frontiers | Frontiers in Artificial Intelligence

TYPE Perspective
PUBLISHED 08 January 2025
DOI 10.3389/frai.2024.1464690

OPEN ACCESS

EDITED BY
Pradeep K. Murukannaiah,
Delft University of Technology, Netherlands

REVIEWED BY
Michiel Van Der Meer,
Idiap Research Institute, Switzerland

*CORRESPONDENCE
Andrea Passerini
✉ andrea.passerini@unitn.it

RECEIVED 14 July 2024
ACCEPTED 19 December 2024
PUBLISHED 08 January 2025

CITATION
Passerini A, Gema A, Minervini P, Sayin B and Tentori K (2025) Fostering effective hybrid human-LLM reasoning and decision making. *Front. Artif. Intell.* 7:1464690.
doi: 10.3389/frai.2024.1464690

Fostering effective hybrid human-LLM reasoning and decision making

Andrea Passerini[1]*, Aryo Gema[2], Pasquale Minervini[2], Burcu Sayin[1] and Katya Tentori[3]

[1]Department of Information Engineering and Computer Science, University of Trento, Trento, Italy, [2]School of Informatics, University of Edinburgh, Edinburgh, United Kingdom, [3]Center for Mind/Brain Sciences, University of Trento, Trento, Italy

The impressive performance of modern Large Language Models (LLMs) across a wide range of tasks, along with their often non-trivial errors, has garnered unprecedented attention regarding the potential of AI and its impact on everyday life. While considerable effort has been and continues to be dedicated to overcoming the limitations of current models, the potentials and risks of human-LLM collaboration remain largely underexplored. In this perspective, we argue that enhancing the focus on human-LLM interaction should be a primary target for future LLM research. Specifically, we will briefly examine some of the biases that may hinder effective collaboration between humans and machines, explore potential solutions, and discuss two broader goals—mutual understanding and complementary team performance—that, in our view, future research should address to enhance effective human-LLM reasoning and decision-making.

KEYWORDS

hybrid intelligence, human-AI collaboration, LLMs, biases, mutual understanding, complementary team performance

1 Introduction

The release of chatGPT has raised unprecedented attention and generated high expectations on the capabilities of AI systems that leverage large language model (LLM) technologies. These systems have demonstrated impressive results across a wide range of tasks (Liu et al., 2023b; Yang et al., 2024), such as language translation (Jiao et al., 2023), text summarization (Pu and Demberg, 2023), question-answering (Bahak et al., 2023), reasoning (Bang et al., 2023), and text generation (Chen et al., 2023b; Jeblick et al., 2022), prompting questions about the potential emergence of "thinking machines" and artificial general intelligence sparks (Bubeck et al., 2023). However, several studies have highlighted the limitations of these systems. Just to provide a few examples, they have been shown to provide entirely fabricated information (Huang et al., 2023), to exhibit sensitivity to small changes in the way questions are posed (Pezeshkpour and Hruschka, 2023), and to agree with human opinions regardless of content (Sharma et al., 2023).

Human beings are also far from being entirely rational, and not in an obvious way. The deviations of human reasoning from normative benchmarks create an intriguing puzzle that is not yet completely understood in Cognitive Science. On the one hand, systematic and persistent biases manifest even in well-motivated and expert individuals engaged in simple, high-stakes probability tasks (Baron, 2023). This suggests that reasoning errors do not stem from carelessness, computational limitations, or lack of education, nor are they necessarily caused by "external constraints" such as inadequate information or time pressure. On the other hand, individuals are often capable of complex inferences (Tenenbaum et al., 2011; Mastropasqua et al., 2010). In particular, evidential reasoning—i.e., the assessments of the perceived impact of evidence—appears to be quite effective, demonstrating greater accuracy and consistency over time compared to corresponding posterior probability judgments (Tentori et al., 2016). This holds true even though, from a formal standpoint, calculating the former is no easier than calculating the latter.

Notably, there are solid reasons to believe that neither LLM-based AI systems nor humans will turn into completely rational agents anytime soon. With regard to the former, the inherent mechanisms of LLMs impose significant constraints on their capabilities. Bender and Koller (2020) and Bender et al. (2021) coined the term "stochastic parrots" to highlight the fact that LLMs focus on form over meaning and stressed the difficulty of getting the latter from the former. More recently, Mahowald et al. (2024) formalized the problem in terms of the distinction between formal and functional linguistic competence, arguing that LLM architectures require substantial modifications to have a chance of achieving the latter. Finally, Xu et al. (2024b) indicated that inconsistencies between LLMs and the real world are, to some extent, inevitable. In a similar vein, efforts to enhance human rationality by using visual aids (Khan et al., 2015), promoting accountability (e.g., Boissin et al., 2023), or shaping external environments (e.g., *nudging*, Thaler and Sunstein, 2009) have often yielded modest results that are not easily generalizable to other contexts (Chater and Loewenstein, 2023). The limited effectiveness of these interventions suggests that the causes of reasoning biases are deeply ingrained in our cognitive processes, and we cannot expect to eradicate them, at least not in the near future.

Humans and LLMs are not only imperfect yet highly capable, but they also differ significantly in their respective strengths and weaknesses (Chang et al., 2023; Shen et al., 2023; Felin and Holweg, 2024; Leivada et al., 2024). Thus, while mere interaction between the two does not guarantee success, a carefully designed human-LLM synergy has the potential to prevent critical problems and achieve results that surpass what either could accomplish alone. Indeed, recent research highlights human-LLM collaboration as a key direction toward realizing genuinely human-centered AI. (Dellermann et al., 2019; Akata et al., 2020; Lawrence, 2024; Wang et al., 2024; Ma et al., 2024; Liao and Wortman Vaughan, 2024). However, in our view, effectively addressing this issue necessitates a significant shift in perspective. The primary challenge we must confront—and one that will increasingly be faced in the future—lies not so much in the specific boundaries of human rationality or the current technological limitation of LLMs, but rather in the nature and severity of biases that can arise from their *interaction*. For this reason, we do not aim to provide an exhaustive list of the many cognitive biases that individuals—and, in some cases, LLMs—exhibit. Instead, we will focus on three major problems of LLMs—*hallucinations, inconsistencies* and *sycophancy*—demonstrating how they can impact the interplay with humans. We will then discuss two key desiderata, *mutual understanding* and *complementary team performance*, which, in our opinion, future research should address more comprehensively to foster effective human-LLM reasoning and decision-making.

2 Potential weaknesses in human-LLM interaction

One of the most well-known problems of LLMs is hallucination, which refers to their distinct possibility of generating outputs that do not align with factual reality or the input context (Huang et al., 2023). Hallucinations in LLMs have several causes, from flawed data sources (Lin et al., 2022b) to architectural biases (Li et al., 2023b; Liu et al., 2023a). To exacerbate the issue, when LLMs engage in hallucination, they maintain an aura of authority and credibility by generating responses that appear coherent and well-formed in terms of natural language structure (Berberette et al., 2024; Su et al., 2024). Such a behavior can easily lead to an *automation bias* (Cummings, 2012), where users tend to over-rely on information and suggestions from automated systems compared to those from their peers. Indeed, while people can easily detect nonsensical or blatantly unrelated outputs from LLMs when they have a good knowledge of the topic, they are more likely to overlook such errors when they lack expertise in the subject. This creates a paradox: one must already possess the correct answer to reliably avoid being misled by LLMs. Nonetheless, expertise itself is not a guarantee that everything will go smoothly. Humans, including professionals such as, for example, physicians, often exhibit a tendency known as *overconfidence* (Hoffrage, 2022), where they tend to overestimate their abilities or the accuracy of their knowledge. Predicting which of these somewhat opposite attitudes would prevail in a given interaction between humans and LLMs is extremely difficult. LLMs could, in principle, counteract overconfidence by providing negative feedback to users. However, what might seem like an easy solution runs into another characteristic of these systems: their tendency toward sycophancy, which is the inclination to please users by generating responses that are agreeable rather than strictly accurate, especially when trained with biased human feedback or tasked with generating content in subjective domains (Sharma et al., 2023; Ranaldi and Pucci, 2023; Wei et al., 2023). Furthermore, overly critical feedback may lead to *algorithm aversion bias* (Dietvorst et al., 2014), where users disregard information that conflicts with their previous beliefs, even when it is actually pertinent and correct. This bias reflects the skepticism with which humans—especially professionals in high-stakes fields like healthcare and law, where accountability is paramount—often view the advanced capabilities of LLMs (Park et al., 2023; Cheong et al., 2024; Choudhury and Chaudhry, 2024; Eigner and Händler, 2024; Watters and Lemanski, 2023). Additionally, algorithm aversion may be fueled by a loss of confidence following unsatisfactory initial interactions (Huang et al., 2024; McGrath et al., 2024). In particular, LLM inconsistency—reflected in their tendency to produce varying outputs for very similar (or even identical) inputs—can easily leave lasting impressions of unreliability. This issue is exacerbated by high prompt sensitivity, where the LLM tend to provide different answers even with slight changes in how questions are phrased (Pezeshkpour and Hruschka, 2023; Voronov et al., 2024; Mao et al., 2024; Sayin et al., 2024). As a consequence, individuals may become increasingly reluctant to utilize LLMs when confronted with important reasoning and decision-making tasks.

