameworks (Copi et al., 2016; Mendelson, 2009; Liu et al., 2023b).

Embodied reasoning. Embodied reasoning emphasizes processes inherently linked to physical interaction with environments, fundamentally relying on spatial intelligence and real-time feedback (Shapiro, 2019; Barsalou, 2008). This category includes predicting and interpreting physical interactions, as well as performing goal-directed behaviors constrained by real-world physical limitations (Huang et al., 2022b; Lee-Cultura & Giannakos, 2020).

2.2. LLM Reasoning Faliures & Common Research Practice

Despite advancements in interpretability research (Dwivedi et al., 2023; Li et al., 2024d), LLMs largely remain *black-box* systems (Luo & Specia, 2024), reflecting the complexity and opacity inherent in human cognition (Castelvecchi, 2016). Consequently, researchers typically evaluate reasoning through behavioral analyses, systematically observing model outputs in response to well-designed prompts or tasks (Ribeiro et al., 2020). Within this framework, ***LLM reasoning failures*** refer to instances where ***model responses significantly deviate from expected logical coherence, contextual relevance, or factual correctness***. Current research typically starts by identifying reasoning failures in ***intuitively simple tests***, which often expose deeper, fundamental limitations. These observed cases are then generalized through ***systematical larger-scale evaluations***, confirming their pervasiveness and broader significance. Through explicitly defining and categorizing these reasoning failures – informal, formal, and embodied – this survey unifies fragmented research findings, elucidates shared patterns, and motivates targeted efforts toward deeper understanding and systematic mitigation of critical reasoning failures in LLMs.

3. Reasoning Informally in Intuitive Applications

Humans naturally possess the ability to reason informally in intuitive, everyday situations. These abilities stem from innate cognitive functions and are further shaped by personal development, social interaction, and lived experience. Though often taken for granted, they form the foundation of much of human reasoning and decision-making. In this section, we first examine studies focused on core cognitive abilities in LLMs and then extend to reasoning in social contexts.

3.1. Individual Cognitive Reasoning

Numerous studies have demonstrated that LLMs frequently display reasoning failures related to human cognitive abilities (Han et al., 2024b; Gong et al., 2024; Galatzer-Levy et al., 2024; Suri et al., 2024). These failures may arise from a lack of certain fundamental human cognitive functions (Han et al., 2024b) or from the manifestation of cognitive biases and reasoning errors similar to those found in humans (Suri et al., 2024; Lampinen et al., 2024). We categorize these LLM reasoning failures as instances ***directly stemming from well-known human cognitive phenomena or past psychological experiments.***

Fundamental Cognitive Skills. Humans naturally possess a set of fundamental cognitive skills that are indispensable for reasoning. Many studies show LLM demonstrate systematic reasoning failures due to lack of these skills.

One critical area is ***human core executive functions*** – a set of essential cognitive processes, including working memory (Baddeley, 2020), inhibitory control (Diamond, 2013; Williams et al., 1999), and cognitive flexibility (Canas et al., 2006), that support human reasoning (Diamond, 2013). Deficiencies in these functions can lead to systematic reasoning failures. Working memory is the ability to hold and manipulate information over short periods. LLMs exhibit limited working memory; when this capacity is exceeded, they can experience systematic failures (Gong et al., 2024; Zhang et al., 2024a). Inhibitory control refers to the ability to suppress impulsive or default responses. LLMs often lack this ability, tending to default outputs (Han et al., 2024b).