Let us now consider a situation that, in principle, would be expected to unfold more smoothly—namely, one in which neither LLMs nor humans are outright incorrect. It might be assumed that accuracy alone would suffice to prevent errors; however, unfortunately, this is not necessarily the case. A well-known bias that could persist or even intensify in interactions where humans feel competent and LLMs provide reliable evidence is *confirmation bias*: the tendency to selectively seek, interpret, and recall information that supports existing beliefs (Nickerson, 1998). Indeed, when users query LLMs based on initial hypotheses and the models provide selective answers mainly based on local context, a vicious cycle can be fueled. A closely related cognitive bias that may similarly be exacerbated in interactions with LLMs is *belief bias*, that is the tendency to conflate the validity of an argument with the confidence placed in its conclusion (Evans et al., 1983). For instance, users might fail to realize that evidence obtained too easily, thanks to the "efficiency" of LLMs in supporting a cherished hypothesis, is not as comprehensive or conclusive with respect to the hypothesis in question as it may seem. Another risk is *overestimating redundant information*: without full control over the sources LLMs draw from, users may overlook redundancy and mistakenly believe they are gaining new evidence to support a particular belief or prediction, when in fact they are not (Bohren, 2016). Similarly, interactions between individuals and LLMs might be susceptible to the so-called *anchoring* to initial hypotheses or inquiries (Tversky and Kahneman, 1974), as well as to *order effects* (Hogarth and Einhorn, 1992). These biases refer, respectively, to the tendency to rely excessively on reference points (even if irrelevant) when making estimates, and to assign greater importance to, or better recall, the first or last pieces of information encountered, at the expense of less available content.

Ex-post evaluation of interactions between human reasoners and LLMs (i.e., the assessment of their interactions after they have taken place) is not immune to errors either. Among the major issues, one cannot help but consider the well-known *hindsight bias*, which is the tendency to perceive events, once they have occurred, as more predictable than they actually were (Arkes, 2013). For instance, individuals might overestimate the accuracy of LLM predictions simply because they overlook how often the original outputs of these models are tentative and inconclusive. Similarly, due to the selective information provided by the models, individuals may underestimate their own initial uncertainties. The concern is that if this misinterpretation of the interaction occurs collaboratively, biases like the one discussed above could be reinforced rather than mitigated.

In conclusion, interactions between the LLM and the user can amplify their inherent weaknesses or even create new ones. This underscores the urgent need for methodological innovations that integrate LLM behaviors with new, interactively designed solutions; without this, they may fail or even backfire.

3 Toward effective human-LLM interaction

In this section, we will first present potential solutions to three major challenges of LLMs: hallucinations, inconsistencies, and sycophancy. We will then discuss how fostering mutual understanding and enhancing complementary team performance are crucial for achieving effective collaboration in reasoning and decision-making between humans and LLMs.

3.1 Detecting and mitigating the impact of hallucinations

Hallucinations are extensively studied in the field of Natural Language Processing (NLP), with various approaches proposed to prevent, detect, or mitigate their occurrence (Huang et al., 2023; Ji et al., 2023a; Rawte et al., 2023; Zhang et al., 2023). Following Huang et al. (2023), we categorize hallucinations into *factuality hallucinations*, where the model generates responses that contradict real-world facts, and *faithfulness hallucinations*, where the model's responses are not aligned with user instructions or the provided context. The latter can be further divided into *intrinsic hallucinations*, involving responses that directly contradict the context, and *extrinsic hallucinations*, in which the generated content cannot be verified or refuted based on the context (Maynez et al., 2020).

One way to improve the factuality of model-generated content is via *retrieval augmented generation* (Lewis et al., 2020), which conditions the generation process on documents retrieved from a corpus such as Wikipedia or Pubmed (Shuster et al., 2021; Xiong et al., 2024; Zakka et al., 2024). However, LLMs can still disregard provided information and rely on their parametric knowledge due to intrinsic mechanisms (Jin et al., 2024; Xu et al., 2024a) or sensitivity to prompts (Liu et al., 2024). Another solution is adapting the generation process—referred to as *decoding*—to produce more factual responses (Lee et al., 2022; Burns et al., 2023; Moschella et al., 2023; Li et al., 2023a; Chuang et al., 2023), and post-editing to refine the originally generated content, leveraging the self-correction capabilities of LLMs (Dhuliawala et al., 2023; Ji et al., 2023b). Decoding can also be adapted to generate outputs that are more faithful to the user instructions or the provided context.

Recent efforts to mitigate faithfulness hallucinations focus on two main areas: *context consistency*, which aims to improve the alignment of model-generated responses with user instructions and the provided context (Tian et al., 2019; van der Poel et al., 2022; Wan et al., 2023; Shi et al., 2023; Gema et al., 2024; Zhao et al., 2024b); and *logical consistency*, which seeks to ensure logically coherent responses in multi-step reasoning tasks (Wang et al., 2023a). Decoding-based methods can be coupled with *post-hoc* hallucination detection approaches (Manakul et al., 2023; Min et al., 2023; Mishra et al., 2024) to define a reward model and adaptively increase the likelihood of hallucination-free generations (Wan et al., 2023; Amini et al., 2024; Lu et al., 2022, 2023; Deng and Raffel, 2023). From the user's perspective, a crucial factor in reducing LLM hallucinations is ensuring that queries are well-constructed, unambiguous, and as specific as possible, since vague or poorly phrased prompts can increase the likelihood of hallucinations (Watson and Cho, 2024).

Although the solutions discussed above can help reduce hallucinations, they will remain, to some extent, inevitable due to the complexity of the world that LLMs attempt to capture (Xu et al., 2024b). A complementary approach is to enhance humans' awareness in managing such occurrences by enabling LLMs to provide uncertainty estimates alongside their outputs. The approaches implemented so far in this line of research fall into three categories (Xiong et al., 2024): *logit-based estimation*, *verbalization-based estimation*, and *consistency-based estimation*. Logit-based estimation requires access to the model logits and typically measures uncertainty by calculating token-level probability or entropy (Guo et al., 2017; Kuhn et al., 2023). Verbalize-based estimation works by directly requesting LLMs to express their uncertainty via prompting strategy (Mielke et al., 2022; Lin et al., 2022a; Xiong et al., 2024; Kadavath et al., 2022). Finally, consistency-based estimation works under the assumption that the most consistent response signifies the least hallucination in the LLM generations (Lin et al., 2023; Chen and Mueller, 2023; Wang et al., 2023b; Zhao et al., 2023). Additionally, recent studies are exploring a new and promising strategy in which LLMs learn to generate citations (Gao et al., 2023; Huang and Chang, 2023). In this way, users can assess the reliability of the outputs provided by LLMs by examining, and potentially directly accessing, their sources.

3.2 Improving robustness

Variability, prompt brittleness, and inconsistencies in LLM outputs across different conditions, domains, and tasks (Gupta et al., 2023; Zhou et al., 2024; Tytarenko and Amin, 2024) pose significant challenges for ensuring effective interaction with humans and can substantially exacerbate their algorithmic aversion. Efforts to enhance the robustness of LLMs have included adjustments during training, as well as *post-hoc* solutions applied after learning has taken place. Regarding the former, recent research has increasingly recognized the value of including domain experts within development teams (e.g., Med-Gemini in healthcare; Saab et al., 2024, FinMA in finance; Xie et al., 2023, and SaulLM; Colombo et al., 2024). Post-training techniques aimed at mitigating prompt sensitivity while preserving performance include *in-context learning adjustments* (Gupta et al., 2023), *task-specific context attribution* (Tytarenko and Amin, 2024), and *batch calibration* (Zhou et al., 2024).

Among the solutions for enhancing LLM robustness are those that directly involve humans, within both perspectives mentioned above. Zhao et al. (2024c) introduced *consistency alignment training* to better align LLM responses with human expectations, fine-tuning LLMs to provide consistent answers to paraphrased instructions Post-training methods involving humans often focus on improving in-context learning examples to be given to the LLMs, by coupling input-output pairs with their corresponding human-generated natural language explanations (He et al., 2024).

Another approach to increasing robustness involves introducing an intermediate step between the user and the model, known as *guardrailing* (Inan et al., 2023; Rebedea et al., 2023), which literally means 'keeping the model on track.' This step evaluates the input and/or output of LLMs to determine if and how certain enforcement actions should be implemented. Common instances include refraining from providing answers that could lead to misuse or blocking responses that contain harmful, inappropriate, or biased content.

3.3 Dealing with sycophancy

Sycophancy is a sort of 'side effect' of the attempt to maximize user satisfaction and the training of LLMs on datasets that include texts generated by humans, where interlocutors often seek to meet each other's expectations. This issue with current LLMs is, of course, not independent of other limitations, but they can exacerbate one another. Indeed, LLMs often hallucinate and become inconsistent in order to appease user prompts, especially when these are misleading. By compelling LLMs not to accommodate these prompts, it could thus lead to a reduction of multiple limitations. On this line, Rrv et al. (2024) showed how popular hallucination mitigation strategies can be effectively used also to reduce the sycophantic behavior of LLMs in factual statement generation.

Other solutions to address sycophancy involve fine-tuning LLMs over aggregated preferences of multiple humans (Sharma et al., 2023), generating synthetic fine-tuning data to change model behavior (Wei et al., 2023) or applying activation editing to steer the internal representations of LLMs toward a less sycophantic direction (Panickssery et al., 2024). To preserve the original capabilities of the LLM as much as possible, Chen et al. (2024) propose *supervised pinpoint tuning*, where fine-tuning is confined to specific LLM modules identified as responsible for the sycophantic behavior.

Finally, Cai et al. (2024) proposed a shift in perspective, termed *antagonistic AI*, a provocative counter-narrative to the prevailing trend of designing AI systems to be agreeable and subservient. According to this approach, human-LLM interactions could benefit from confrontational LLMs that challenge users, even to the point of being blunt if necessary. More specifically, the authors argue that forcing users to confront their own assumptions would, at least in certain situations, promote critical thinking. This intriguing proposal has yet to be implemented or undergo empirical testing. Complementary to this, Tessler et al. (2024) demonstrated that LLMs can assist humans in finding common ground during democratic deliberation by facilitating effective perspective-taking among group members. We believe these approaches could indeed help people identify potential pitfalls in their reasoning and decision-making processes if complemented by cognition-aware interaction strategies to avoid exacerbating algorithmic aversion bias.

3.4 Fostering mutual understanding

The blossoming area of Explainable AI (XAI; Miller, 2018; Gunning et al., 2019; Longo et al., 2024) aims at addressing the problem of explaining the outputs of black-box models to humans, focusing either on single predictions (*local explainability*) or the entire model (*global explainability*). *Explanatory interactive learning* (Teso and Kersting, 2019) builds upon XAI approaches to allow humans to guide machines in learning meaningful predictive patterns while avoiding confounders and shortcuts. However, XAI faces several challenges, from the lack of faithfulness in the generated explanations (Camburu et al., 2019) to the impact of human cognitive biases on evaluating these explanations (Bertrand et al., 2022), including the risk of increasing automation bias (Bansal et al., 2021; Buçinca et al., 2021). The opposite direction—helping machines to understand humans—is equally challenging. Eliciting human knowledge proves inherently difficult, as it is often implicit, incomplete, or incorrect (Patel et al., 1999). Expert judgments, although intuitive, depend on rich mental models that manage incomplete or conflicting information, complicating the representation of this knowledge for machine learning models (Klein et al., 2017; Militello and Anders, 2019).

Compared to other black-box models, LLM-based architectures offer both advantages and disadvantages in terms of mutual understanding. A clear advantage is their use of natural language for communication, enabling conversational sessions where human feedback is integrated into subsequent interactions. However, this natural mode of interaction can be misleading for human partners. Indeed, empirical studies show that users increase their trust in LLM responses when these are accompanied by explanations, even if the responses are deceptive (Sharma et al., 2024). Although various attempts to foster human-LLM alignment through training and interaction strategies have been made (Wang et al., 2023c), LLMs still represent concepts through distributional semantics (Lenci and Sahlgren, 2023), which differs significantly from human semantic understanding (Bender and Koller, 2020). One consequence is that, like many other sub-symbolic machine learning models, LLMs are prone to shortcut learning (Du et al., 2023), a tendency to rely on non-robust features that are spuriously correlated with ground-truth supervision in the training data, yet fail to generalize in out-of-distribution scenarios. XAI approaches are starting to shed light on the reasoning mechanisms of LLMs (Zhao et al., 2024a), but further research is needed for them to produce reliable proxies of the trustworthiness of LLM outputs.

Finally, effective interaction between humans and LLMs requires a form of mutual understanding that involves a theory of mind (ToM; Premack and Woodruff, 1978)—the ability to infer what others are thinking and how this differs from our own thoughts, a crucial precondition for effective communication and cooperation. Recent studies (van Duijn et al., 2023; Kosinski, 2024; Strachan et al., 2024) have shown that larger LLMs, such as GPT-4, made significant progress in ToM, performing on par with, and sometimes even surpassing, humans under certain conditions. However, this competence primarily reflects an ability to simulate human-like responses rather than a genuine mastery of the cognitive processes involved in ToM reasoning. Achieving authentic ToM in LLMs will require further advancements, such as leveraging external memory systems (Li and Qiu, 2023; Schuurmans, 2023) and, eventually, developing machine metacognition (Johnson et al., 2024).

3.5 Targeting complementary team performance

Machine learning methods are typically evaluated in terms of their performance as standalone entities. LLMs are no exceptions to this rule and most research focuses on improving their performance over pre-defined benchmarks (Hendrycks et al., 2021; Liang et al., 2022; Petroni et al., 2021; Chiang et al., 2024). A recent trend has started to question this perspective, advocating for explicit inclusion of the human component in the development and use of these systems (Donahue et al., 2022; Hemmer et al., 2021; Guszcza et al., 2022; Sayin et al., 2023). The notion of *complementary team performance* (CTP; Bansal et al., 2021) has been introduced to evaluate whether team accuracy is higher than either the human or the AI working alone (Hemmer et al., 2021, 2024; Campero et al., 2022). Quite interestingly, studies have shown that human-AI teams can outperform humans but often do not exceed the performance of AI alone (Bansal et al., 2021; Hemmer et al., 2021), highlighting the complexity of achieving good CTP in practice.