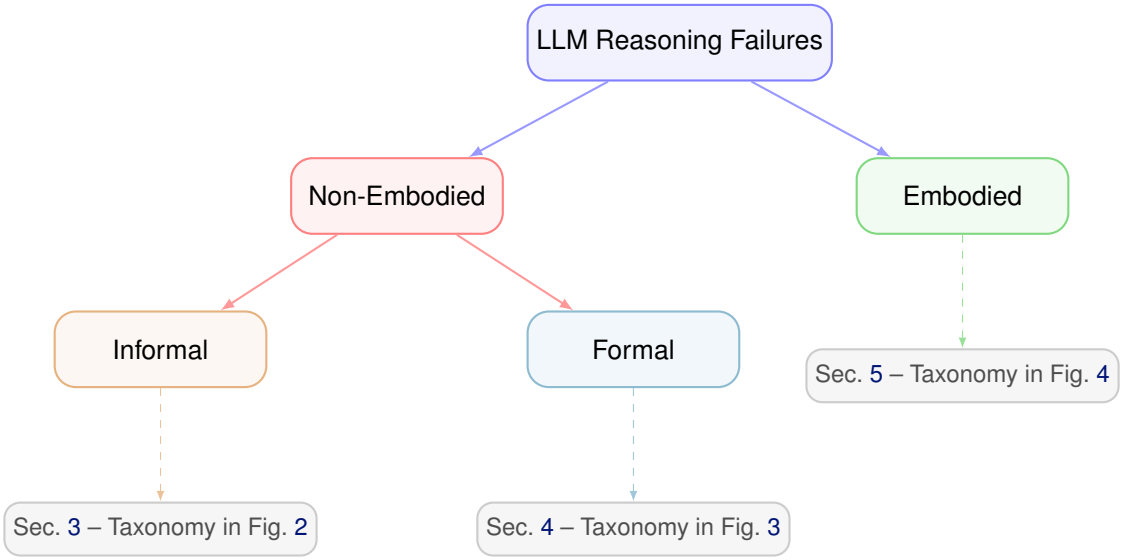

Figure 1. Overall taxonomy of LLM Reasoning Failures.

Cognitive flexibility is the capacity to shift between tasks or adapt to new rules and perspectives. Many models struggle with this, particularly in rapid task switching and adapting to changing instructions (Kennedy & Nowak, 2024).

Another aspect is ***abstract reasoning*** (Guinungco & Roman, 2020), the cognitive ability to recognize patterns and relationships of theoretical or intangible concepts and ideas. Several studies show that LLMs consistently underperform on abstract reasoning benchmarks compared to humans (Xu et al., 2023c; Gendron et al., 2023; Galatzer-Levy et al., 2024).

Efforts to enhance these capabilities include prompting techniques (Wei et al., 2022b), retrieval augmentation (Xu et al., 2023b), and architectural changes (e.g., more human-like attention mechanisms) (Wu et al., 2024d). Yet, many tasks humans perform effortlessly still remain challenging for LLMs. Future works should aim to bridge this gap by uncovering the underlying causes and guiding the development of more cognitively aligned models.

Cognitive Biases. Cognitive biases, extensively studied in human reasoning (Tversky & Kahneman, 1974; 1981), refer to systematic deviations from rational judgment, often arising from mental shortcuts, limited cognitive resources, or contextual influences, leading to predictable errors. LLMs exhibit similar cognitive biases that systematically affect their reasoning (Hagendorff, 2023); these biases are deeply ingrained, permeating a wide range of downstream tasks, making identification and mitigation critical for real-world reliability.

A key aspect is ***the content of the information***. LLMs process abstract or unfamiliar topics less effectively (Lampinen et al., 2024) – a phenomenon known as the "content effect" –and tend to favor information that aligns with preceding context or implied assumptions, reflecting confirmation bias (O'Leary, 2025; Shi et al., 2024; Malberg et al., 2024). They also exhibit group attribution bias (Hamilton & Gifford, 1976; Allison & Messick, 1985) and negativity bias (Rozin & Royzman, 2001), prioritizing popular content (Echterhoff et al., 2024) and negative inputs (Yu et al., 2024c; Malberg et al., 2024).

Even the content remains the same, ***the presentation of the information*** can introduce bias. LLMs are vulnerable to anchoring bias (Lieder et al., 2018), where their reasoning disproportionately relies on initial inputs. They also exhibit framing effects (Druckman, 2001), where different phrasing or structuring of the same information skews their outputs (Jones & Steinhardt, 2022; Echterhoff et al., 2024; Suri et al., 2024; Nguyen, 2024; Malberg et al., 2024; Lou & Sun, 2024). Furthermore, the perspective (e.g., first-person) (Cohn et al., 2024), length (Koo et al., 2023), and inclusion of irrelevant or distracting information (Shi et al., 2023) can all impact reasoning, leading to LLMs being systematically biased similar to humans.

These human-like biases may stem from reinforcement learning from human feedback (RLHF) and instruction tuning (Itzhak et al., 2023). While most current mitigation strategies rely on prompting (Echterhoff et al., 2024) – such as inserting debiasing instructions – they remain surface-level and lack fundamental control. Interestingly, inducing specific personalities in LLMs has been shown to modulate cognitive biases (Shi et al., 2024). Future research should aim for a deeper understanding of LLM behavior and more

principled methods for identifying and controlling cognitive biases.

3.2. Implicit Social Reasoning

Certain weaknesses in LLM cognitive reasoning become evident only within the context of a "society" involving others. 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 – such as beliefs, intentions, and emotions – to oneself and others, and to understand that other's mental states may differ from one's own (Frith & Frith, 2005). This capacity is crucial for human social reasoning, enabling humans to interpret behaviors, predict actions, and judge and navigate complex interpersonal interactions. Human ToM typically emerges naturally in early childhood with key developmental milestones, such as passing false belief tasks (understand that others' beliefs may be incorrect or different) (Wimmer & Perner, 1983).