Within the machine learning community, researchers have developed *ad hoc* learning strategies to improve CTP. The most popular is *selective classification* (Geifman and El-Yaniv, 2017), where the machine selectively abstains from providing predictions it deems too uncertain. Several selective classification strategies have been proposed in the NLP community, especially in question-answering tasks (Xin et al., 2021; Varshney et al., 2022). A limitation of selective classification is that it does not take into account the characteristics of the person to whom the prediction is deferred. *Learning to defer* (Madras et al., 2018) is an advancement over selective classification, in which human expertise is being modeled and accounted for in choosing when to abstain. *Learning to complement* (Wilder et al., 2021) further extends this line of research by designing a training strategy that directly optimizes team performance. The next challenging yet crucial step will be to adapt these strategies to handle arbitrary users and general-purpose human-LLM reasoning and decision-making tasks.

A major limitation of current solutions for learning to defer/complement is that they rely on a *separation of responsibilities* between the human and the machine. Banerjee et al. (2024) argued that this is suboptimal because it leaves humans completely unassisted in the (presumably difficult) cases where the machine defers, while fostering their automation bias when the machine does not defer. The authors proposed an alternative strategy, *learning to guide*, in which the machine is trained to provide helpful hints to assist the user in making the right decision.

Other promising research directions include adapting strategies that have been developed and proven effective in other areas of AI to LLMs. Among these is *conformal prediction* (Angelopoulos and Bates, 2023), which allows a model to return prediction sets that, according to a user-specified probability, are guaranteed to contain the ground truth. This has been empirically shown to improve human decision-making (Straitouri et al., 2023; Cresswell et al., 2024), and it is beginning to be extended to LLM architectures [*conformal language modeling* (Quach et al., 2024)]. Another approach is *mixed-initiative interaction* (Allen et al., 1999; Barnes et al., 2015), where each agent contributes its strengths to the task, with its level of engagement dynamically adjusted to the specific issue at hand. Recent studies have introduced methods for formalizing prompt construction to enable controllable mixed-initiative dialogue generation (Chen et al., 2023a). Finally, *argumentative decision making* (Amgoud and Prade, 2009) applies argumentation theory to enhance team performance by structuring interactions as sequences of arguments and counter-arguments. Recently, argumentative LLMs (Freedman et al., 2024) have been proposed and tested as a method using LLMs to construct formal argumentation frameworks that support reasoning in decision-making.

4 Conclusion

A human-centered approach to AI has been increasingly promoted by governmental institutions (European Commission, 2020), with legal requirements in many countries mandating human oversight for high-stakes applications (Government of Canada, 2019; European Commission, 2021). Building on this perspective, we have discussed a range of strategies through which the main limitations of current LLMs could be addressed and proposed two fundamental desiderata—mutual understanding and complementary team performance—that, in our view, should guide future research on LLMs and beyond. Indeed, while this manuscript focuses on LLMs due to their widespread adoption, including among lay users, many of the points raised may well apply to multimodal and general-purpose foundation models (Sun et al., 2024) when interacting with humans.

The advocated shift in perspective would require greater involvement of cognitive scientists in shaping approaches to overcome LLM limitations and assess their effectiveness, significantly altering priorities regarding problems and goals for the success of LLMs. Future work could explore new evaluation metrics inspired by cognitive science to better measure the effectiveness of these approaches. Indeed, only by combining the knowledge and exploiting the strengths of both humans and LLMs can we have a real chance to achieve a true partnership—one that is not only more effective in reducing human-machine biases but also more transparent and fair.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

AP: Conceptualization, Funding acquisition, Supervision, Writing - original draft, Writing - review & editing. AG: Conceptualization, Writing - original draft. BS: Conceptualization, Writing - original draft. PM: Supervision, Writing - review & editing, Conceptualization. KT: Conceptualization, Writing - original draft, Writing - review & editing, Supervision.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. Funded by the European Union, Grant Agreement No. 101120763-TANGO. AP acknowledges the support of the MUR PNRR project FAIR-Future AI Research (PE00000013) funded by the NextGenerationEU. AG was supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

References

Akata, Z., Balliet, D., de Rijke, M., Dignum, F., Dignum, V., Eiben, G., et al. (2020). A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer* 53, 18–28. doi: 10.1109/MC.2020.2996587

Allen, J., Guinn, C., and Horvtz, E. (1999). Mixed-initiative interaction. *IEEE Intell. Syst. Their Appl.* 14, 14–23.

Amgoud, L., and Prade, H. (2009). Using arguments for making and explaining decisions. *Artif. Intell.* 173, 413–436. doi: 10.1016/j.artint.2008.11.006

Amini, A., Vieira, T., and Cotterell, R. (2024). Variational best-of-N alignment. *CoRR, abs/2407.06057*. doi: 10.48550/arXiv.2407.06057

Angelopoulos, A. N., and Bates, S. (2023). *Conformal Prediction: A Gentle Introduction*. Norwell, MA: Now Foundations and Trends.

Arkes, H. (2013). The consequences of the hindsight bias in medical decision making. *Curr. Direct. Psychol. Sci.* 22, 356–360. doi: 10.1177/0963721413489988

Bahak, H., Taheri, F., Zojaji, Z., and Kazemi, A. (2023). Evaluating chatGPT as a question answering system: a comprehensive analysis and comparison with existing models. *arXiv, abs/2312.07592*. doi: 10.48550/arXiv.2312.07592

Banerjee, D., Teso, S., Sayin, B., and Passerini, A. (2024). *Learning to Guide Human Decision Makers With Vision-Language Models*.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., et al. (2023). "A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity," in *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds. J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, et al. (Nusa Dua: Association for Computational Linguistics), 675–718.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., et al. (2021). Does the whole exceed its parts? the effect of ai explanations on complementary team performance. *arXiv, abs/2006.14779*. doi: 10.48550/arXiv.2006.14779

Barnes, M. J., Chen, J. Y., and Jentsch, F. (2015). "Designing for mixed-initiative interactions between human and autonomous systems in complex environments," in *2015 IEEE International Conference on Systems, Man, and Cybernetics* (Hong Kong: IEEE), 1386–1390.

Baron, J. (2023). *Thinking and Deciding, 5th Edn*. Cambridge: Cambridge University Press.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). "On the dangers of stochastic parrots: can language models be too big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21* (New York, NY: Association for Computing Machinery), 610–623.

Bender, E. M., and Koller, A. (2020). "Climbing towards NLU: on meaning, form, and understanding in the age of data," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, eds. D. Jurafsky, J. Chai, N. Schluter and J. Tetreault (Association for Computational Linguistics), 5185–5198.

Berberette, E., Hutchins, J., and Sadovnik, A. (2024). Redefining "hallucination" in LLMS: towards a psychology-informed framework for mitigating misinformation. *arXiv, abs/2402.01769*. doi: 10.48550/arXiv.2402.01769

Bertrand, A., Belloum, R., Eagan, J. R., and Maxwell, W. (2022). "How cognitive biases affect XAI-assisted decision-making: a systematic review," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, AIES '22* (New York, NY: Association for Computing Machinery), 78–91.

Bohren, J. A. (2016). Informational herding with model misspecification. *J. Econ. Theor.* 163, 222–247. doi: 10.1016/j.jet.2016.01.011

Boissin, E., Caparos, S., and De Neys, W. (2023). No easy fix for belief bias during syllogistic reasoning? *J. Cogn. Psychol.* 35, 1–21. doi: 10.1080/20445911.2023.2181734

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., et al. (2023). Sparks of artificial general intelligence: early experiments with GPT-4. *arXiv, abs/2303.12712*. doi: 10.48550/arXiv.2303.12712

Buçinca, Z., Malaya, M. B., and Gajos, K. Z. (2021). To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proc. ACM Hum. Comput. Interact.* 5:3449287. doi: 10.1145/3449287

Burns, C., Ye, H., Klein, D., and Steinhardt, J. (2023). *Discovering Latent Knowledge in Language Models Without Supervision*.