ToM has long been studied in psychology and cognitive science given its central role in human cognition. More recently, researchers have begun applying these same principles to LLMs to evaluate 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 (e.g., infer what another individual perceives) (Sap et al., 2022), and unexpected content tasks (e.g., predict what someone believes is inside a mislabeled container they haven't seen opened) (Pi et al., 2024). Even advanced models such as GPT-4 struggle with these tasks that are trivial for human children. Additionally, minor modifications in task phrasing often lead to drastic drops in performance, showing ToM reasoning is quite brittle (Ullman, 2023; Kosinski, 2023; Pi et al., 2024; Shapira et al., 2023).

While newer models, such as o1-mini and GPT-4o, have shown improvements in tracking what information others are aware of (Gu et al., 2024; Zhou et al., 2023d), they still fall short in predicting others' behaviors, making appropriate judgments, and translating this understanding into coherent actions. They also exhibit limitations in emotional reasoning, including difficulties in emotional intelligence (EI) (Sabour et al., 2024; Hu et al., 2025; Amirizaniani et al., 2024a; Vzorinab et al., 2024), affective bias (Chochlakis et al., 2024), and cultural variations in emotional understanding (Havaldar et al., 2023).

While prompting techniques like Chain-of-Thought (CoT) show some potential (Gandhi et al., 2024), fundamental gaps remain. This suggests that the limitations may arise more deeply from LLMs' architecture, training paradigms, and lack of embodied cognition (Strachan et al., 2024; Sclar et al., 2023). Since the ability to infer others' mental states is essential for social reasoning, we encourage future research to move beyond inference-time prompting and explore deeper mechanisms for control and alignment.

Social Norms and Moral Values. Beyond ToM, another critical failure in LLMs' social reasoning involves their handling of ***social norms, moral values, and ethical principles***. These elements guide human decision-making, shaping what is considered appropriate, just, or acceptable in various social contexts. While humans develop moral and ethical reasoning through experience, LLMs, trained purely on text, frequently exhibit systematic reasoning failures, leading to unreliable ***social, moral, and ethical reasoning***.

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 different languages (Agarwal et al., 2024). Additionally, fine-tuning can further exacerbate these inconsistencies, sometimes prioritizing task-specific optimization over ethical coherence (Yu et al., 2024a).

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

3.3. Explicit Social Reasoning

In reasoning, "society" can refer to not only an abstract concept but also 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 that is challenging for current LLMs.

In MAS, 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, which LLMs frequently fail (Li et al., 2023a; Cross et al., 2024; Guo et al., 2024c; Han et al., 2024c) as they often rely on local or recent information (Piatti et al., 2024; Zhang et al., 2023; Han et al., 2024c). Furthermore,

inefficient communication and ToM within MAS (Guo et al., 2024c; Agashe et al., 2024) lead to misinterpretations and inaccurate representations of other agents, causing strategic misalignments (Pan et al., 2025; Li et al., 2023a; Cross et al., 2024; Han et al., 2024c). Additionally, 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 result from both capabilities of individual LLMs and the design of MAS (Pan et al., 2025). For efficient and reliable real-world usage, future research should focus on improving the communication abilities of individual models and designing more robust systems with stringent specifications and verification processes.

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, leading to flawed thought processes and harmful decisions. Logic spans a spectrum – from implicit structures in natural languages (Iwańska, 1993), to symbols (Lewis et al., 1959) and math (Shoenfield, 2018) – progressing relatively from *informal* to *formal*. This section follows that progression.

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 just 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 – a step usually easy for humans. Studies show systematic failures in basic two-hop reasoning – combining only two facts across documents – and worsening performance as compositional depth increases (Balesni et al., 2024; Zhao & Zhang, 2024; Xu et al., 2024b). This weakness extends beyond basic tasks, to compositions of math problems (Zhao et al., 2024b; Hosseini et al., 2024) (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).

Chain-of-thought (CoT) prompting (Wei et al., 2022b) improves on this by making reasoning steps explicit at inference time (Balesni et al., 2024). Still, latent compositionality is more efficient in practice yet harder to achieve (Yang et al., 2024c). Toward this, Li et al. (2024e) identify 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 the training data (Yu et al., 2024b).

Specific Logical Relations. Both reversal curse and compositional reasoning reflect foundamental 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. 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 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 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

Whereas 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 static benchmark reliability (Zhou et al., 2023c; Zheng et al., 2024b; Xu et al., 2024a; Patel et al., 2021), these studies introduce ***logic-preserving*** transformations to existing tasks, 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 or events (Chen et al., 2024c; Yamin et al., 2024), or superficially editing contexts (e.g., character names) (Jiang et al., 2024b; Mirzadeh et al., 2024; Shi et al., 2023; Wang & Zhao, 2024). Such modifications keep tasks essentially the same. Performance drops thus reveal unstable reasoning and reduced trustworthiness.