Cai, A., Arawjo, I., and Glassman, E. L. (2024). Antagonistic AI. *ArXiv, abs/2402.07350*. doi: 10.48550/arXiv.2402.07350

Camburu, O.-M., Giunchiglia, E., Foerster, J., Lukasiewicz, T., and Blunsom, P. (2019). *Can I Trust the Explainer? Verifying Post-hoc Explanatory Methods*.

Campero, A., Vaccaro, M., Song, J., Wen, H., Almaatouq, A., and Malone, T. W. (2022). A test for evaluating performance in human-computer systems. *arXiv, abs/2206.12390*. doi: 10.48550/arXiv.2206.12390

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., et al. (2023). A survey on evaluation of large language models. *arXiv, abs/2307.03109*. doi: 10.48550/arXiv.2307.03109

Chater, N., and Loewenstein, G. (2023). The I-frame and the S-frame: how focusing on individual-level solutions has led behavioral public policy astray. *Behav. Brain Sci.* 46:e147. doi: 10.1017/s0140525x22002023

Chen, J., and Mueller, J. (2023). Quantifying uncertainty in answers from any language model via intrinsic and extrinsic confidence assessment. *arXiv preprint arXiv:2308.16175*. doi: 10.48550/arXiv.2308.16175

Chen, M., Yu, X., Shi, W., Awasthi, U., and Yu, Z. (2023a). "Controllable mixed-initiative dialogue generation through prompting," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, eds. A. Rogers, J. Boyd-Graber, and N. Okazaki (Toronto, ON: Association for Computational Linguistics), 951–966.

Chen, N., Wang, Y., Jiang, H., Cai, D., Li, Y., Chen, Z., et al. (2023b). "Large language models meet Harry Potter: a dataset for aligning dialogue agents with characters," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, eds. H. Bouamor, J. Pino, and K. Bali (Singapore: Association for Computational Linguistics), 8506–8520.

Chen, W., Huang, Z., Xie, L., Lin, B., Li, H., Lu, L., et al. (2024). "From yes-men to truth-tellers: addressing sycophancy in large language models with pinpoint tuning," in *Forty-first International Conference on Machine Learning*.

Cheong, I., Xia, K., Feng, K. J. K., Chen, Q. Z., and Zhang, A. X. (2024). "(A)I am not a lawyer, but...: engaging legal experts towards responsible LLM policies for legal advice," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24* (New York, NY: Association for Computing Machinery), 2454–2469.

Chiang, W., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., et al. (2024). Chatbot arena: an open platform for evaluating LLMs by human preference. *CoRR, abs/2403.04132*. doi: 10.48550/arXiv.2403.04132

Choudhury, A., and Chaudhry, Z. (2024). Large language models and user trust: consequence of self-referential learning loop and the deskilling of health care professionals. *J. Med. Internet Res.* 26:56764. doi: 10.2196/56764

Chuang, Y., Xie, Y., Luo, H., Kim, Y., Glass, J. R., and He, P. (2023). DoLa: decoding by contrasting layers improves factuality in large language models. *CoRR, abs/2309.03883*. doi: 10.48550/arXiv.2309.03883

Colombo, P., Pires, T. P., Boudiaf, M., Culver, D., Melo, R., Corro, C., et al. (2024). SaulLM-7B: a pioneering large language model for law. *CoRR, abs/2403.03883*. doi: 10.48550/arXiv.2403.03883

Cresswell, J. C., Sui, Y., Kumar, B., and Vouitsis, N. (2024). Conformal prediction sets improve human decision making. *arXiv, 2401.13744*. doi: 10.48550/arXiv.2401.13744

Cummings, M. (2012). "Automation bias in intelligent time critical decision support systems," in *Collection of Technical Papers—AIAA 1st Intelligent Systems Technical Conference*.

Dellermann, D., Ebel, P., Sollner, M., and Leimeister, J. M. (2019). Hybrid intelligence. *Bus. Inform. Syst. Eng.* 61, 637–643. doi: 10.1007/s12599-019-00595-2

Deng, H., and Raffel, C. (2023). "Reward-augmented decoding: efficient controlled text generation with a unidirectional reward model," in *EMNLP* (Association for Computational Linguistics), 11781–11791.

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., et al. (2023). Chain-of-verification reduces hallucination in large language models. *CoRR, abs/2309.11495*. doi: 10.48550/arXiv.2309.11495

Dietvorst, B. J., Simmons, J. P., Massey, C., Dietvorst, B. J., and Simmons, J. (2014). Algorithm aversion: people erroneously avoid algorithms after seeing them ERR. *J. Exp. Psychol.* 144, 114–126. doi: 10.1037/xge0000033

Donahue, K., Chouldechova, A., and Kenthapadi, K. (2022). "Human-algorithm collaboration: achieving complementarity and avoiding unfairness," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22* (New York, NY: Association for Computing Machinery), 1639–1656.

Du, M., He, F., Zou, N., Tao, D., and Hu, X. (2023). Shortcut learning of large language models in natural language understanding. *Commun. ACM* 67, 110—120. doi: 10.1145/3596490

Eigner, E., and Händler, T. (2024). Determinants of LLM-assisted decision-making. *arXiv, abs/2402.17385*. doi: 10.48550/arXiv.2402.17385

European Commission (2020). *White Paper on Artificial Intelligence: a European Approach to Excellence and Trust. White Paper COM(2020) 65 Final*. Brussels: European Commission.

European Commission (2021). *Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*.

Evans, J., Barston, J., and Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Mem. Cogn.* 11, 295–306.

Felin, T. and Holweg, M. (2024). *Theory Is All You Need: AI, Human Cognition, and Decision Making*.

Freedman, G., Dejl, A., Gorur, D., Yin, X., Rago, A., and Toni, F. (2024). Argumentative large language models for explainable and contestable decision-making. *arXiv, 2405.02079*. doi: 10.48550/arXiv.2405.02079

Gao, T., Yen, H., Yu, J., and Chen, D. (2023). "Enabling large language models to generate text with citations," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, eds. H. Bouamor, J. Pino, and K. Bali (Singapore: Association for Computational Linguistics), 6465–6488.

Geifman, Y., and El-Yaniv, R. (2017). "Selective classification for deep neural networks," in *Advances in Neural Information Processing Systems, Vol. 30*, eds. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (Red Hook, NY: Curran Associates, Inc).

Gema, A. P., Jin, C., Abdulaal, A., Diethe, T., Teare, P., Alex, B., et al. (2024). DeCoRe: decoding by contrasting retrieval heads to mitigate hallucinations. *arXiv preprint arXiv:2410.18860*. doi: 10.48550/arXiv.2410.18860

Government of Canada (2019). *Directive on Automated Decision-Making*. Available at: https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., and Yang, G.-Z. (2019). XAI–explainable artificial intelligence. *Sci. Robot*. 4:eaay7120. doi: 10.1126/scirobotics.aay7120

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). "On calibration of modern neural networks," in *International Conference on Machine Learning* (Sydney), 1321–1330.

Gupta, K., Roychowdhury, S., Kasa, S. R., Kasa, S. K., Bhanushali, A., Pattisapu, N., et al. (2023). How robust are LLMs to in-context majority label bias? *arXiv, abs/2312.16549*. doi: 10.48550/arXiv.2312.16549

Guszcza, J., Danks, D., Fox, C., Hammond, K., Ho, D., Imas, A., et al. (2022). Hybrid intelligence: a paradigm for more responsible practice. *SSRN Electr. J*. 2022:4301478. doi: 10.2139/ssrn.4301478

He, X., Wu, Y., Camburu, O.-M., Minervini, P., and Stenetorp, P. (2024). "Using natural language explanations to improve robustness of in-context learning," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds. L.-W. Ku, A. Martins, and V. Srikumar (Bangkok: Association for Computational Linguistics), 13477–13499.

Hemmer, P., Schemmer, M., Kühl, N., Vössing, M., and Satzger, G. (2024). Complementarity in human-ai collaboration: concept, sources, and evidence. *arXiv, abs/2404.00029*. doi: 10.48550/arXiv.2404.00029

Hemmer, P., Schemmer, M., Vössing, M., and Kühl, N. (2021). "Human-AI complementarity in hybrid intelligence systems: a structured literature review," in *Pacific Asia Conference on Information Systems* (Dubai).

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., et al. (2021). *Measuring Massive Multitask Language Understanding*. Bangkok: Association for Computational Linguistics.

Hoffrage, U. (2022). "Overconfidence," in *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory, 3 Edn* (Routledge/Taylor & Francis Group), 291–314.

Hogarth, R. M., and Einhorn, H. J. (1992). Order effects in belief updating: the belief-adjustment model. *Cogn. Psychol*. 24, 1–55.

Huang, J., and Chang, K. C.-C. (2023). Citation: a key to building responsible and accountable large language models. *arXiv preprint arXiv:2307.02185*. doi: 10.48550/arXiv.2307.02185

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., et al. (2023). A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *arXiv, abs/2311.05232*. doi: 10.48550/arXiv.2311.05232

Huang, S.-H., Lin, Y.-F., He, Z., Huang, C.-Y., and Huang, T.-H. K. (2024). "How does conversation length impact user's satisfaction? a case study of length-controlled conversations with LLM-powered chatbots," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA '24*, New York, NY: Association for Computing Machinery.

Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., et al. (2023). Llama guard: LLM-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*. doi: 10.48550/arXiv.2312.06674

Jeblick, K., Schachtner, B., Dexl, J., Mittermeier, A., Stüber, A. T., Topalis, J., et al. (2022). ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *arXiv, abs/2212.14882*. doi: 10.1007/s00330-023-10213-1

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., et al. (2023a). Survey of hallucination in natural language generation. *ACM Comput. Surv*. 55:248. doi: 10.1145/3571730

Ji, Z., Yu, T., Xu, Y., Lee, N., Ishii, E., and Fung, P. (2023b). "Towards mitigating LLM hallucination via self reflection," in *EMNLP (Findings)* (Singapore: Association for Computational Linguistics), 1827–1843.

Jiao, W., Wang, W., tse Huang, J., Wang, X., Shi, S., and Tu, Z. (2023). *Is ChatGPT a Good Translator? Yes With GPT-4 as the Engine*.

Jin, Z., Cao, P., Chen, Y., Liu, K., Jiang, X., Xu, J., et al. (2024). Tug-of-war between knowledge: exploring and resolving knowledge conflicts in retrieval-augmented language models. *arXiv preprint arXiv:2402.14409*. doi: 10.48550/arXiv.2402.14409

Johnson, S. G. B., Karimi, A.-H., Bengio, Y., Chater, N., Gerstenberg, T., Larson, K., et al. (2024). Imagining and building wise machines: the centrality of AI metacognition. *arXiv, 2411.02478*. doi: 10.48550/arXiv.2411.02478

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., et al. (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*. doi: 10.48550/arXiv.2207.05221

Khan, A., Breslav, S., Glueck, M., and Hornbæk, K. (2015). Benefits of visualization in the mammography problem. *Int. J. Hum. Comput. Stud*. 83, 94–113. doi: 10.1016/j.ijhcs.2015.07.001

Klein, G., Shneiderman, B., Hoffman, R. R., and Ford, K. M. (2017). Why expertise matters: a response to the challenges. *IEEE Intell. Syst*. 32, 67–73. doi: 10.1109/MIS.2017.4531230

Kosinski, M. (2024). *Evaluating Large Language Models in Theory of Mind Tasks*.

Kuhn, L., Gal, Y., and Farquhar, S. (2023). "Semantic uncertainty: linguistic invariances for uncertainty estimation in natural language generation," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1–5, 2023*. Kigali.

Lawrence, C. (2024). *Human-centric AI: A Road Map to Human-AI Collaboration*. Available at: https://neclab.eu/technology/case-studies/human-centricity-ai-a-road-map-to-human-ai-collaboration

Lee, N., Ping, W., Xu, P., Patwary, M., Shoeybi, M., and Catanzaro, B. (2022). Factuality enhanced language models for open-ended text generation. *CoRR, abs/2206.04624*. doi: 10.48550/arXiv.2206.04624

Leivada, E., Dentella, V., and Günther, F. (2024). Evaluating the language abilities of large language models vs. humans: three caveats. *Biolinguistics*. doi: 10.5964/bioling.14391

Lenci, A., and Sahlgren, M. (2023). *Distributional Semantics. Studies in Natural Language Processing*. Cambridge: Cambridge University Press.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv. Neural Inform. Process. Syst*. 33, 9459–9474. doi: 10.48550/arXiv.2005.11401

Li, K., Patel, O., Vi'egas, F., Pfister, H.-R., and Wattenberg, M. (2023a). Inference-time intervention: eliciting truthful answers from a language model. *arXiv, abs/2306.03341*. doi: 10.48550/arXiv.2306.03341

Li, X., and Qiu, X. (2023). "MoT: Memory-of-thought enables ChatGPT to self-improve," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, eds. H. Bouamor, J. Pino, and K. Bali (Singapore: Association for Computational Linguistics), 6354–6374.

Li, Z., Zhang, S., Zhao, H., Yang, Y., and Yang, D. (2023b). BatGPT: a bidirectional autoregessive talker from generative pre-trained transformer. *arXiv, 2307.00360*. doi: 10.48550/arXiv.2307.00360

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., et al. (2022). Holistic evaluation of language models. *CoRR, abs/2211.09110*. doi: 10.48550/arXiv.2211.09110

Liao, Q. V., and Wortman Vaughan, J. (2024). *AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap*. Harvard Data Science Review. Available at: https://hdsr.mitpress.mit.edu/pub/aelql9qy (accessed December 14, 2024).

Lin, S., Hilton, J., and Evans, O. (2022a). Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*. doi: 10.48550/arXiv.2205.14334

Lin, S., Hilton, J., and Evans, O. (2022b). "TruthfulQA: measuring how models mimic human falsehoods," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds. S. Muresan, P. Nakov, and A. Villavicencio (Dublin: Association for Computational Linguistics), 3214–3252.

Lin, Z., Trivedi, S., and Sun, J. (2023). Generating with confidence: uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*. doi: 10.48550/arXiv.2305.19187

Liu, B., Ash, J. T., Goel, S., Krishnamurthy, A., and Zhang, C. (2023a). Exposing attention glitches with flip-flop language modeling. *arXiv, 2306.00946*. doi: 10.48550/arXiv.2306.00946

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., et al. (2024). Lost in the middle: how language models use long contexts. *Trans. Assoc. Comput. Linguist*. 12, 157–173. doi: 10.1162/tacl_a_00638

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., et al. (2023b). Summary of chatGPT-related research and perspective towards the future of large language models. *Meta-Radiology* 1:100017. doi: 10.1016/j.metrad.2023.100017

Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Ser, J. D., et al. (2024). Explainable artificial intelligence (XAI) 2.0: a manifesto of open challenges and interdisciplinary research directions. *Inform. Fus*. 106:102301. doi: 10.1016/j.inffus.2024.102301

Lu, X., Brahman, F., West, P., Jung, J., Chandu, K., Ravichander, A., et al. (2023). "Inference-time policy adapters (IPA): tailoring extreme-scale LMS without fine-tuning," in *EMNLP* (Singapore: Association for Computational Linguistics), 6863–6883.

Lu, X., Welleck, S., West, P., Jiang, L., Kasai, J., Khashabi, D., et al. (2022). "Neurologic a*esque decoding: Constrained text generation with lookahead heuristics," in *NAACL-HLT* (Seattle: Association for Computational Linguistics), 780–799.