Math Word Problem (MWP) Benchmarks. Certain benchmarks inherently possess logical structures that facilitate targeted perturbations. MWPs exemplify this, as their logic is 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 reversing 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., 2025a) – 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 optionally some starter code. Common transformations include syntactically editing doc strings (Xia et al., 2024; Wang et al., 2022; Sarker et al., 2024), renaming functions or variables (Wang et al., 2022; Hooda et al., 2024), or 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 provided contexts, testing whether LLMs identify and adapt to them (Miceli-Barone et al., 2023; Dinh et al., 2023), thereby evaluating deeper reliability.

Mitigations & Extensions. The failures are attributed to ***lack of robustness*** or ***overfitting to public datasets***. Robustness-related issues are commonly mitigated by applying perturbations to diversifying training data (Patel et al., 2021), thus enhancing resilience to variations. Overfitting concerns are addressed through dynamically evolving (Jain et al., 2024a; White et al., 2024) or privately maintained datasets (Rajore et al., 2024). Beyond individual benchmarks, Hong et al. (2024) automates a set of transformations for math and coding benchmarks, and Wu et al. (2024e) alters common assumptions of well-known tasks.

4.3. Arithmetic & Mathematics

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

Counting. Despite its simplicity, counting poses 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., 2024f; Shin & Kaneko, 2024), positional encoding (Chang & Bisk, 2024), and training data composition (Allen-Zhu & Li, 2024). Mitigations via supervised fine-tuning (Zhang & He, 2024) and engaged reasoning (Xu & Ma, 2024) have been proposed, yet robust counting remains elusive.

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 notably in middle-digits (Deng et al., 2024). Surprisingly, LLMs fail at simpler tasks (determining the last digit) but succeed in harder ones (first digit identification) (Gambardella et al., 2024). Those inconsistencies lead to failures for practical tasks like temporal reasoning (Su et al., 2024).

These issues are attributed to 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., 2023b), digit-order reversals (to place attention on least significant digits, aligning with how humans do multiplication) (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).

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 task templates (Nezhurina et al., 2024) to large-scale evaluations on a domain of math (Wei et al., 2023b; Boye & Moell, 2025; Fan et al., 2024). Additionally, LLMs exhibit susceptibility when faced with

unsolvable or faulty MWPs (Ma et al., 2024a; Rahman et al., 2024; Tian et al., 2024). LLMs struggle even in *assessing* reasoning process on MWPs (Zhang et al., 2024g), an arguably easier task. Given these persistent challenges, current efforts in MWPs prioritize developing general methods to improve overall performance rather than addressing individual failure types.

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 upon information in the physical world with an accurate understanding of spatial relationships, object dynamics, and physical principles (Lee-Cultura & Giannakos, 2020). While this comes naturally to humans (Varela et al., 2017) – and even to many animals (Andrews & Monsó, 2021) – it remains a significant challenge for LLMs. Without a true grounding in the physical world, LLMs often struggle with basic physical reasoning, leading to systematic errors and unrealistic predictions (Wang et al., 2023c; Ghaffari & Krishnaswamy, 2024b).

Despite growing interest in spatial intelligence, research into LLMs' physical reasoning failures is still sparse. This section will outline key failures in 1D text-based reasoning, 2D vision-based perception, 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, forces, and people 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., 2022; Shu et al., 2023; Kondo et al., 2023), ***spatial relationships*** (e.g., above, inside, next to) (Liu et al., 2022; 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 and early interaction with the environment. For LLMs, however, this intuitive grasp of physical dynamics is difficult to achieve, 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 and demand some real-world understanding, LLMs exhibit systematic failures.

Physics & Scientific Reasoning. Beyond basic physical commonsense, LLMs also struggle with formal physics reasoning and scientific problem-solving. Complex physics and scientific problems require not just factual recall and intuition but multi-step logical deduction, quantitative reasoning, and correct use of physical laws – areas where even today's most advanced models like o1 (Jaech et al., 2024) and o3-mini (OpenAI, 2025) have notable deficits (Zhang et al., 2025; Xu et al., 2025; Gupta, 2023; Chung et al., 2025). Despite possessing extensive scientific knowledge, LLMs often fail to structure and apply it effectively, leading to systematic reasoning failures in complex problem-solving and scientific discovery (Jaiswal et al., 2024; Ouyang et al., 2023; Chen et al., 2025).

5.2. 2D – Perception-Based Physical Reasoning Failures.