Ma, S., Chen, Q., Wang, X., Zheng, C., Peng, Z., Yin, M., et al. (2024). Towards human-AI deliberation: Design and evaluation of LLM-empowered deliberative AI for AI-assisted decision-making. *arXiv, 2403.16812*. doi: 10.48550/arXiv.2403.16812

Madras, D., Pitassi, T., and Zemel, R. (2018). "Predict responsibly: improving fairness and accuracy by learning to defer," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18* (Red Hook, NY: Curran Associates Inc), 6150–6160.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends Cogn. Sci*. 28, 517–540. doi: 10.1016/j.tics.2024.01.011

Manakul, P., Liusie, A., and Gales, M. (2023). "SelfCheckGPT: zero-resource black-box hallucination detection for generative large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, eds. H.

Bouamor, J. Pino, and K. Bali (Singapore: Association for Computational Linguistics), 9004–9017.

Mao, J., Middleton, S., and Niranjan, M. (2024). "Do prompt positions really matter?" in *Findings of the Association for Computational Linguistics: NAACL 2024*, eds. K. Duh, H. Gomez, S. Bethard (Mexico City: Association for Computational Linguistics), 4102–4130.

Mastropasqua, T., Crupi, V., and Tentori, K. (2010). Broadening the study of inductive reasoning: confirmation judgments with uncertain evidence. *Mem. Cogn.* 38, 941–950. doi: 10.3758/MC.38.7.941

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. T. (2020). "On faithfulness and factuality in abstractive summarization," in *ACL*, eds. D. Jurafsky, J. Chai, N. Schluter, Joel Tetreault (Association for Computational Linguistics), 1906–1919.

McGrath, M. J., Cooper, P. S., and Duenser, A. (2024). Users do not trust recommendations from a large language model more than AI-sourced snippets. *Front. Comput. Sci.* 6:1456098. doi: 10.3389/fcomp.2024.1456098

Mielke, S. J., Szlam, A., Dinan, E., and Boureau, Y.-L. (2022). Reducing conversational agents' overconfidence through linguistic calibration. *Trans. Assoc. Comput. Linguist.* 10, 857–872. doi: 10.1162/tacl_a_00494

Militello, L. G., and Anders, S. H. (2019). "Incident-based methods for studying expertise," in *The Oxford Handbook of Expertise* (Oxford Academic), 429–450.

Miller, T. (2018). Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell.* 267, 1–38. doi: 10.1016/j.artint.2018.07.007

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P., et al. (2023). "FActScore: fine-grained atomic evaluation of factual precision in long form text generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, eds. H. Bouamor, J. Pino, and K. Bali (Singapore: Association for Computational Linguistics), 12076–12100.

Mishra, A., Asai, A., Balachandran, V., Wang, Y., Neubig, G., Tsvetkov, Y., et al. (2024). Fine-grained hallucination detection and editing for language models. *arXiv, abs/2401.06855*. doi: 10.48550/arXiv.2401.06855

Moschella, L., Maiorca, V., Fumero, M., Norelli, A., Locatello, F., and Rodolá, E. (2023). *Relative Representations Enable Zero-Shot Latent Space Communication*. Kigali: ICLR.

Nickerson, R. S. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* 2, 175–220.

Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. M. (2024). *Steering Llama 2 via Contrastive Activation Addition*.

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., and Hendrycks, D. (2023). AI deception: a survey of examples, risks, and potential solutions. *arXiv, abs/2308.14752*. doi: 10.48550/arXiv.2308.14752

Patel, V. L., Arocha, J. F., and Kaufman, D. R. (1999). *Expertise and Tacit Knowledge in Medicine. Tacit Knowledge in Professional Practice*. London: Psychology Press, 75–99.

Petroni, F., Piktus, A., Fan, A., Lewis, P. S. H., Yazdani, M., Cao, N. D., et al. (2021). "KILT: a benchmark for knowledge intensive language tasks," in *NAACL-HLT*, eds. K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Association for Computational Linguistics), 2523–2544.

Pezeshkpour, P., and Hruschka, E. (2023). Large language models sensitivity to the order of options in multiple-choice questions. *CoRR, abs/2308.11483*. doi: 10.48550/arXiv.2308.11483

Premack, D., and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* 1, 515–526.

Pu, D., and Demberg, V. (2023). "ChatGPT vs. human-authored text: insights into controllable text summarization and sentence style transfer," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, eds. V. Padmakumar, G. Vallejo, and Y. Fu (Toronto, ON: Association for Computational Linguistics), 1–18.

Quach, V., Fisch, A., Schuster, T., Yala, A., Sohn, J. H., Jaakkola, T., et al. (2024). Conformal language modeling. *arXiv, abs/2306.10193*. doi: 10.48550/arXiv.2306.10193

Ranaldi, L., and Pucci, G. (2023). When large language models contradict humans? large language models' sycophantic behaviour. *arXiv, abs/2311.09410*. doi: 10.48550/arXiv.2311.09410

Rawte, V., Sheth, A. P., and Das, A. (2023). A survey of hallucination in large foundation models. *CoRR, abs/2309.05922*. doi: 10.48550/arXiv.2309.05922

Rebedea, T., Dinu, R., Sreedhar, M., Parisien, C., and Cohen, J. (2023). NeMo guardrails: a toolkit for controllable and safe LLM applications with programmable rails. *arXiv preprint arXiv:2310.10501*.doi: 10.48550/arXiv.2310.10501

Rrv, A., Tyagi, N., Uddin, M. N., Varshney, N., and Baral, C. (2024). Chaos with keywords: exposing large language models sycophancy to misleading keywords and evaluating defense strategies. *arXiv, abs/2406.03827*. doi: 10.48550/arXiv.2406.03827

Saab, K., Tu, T., Weng, W.-H., Tanno, R., Stutz, D., Wulczyn, E., et al. (2024). Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*. doi: 10.48550/arXiv.2404.18416

Sayin, B., Minervini, P., Staiano, J., and Passerini, A. (2024). "Can LLMs correct physicians, yet? Investigating effective interaction methods in the medical domain," in *Proceedings of the 6th Clinical Natural Language Processing Workshop* (Mexico City: Association for Computational Linguistics), 218–237.

Sayin, B., Yang, J., Passerini, A., and Casati, F. (2023). Value-based hybrid intelligence. *Front. Artif. Intell. Appl.* 368, 366–370. doi: 10.3233/FAIA230100

Schuurmans, D. (2023). *Memory Augmented Large Language Models Are Computationally Universal*.

Sharma, M., Siu, H. C., Paleja, R., and Peña, J. D. (2024). *Why Would You Suggest That? Human Trust in Language Model Responses*.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D. K., Askell, A., Bowman, S. R., et al. (2023). Towards understanding sycophancy in language models. *arXiv, abs/2310.13548*. doi: 10.48550/arXiv.2310.13548

Shen, T., Jin, R., Huang, Y., Liu, C., Dong, W., Guo, Z., et al. (2023). Large language model alignment: a survey. *arXiv, 2309.15025*. doi: 10.48550/arXiv.2309.15025

Shi, W., Han, X., Lewis, M., Tsvetkov, Y., Zettlemoyer, L., and Yih, S. (2023). Trusting your evidence: hallucinate less with context-aware decoding. *arXiv, abs/2305.14739*. doi: 10.48550/arXiv.2305.14739

Shuster, K., Poff, S., Chen, M., Kiela, D., and Weston, J. (2021). "Retrieval augmentation reduces hallucination in conversation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, eds. M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih (Punta Cana: Association for Computational Linguistics), 3784–3803.

Strachan, J., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., et al. (2024). Testing theory of mind in large language models and humans. *Nat. Hum. Behav.* 24:1882. doi: 10.1038/s41562-024-01882-z

Straitouri, E., Wang, L., Okati, N., and Rodriguez, M. G. (2023). "Improving expert predictions with conformal prediction," in *Proceedings of the 40th International Conference on Machine Learning, ICML'23* (Honolulu, HI).