What's Wrong with the Picture? "What's Wrong with the Picture?" is a well-known visual reasoning game, where participants analyze a static image to identify abnormalities. Similar strategies have been applied to vision-language models (VLMs) to reveal their surprising failures on simple tasks such as detecting anomalies (Bitton-Guetta et al., 2023; Zhou et al., 2023b), identifying object overlaps and counts (Rahmanzadehgervi et al., 2024), understanding image content and spatial relations (Liu et al., 2023a; Zhao et al., 2024a).

These errors likely stem from two key issues. First, models disproportionately trust text or common scenarios from their training data rather than accurately describing what is actually shown, leading to failures in tasks such as identifying abnormalities (Deng et al., 2025; Bitton-Guetta et al., 2023; Zhou et al., 2023b). Second, some failures may be explained by the binding problem in cognitive science, where the brain– or a model–struggles to process multiple distinct objects simultaneously due to limited shared resources, often leading to confusion or interference between them (Campbell et al., 2025).

2D Physics and Physical Commonsense. Building on tasks that detect simple anomalies or object properties in static images, a deeper challenge emerges when models must reason about physical commonsense and physics in visually grounded settings. Despite visual input, VLMs continue to struggle with physical commonsense (Ghaffari & Krishnaswamy, 2024a; Schulze Buschoff et al., 2025; Dagan et al., 2023; Balazadeh et al., 2024; Chow et al., 2025; Bear et al., 2021) and advanced physics (Ates et al., 2020; Anand et al., 2024), exhibiting performance gaps similar to those seen in text-only settings discussed in Section 5.1.

Visual Input for Spatial Reasoning. Noting that real-world spatial reasoning requires evolving spatial relationships rather than isolated snapshots, recent works introduce 2D-based simulated environments to test models' understanding of ***motion and object interactions*** (e.g., predicting

how objects move after impact) (Cherian et al., 2024), ***prediction and manipulation*** tasks (e.g., determining where and how to place objects to achieve stability) (Ghaffari & Krishnaswamy, 2024b), ***spatial communication and alignment*** (e.g., communicate spatial locations) (Kar et al., 2025), and ***embodied planning and decision-making*** (e.g., suggesting actions in multi-step interactive tasks) (Chia et al., 2024; Paglieri et al., 2024). While VLMs exhibit some basic spatial knowledge, they consistently struggle to compose and apply it in dynamic, agentic tasks, revealing a gap in structured spatial reasoning.

5.3. 3D – Real-World Physical Reasoning Failures

Real embodied reasoning requires an agent to actively interact with its environment–whether through physical robotics or interactive simulations that go beyond static images or simple 2D snapshots. The agent should process real-time goals and feedback, and execute physical actions. Unlike 1D (text-only) and 2D (image-based) tasks, 3D reasoning focuses on the concept of ***action*** rather than passively analyzing. Despite recent progress in robotics and embodied AI, LLMs and VLMs still face fundamental challenges, showing inaccurate spatial modeling, unrealistic affordance prediction, tool-use failures, and unsafe behavior. This section highlights key failure cases from both simulated and real-world studies.

Real-World Failures in Affordance and Planning. A key failure in embodied environments is the models' inability to generate feasible and coherent action plans. LLMs and VLMs often produce physically impossible or inefficient actions due to affordance error (i.e., incorrect inference about object action possibilities) (Ahn et al., 2022; Li et al., 2025; Hu et al., 2024; Huang et al., 2022a; Jin et al., 2024) as well as logically disordered or looping actions due to limitations in causal real-world understanding (Jin et al., 2024; Hu et al., 2024).

A critical factor underlying these failures is the autoregressive nature of LLMs. Naive LLMs and VLMs generate step by step plans and lack the mechanisms to check and handle 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 can significantly reduce these errors (Liang et al., 2023; Wang et al., 2023a). This finding underscores that embodied reasoning isn't just about immediate spatial or manipulation reasoning, but also about sustained goal-directed behavior over time.

Spatial and Tool-Use Reasoning. Even when LLMs successfully decompose tasks and generate seemingly valid plans, failures still arise due to poor spatial reasoning (Dao & Vu, 2025; Mecattaf et al., 2024) and an inability to generalize tool-use strategies (Xu et al., 2023a). 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 errors in both spatial awareness and interaction with physical environments. A key contributor to these failures is the absence of a robust internal world model (Dao & Vu, 2025; Wu et al., 2025), which often forces LLMs to rely on external aids–such as explicit spatial prompts–to compensate for their limited spatial reasoning. To advance embodied intelligence, future research must focus on strengthening LLMs' internal representations of space –such as spatial memory and quantitative spatial understanding.

Safety and Long-Term Autonomy. Ensuring the safety and reliability of LLM-driven embodied agents remains a major challenge. LLM-generated robotic task plans are highly sensitive to prompt design (Liang et al., 2023) and vulnerable to adversarial manipulation (Zhang et al., 2024c). Moreover, these systems don't act appropriately in real world (Rezaei et al., 2025). These findings highlight the urgent need for more robust, self-correcting, and safety-aware embodied AI systems before real-world deployment.