Su, W., Wang, C., Ai, Q., Hu, Y., Wu, Z., Zhou, Y., et al. (2024). Unsupervised real-time hallucination detection based on the internal states of large language models. *arXiv, abs/2403.06448*. doi: 10.48550/arXiv.2403.06448

Sun, J., Zheng, C., Xie, E., Liu, Z., Chu, R., Qiu, J., et al. (2024). *A Survey of Reasoning With Foundation Models*.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: atatistics, structure, and abstraction. *Science* 331, 1279–1285. doi: 10.1126/science.1192788

Tentori, K., Chater, N., and Crupi, V. (2016). Judging the probability of hypotheses versus the impact of evidence: which form of inductive inference is more accurate and time-consistent? *Cogn. Sci.* 40, 758–778. doi: 10.1111/cogs.12259

Teso, S., and Kersting, K. (2019). "Explanatory interactive machine learning," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES '19* (New York, NY: Association for Computing Machinery), 239–245.

Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Koster, R., et al. (2024). AI can help humans find common ground in democratic deliberation. *Science* 386:eadq2852. doi: 10.1126/science.adq2852

Thaler, R., and Sunstein, C. (2009). Nudge: improving decisions about health, wealth, and happiness. *J. Cogn. Psychol.* 35, 401–421. doi: 10.1007/s10602-008-9056-2

Tian, R., Narayan, S., Sellam, T., and Parikh, A. P. (2019). Sticking to the facts: confident decoding for faithful data-to-text generation. *CoRR, abs/1910.08684*. doi: 10.48550/arXiv.1910.08684

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–1131.

Tytarenko, S., and Amin, M. R. (2024). Breaking free transformer models: task-specific context attribution promises improved generalizability without fine-tuning pre-trained LLMs. *arXiv, abs/2401.16638*. doi: 10.48550/arXiv.2401.16638

van der Poel, L., Cotterell, R., and Meister, C. (2022). "Mutual information alleviates hallucinations in abstractive summarization," in *EMNLP*, eds. Y. Goldberg, Z. Kozareva, Y. Zhang (Abu Dhabi: Association for Computational Linguistics), 5956–5965.

van Duijn, M., van Dijk, B., Kouwenhoven, T., de Valk, W., Spruit, M., and van der Putten, P. (2023). "Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7–10 on advanced tests," in *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, eds. J. Jiang, D. Reitter, and S. Deng (Singapore: Association for Computational Linguistics), 389–402.

Varshney, N., Mishra, S., and Baral, C. (2022). "Towards improving selective prediction ability of NLP systems," in *Proceedings of the 7th Workshop on Representation Learning for NLP*, eds. S. Gella, H. He, B. P. Majumder, B. Can, E. Giunchiglia, S. Cahyawijaya, et al. (Dublin: Association for Computational Linguistics), 221–226.

Voronov, A., Wolf, L., and Ryabinin, M. (2024). Mind your format: towards consistent evaluation of in-context learning improvements. *CoRR, abs/2401.06766*. doi: 10.48550/arXiv.2401.06766

Wan, D., Liu, M., McKeown, K. R., Dreyer, M., and Bansal, M. (2023). "Faithfulness-aware decoding strategies for abstractive summarization," in *EACL*, eds. A. Vlachos, I. Augenstein (Dubrovnik: Association for Computational Linguistics), 2856–2872.

Wang, J., Ma, W., Sun, P., Zhang, M., and Nie, J.-Y. (2024). Understanding user experience in large language model interactions. *arXiv, abs/2401.08329*. doi: 10.48550/arXiv.2401.08329

Wang, P., Wang, Z., Li, Z., Gao, Y., Yin, B., and Ren, X. (2023a). "SCOTT: self-consistent chain-of-thought distillation," in *ACL (1)*, eds A. Rogers, J. Boyd-Graber, N. Okazaki (Toronto: Association for Computational Linguistics), 5546–5558.

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., et al. (2023b). "Self-consistency improves chain of thought reasoning in language models," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1–5, 2023*. Kigali.

Wang, Y., Zhong, W., Li, L., Mi, F., Zeng, X., Huang, W., et al. (2023c). *Aligning Large Language Models With Human: A Survey*.

Watson, W., and Cho, N. (2024). Hallucibot: is there no such thing as a bad question? *arXiv preprint arXiv:2404.12535*. doi: 10.48550/arXiv.2404.12535

Watters, C., and Lemanski, M. K. (2023). Universal skepticism of chatgpt: a review of early literature on chat generative pre-trained transformer. *Front. Big Data* 6. doi: 10.3389/fdata.2023.1224976

Wei, J. W., Huang, D., Lu, Y., Zhou, D., and Le, Q. V. (2023). Simple synthetic data reduces sycophancy in large language models. *arXiv, abs/2308.03958*. doi: 10.48550/arXiv.2308.03958

Wilder, B., Horvitz, E., and Kamar, E. (2021). "Learning to complement humans," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI'20* (Yokohama).

Xie, Q., Han, W., Zhang, X., Lai, Y., Peng, M., Lopez-Lira, A., et al. (2023). "PIXIU: a comprehensive benchmark, instruction dataset and large language model for finance," in *Advances in Neural Information Processing Systems, Volume 36*, eds. A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Red Hook, NY: Curran Associates, Inc.), 33469–33484.

Xin, J., Tang, R., Yu, Y., and Lin, J. (2021). "The art of abstention: selective prediction and error regularization for natural language processing," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, eds. C. Zong, F. Xia, W. Li, and R. Navigli (Association for Computational Linguistics), 1040–1051.

Xiong, G., Jin, Q., Lu, Z., and Zhang, A. (2024). Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*. doi: 10.48550/arXiv.2402.13178

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., et al. (2024). "Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs," in *The Twelfth International Conference on Learning Representations*.

Xu, R., Qi, Z., Wang, C., Wang, H., Zhang, Y., and Xu, W. (2024a). Knowledge conflicts for LLMs: a survey. *arXiv preprint arXiv:2403.08319*. doi: 10.48550/arXiv.2403.08319

Xu, Z., Jain, S., and Kankanhalli, M. (2024b). Hallucination is inevitable: an innate limitation of large language models. *arXiv*. doi: 10.48550/arXiv.2401.11817

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., et al. (2024). Harnessing the power of LLMs in practice: a survey on chatGPT and beyond. *ACM Trans. Knowl. Discov. Data* 18:3649506. doi: 10.1145/3649506

Zakka, C., Shad, R., Chaurasia, A., Dalal, A. R., Kim, J. L., Moor, M., et al. (2024). Almanac–retrieval-augmented language models for clinical medicine. *NEJM AI* 1:AIoa2300068. doi: 10.1056/AIoa2300068

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., et al. (2023). Siren's song in the AI ocean: a survey on hallucination in large language models. *CoRR, abs/2309.01219*. doi: 10.48550/arXiv.2309.01219

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., et al. (2024a). Explainability for large language models: a survey. *ACM Trans. Intell. Syst. Technol.* 15:3639372. doi: 10.1145/3639372

Zhao, R., Li, X., Joty, S., Qin, C., and Bing, L. (2023). "Verify-and-edit: a knowledge-enhanced chain-of-thought framework," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds. A. Rogers, J. Boyd-Graber, and N. Okazaki (Toronto, ON: Association for Computational Linguistics), 5823–5840.

Zhao, Y., Devoto, A., Hong, G., Du, X., Gema, A. P., Wang, H., et al. (2024b). Steering knowledge selection behaviours in LLMs via sae-based representation engineering. *arXiv preprint arXiv:2410.15999*. doi: 10.48550/arXiv.2410.15999

Zhao, Y., Yan, L., Sun, W., Xing, G., Wang, S., Meng, C., et al. (2024c). Improving the robustness of large language models via consistency alignment. *arXiv, abs/2403.14221*. doi: 10.48550/arXiv.2403.14221

Zhou, H., Wan, X., Proleev, L., Mincu, D., Chen, J., Heller, K., et al. (2024). Batch calibration: rethinking calibration for in-context learning and prompt engineering. *arXiv, abs/2309.17249*. doi: 10.48550/arXiv.2309.17249