6. 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. 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.

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 abili-

ties 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.

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. Do llms exhibit human-like reasoning? evaluating theory of mind in llms for open-ended responses. *arXiv preprint arXiv:2406.05659*, 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, 2024b. URL https://arxiv.org/abs/2406.05659.

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. Synthetic vision: Training vision-language models to understand physics. *arXiv preprint arXiv:2412.08619*, 2024.

Balesni, M., Korbak, T., and Evans, O. The two-hop curse: Llms trained on a->b, b->c fail to learn a–>c, 2024. URL https://arxiv.org/abs/2411.16353.

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.

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.

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.

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.

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.

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.

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.

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*, 2025.

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.

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.

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.

Druckman, J. N. Evaluating framing effects. *Journal of economic psychology*, 22(1):91–101, 2001.

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.

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.

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., 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.

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.

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.

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, 2024b. URL https://arxiv.org/abs/2409.15454.

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.

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.

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, 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, 2025a. 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, 2025b.

Huang, S., Song, P., George, R. J., and Anandkumar, A. Leanprogress: Guiding search for neural theorem proving via proof progress prediction, 2025c. URL https://arxiv.org/abs/2502.17925.

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.

Itzhak, I., Stanovsky, G., Rosenfeld, N., and Belinkov, Y. Instructed to bias: instruction-tuned language models exhibit emergent cognitive bias. *arXiv preprint arXiv:2308.00225*, 2023.

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

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.

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, 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, 2025.

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.

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.

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.

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., 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.

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, 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.

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, M., Wang, J., Lin, T., Ma, Q., Fang, Z., and Wu, Y. An empirical study of the code generation of safety-critical software using llms. *Applied Sciences*, 14(3), 2024a. ISSN 2076-3417. doi: 10.3390/app14031046. URL https://www.mdpi.com/2076-3417/14/3/1046.

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*, 2022.

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*, 2024b.

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.

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.

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

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.

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

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.

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 multilayered 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.

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.

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.

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.

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.

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.

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.

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.

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

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.

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.

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

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, 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.

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.

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.

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, 2024. URL https://arxiv.org/abs/2410.00988.

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.

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.

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.

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, 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, 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, 2025.

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.

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, 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, 2023b. 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.

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, K., Choi, Y. M., Song, Z., He, T., Wang, W. Y., and Li, L. Hire a linguist!: Learning endangered languages with in-context linguistic descriptions, 2024d. URL https://arxiv.org/abs/2402.18025.

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 2024e. 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, 2024f. 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*, 2025.

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 2024g. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.524. URL https://aclanthology.org/2024.findings-acl.524/.

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*, 2024h.

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, 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.

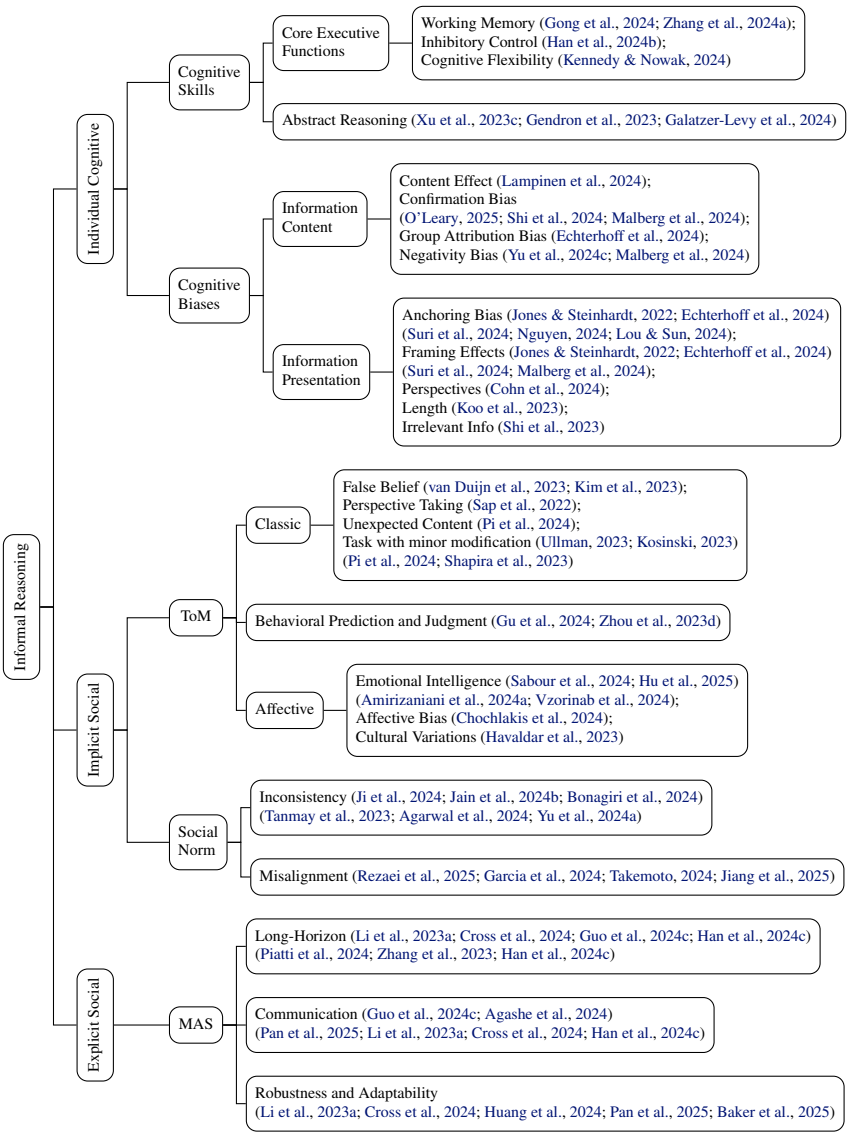

Figure 2. Taxonomy of Informal LLM Reasoning Failures.

A. 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 in Section 2, 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).

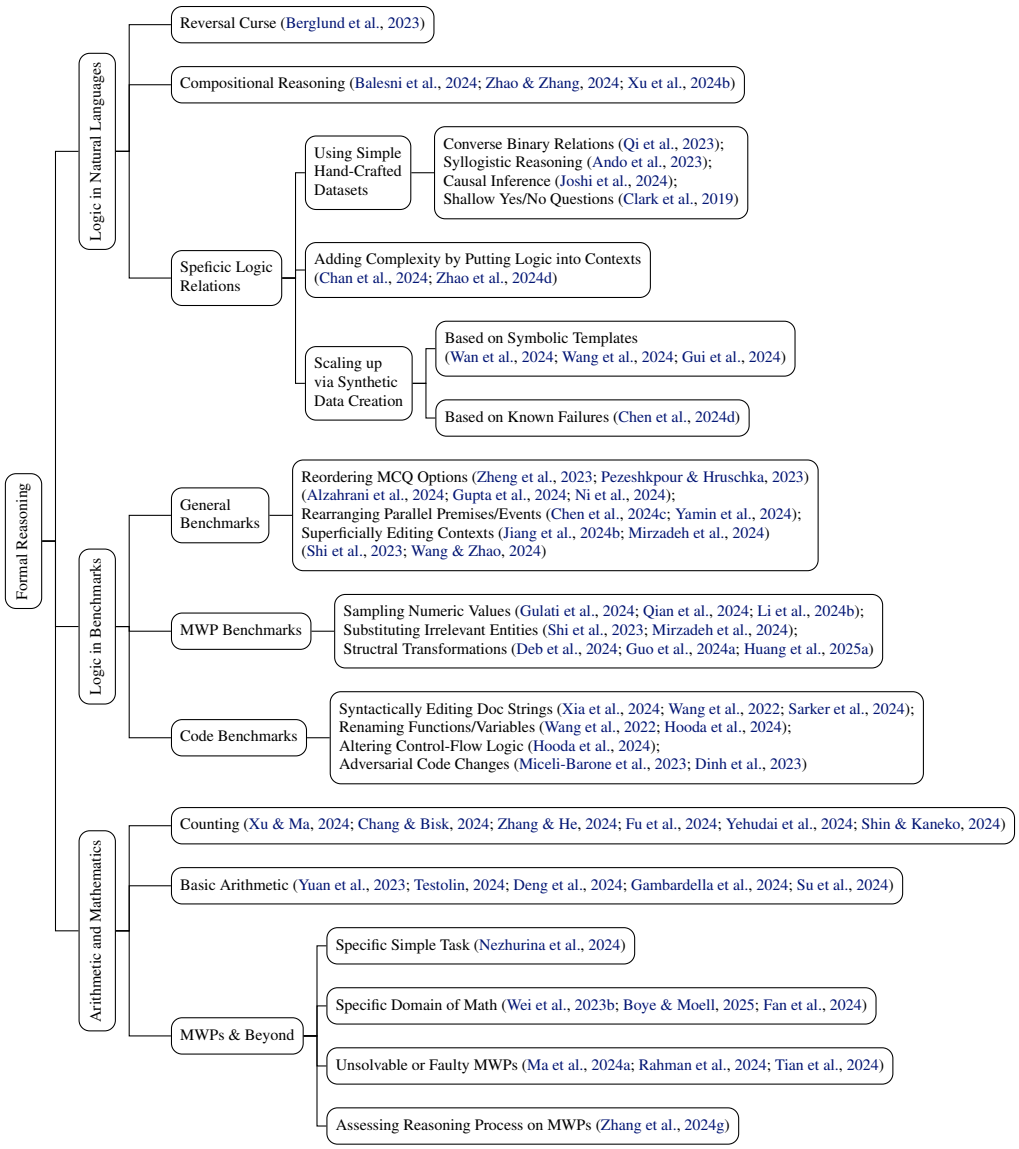

Figure 3. Taxonomy of Formal LLM Reasoning Failures.

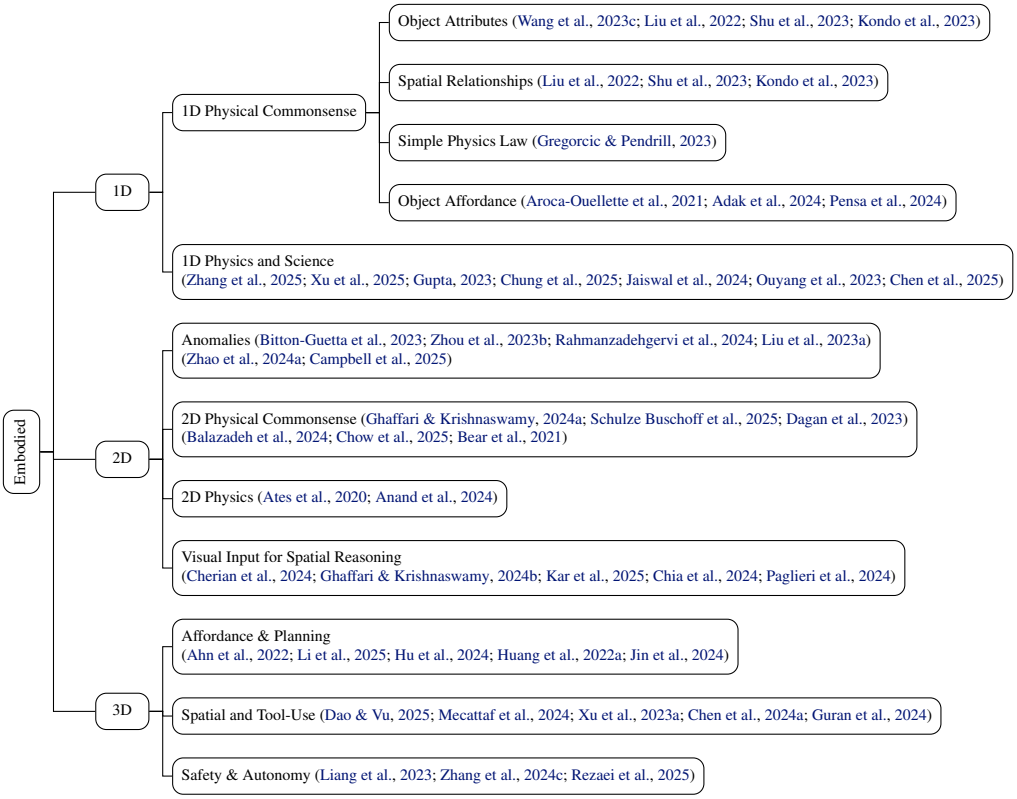

Figure 4. Taxonomy of Embodied LLM Reasoning Failures.

B. Artifacts

Upon the release of this survey, we will make public a comprehensive collection of categorized works in the field of LLM reasoning failures, to facilitate future research by providing an easy 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.

C. 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. 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 autoformalization (Wu et al., 2022; Jiang et al., 2023a; Murphy et al., 2024), efficient proof search (Lample et al., 2022; Huang et al., 2025c; 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 (Liu et al., 2024a; Kasibatla et al., 2024; Thompson et al., 2025).

D. 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., 2024h; Yao et al., 2023; Wen et al., 2024) – 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., 2025b). 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., 2024b), mitigate(Han et al., 2024a; Owens et al., 2024), and regulate (Zheng et al., 2024a; Jiang et al., 2023b) these biases in order to promote fairness and justice. Nevertheless, significant challenges persist. Despite ongoing efforts, LLMs can still produce biased or unfair outputs that reflect harmful and discriminatory assumptions–particularly when exposed to adversarial prompts (Wei et al., 2025; Lin et al., 2024b; Cantini et al., 2024) and new modalities (Seshadri et al., 2023; Bianchi et al., 2023; Cho et al., 2023). Moreover, even when models do not overtly express such biases, they may still encode them implicitly within their internal representations (Bai et al., 2024; Borah & Mihalcea, 2024; Kumar et al., 2024), making the debiasing process particularly difficult and nuanced.

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