Challenges in Guardrailing Large Language Models for Science

Nishan Pantha[1], Muthukumaran Ramasubramanian[1], Iksha Gurung[1], Manil Maskey[2], Rahul Ramachandran[2]

[1]Earth Systems Science Center, University of Alabama in Huntsville, Huntsville, AL 35801, USA
[2]Marshall Space Flight Center, NASA, Huntsville, AL 35899, USA

Abstract

The rapid development in large language models (LLMs) has transformed the landscape of natural language processing and understanding (NLP/NLU), offering significant benefits across various domains. However, when applied to scientific research, these powerful models exhibit critical failure modes related to scientific integrity and trustworthiness. Existing general-purpose LLM guardrails are insufficient to address these unique challenges in the scientific domain. We propose a comprehensive taxonomic framework for LLM guardrails encompassing four key dimensions: trustworthiness, ethics & bias, safety, and legal compliance. Our framework includes structured implementation guidelines for scientific research applications, incorporating white-box, black-box, and gray-box methodologies. This approach specifically addresses critical challenges in scientific LLM deployment, including temporal sensitivity, knowledge contextualization, conflict resolution, and intellectual property protection.

Index Terms

NLP, NLU, LLM, AI

I. Introduction

The advent of large language models (LLMs) has revolutionized the field of natural language processing and understanding (NLP/NLU) [1] [2] [3], leading to numerous applications, particularly in chat systems. Users interact with these systems in both open-ended and close-ended question-answering (QA) modes, leveraging the models' capabilities to generate human-like responses and perform complex language tasks. However, deploying LLMs in real-world applications introduces safety, ethics, and reliability challenges. As a result, extensive research has focused on incorporating LLM guardrails to ensure responsible use and prevent failure modes [4], [5].

LLM guardrails are mechanisms designed to enforce safety and various standards in LLM applications by monitoring and controlling user interactions [6]. These guardrails operate through both intrinsic model-level constraints and explicit rule-based systems, ensuring LLMs function within predefined principles while validating response structure, type, and quality. This is crucial due to the inherent unpredictability of LLMs, which can generate biased, misleading, or harmful outputs. Effective governance and safety measures are essential to maintain trust in generative AI technologies like these, as they become more integrated into daily applications. Some critical dimensions and properties of LLM guardrails, applicable to a wide range of domains [6], include (but are not limited to):

- Mitigating factual hallucinations
- Ensuring fairness in data handling
- Enforcing data privacy, confidentiality, and regulatory standards
- Enhancing model robustness against adversarial attacks
- Detecting and filtering toxic content
- Complying with legal and ethical standards
- Identifying and Handling out-of-distribution inputs and outputs.
- Accurately quantifying and communicating output uncertainty

These properties become even more critical in sensitive domains, such as scientific research. High standards of factual accuracy and adherence to content moderation are essential to prevent the generation of inappropriate or misleading responses. The scientific field demands precision, as even minor inaccuracies or biases can have significant consequences [7], [8], from misleading research directions to compromising experimental reproducibility, affecting public trust and the advancement of knowledge [9]. Therefore, developing and implementing systematic techniques to evaluate, analyze, and enhance the performance of LLM guardrails is crucial in these contexts.

Recent years have also seen rapid developments in LLMs and AI systems such as GPT-4 [10], Llama-3 [11], Claude [12], Mistral [13], and Gemini [14] transforming the landscape of the scientific domain, exhibiting remarkable capabilities in scientific knowledge processing and data discovery, content generation, and data assimilation at scale [15], [16]. These models also have the potential to significantly enhance scientific workflows, from accelerating literature reviews and automating data analysis to aiding in the writing and synthesis of research findings, redefining the bounds of knowledge consumption. [17].

However, their positive impact depends on the reliability and accuracy of their outputs. Given the sensitive nature of scientific inquiry, these outputs need to adhere to strict standards.

Establishing comprehensive guidelines for deploying LLM guardrails in scientific research is therefore crucial. These guidelines should aim to safeguard the integrity of scientific processes, ensuring that the information generated by LLMs is factually accurate, consistent, and aligned with established scientific principles and values. Moreover, the ethical implications of LLM use in science must be thoroughly examined and addressed. Moreover, ethical considerations such as the potential for amplifying biases, privacy concerns, and societal impacts must be carefully addressed [18], [19]. Effective guidelines should incorporate mechanisms to identify and mitigate these ethical risks, ensuring that the deployment of LLMs in scientific research upholds the highest standards of scientific integrity, fairness, and social responsibility.

This paper explores the key categories and dimensions of LLM guardrails essential for scientific research, including aspects like scientific integrity and biases. Implementing these safeguards requires developing specialized evaluation, analysis, and enhancement techniques. Our aim is to prevent harmful content generation and ensure that LLMs serve as reliable, ethically sound tools in scientific inquiry. By doing so, we seek to utilize LLMs while preserving the accuracy fundamental to the scientific community. In the following sections, we present an overview of aspects of existing LLM guardrail frameworks, emphasizing their roles in supporting trustworthy AI applications in scientific environments. This overview sets the stage for a deeper examination of specific dimensions and challenges in deploying LLM guardrails for scientific applications.

II. Overview of LLM Guardrails

Implementing LLM guardrails is crucial to minimize risks, especially when these models are applied in sensitive areas like scientific research. Such guardrails help ensure that LLMs operate within established standards, addressing safety, ethical use, and reliability concerns. In high-stakes fields like science, these measures are vital for maintaining trust and ensuring the responsible use of AI technologies

A. General Dimensions for LLM Guardrails

Recent surveys, including those by Dong et al. (2024) [6], have highlighted various dimensions for developing effective LLM guardrails. Dong et al. (2024) emphasize the importance of black-box and post-hoc strategies, which involve continuously monitoring and filtering LLM inputs and outputs to safeguard against potential failures. Frameworks such as Llama Guard [20], Nvidia NeMo [21], LMQL [22], Guidance [23], and Guardrails AI [24] are examples of systems designed to enforce these safety measures, ensuring that LLMs operate within established guidelines and standards across various applications.

For example, Llama Guard [20] focuses on enhancing human-AI conversation safety by classifying outputs based on user-defined categories. Nvidia NeMo [21] employs a more formal approach, using sentence transformers to guide LLMs within strict dialogical boundaries. On the other hand, Guardrails AI provides structure, type, and quality guarantees for LLM outputs, though its applicability is limited to text-based scenarios. The table I summarizes different properties of these guardrails illustrating their respective strengths and limitations in addressing different aspects of LLM guardrails such as safety, fairness, privacy, robustness, and legality.

TABLE I: Abilities among different Guardrails, Dong *et al.* [6]

	Llama Guard	Nvidia NeMo Guardrails	AI TruLens	Guidance AI	LMQL	
Hallucination	✓	✓	✓	✓	✓	✓
Fairness	✓	-	✓	✓	-	-
Privacy	-	✓	-	-	-	-
Robustness	-	-	-	-	-	-
Toxicity	✓	✓	✓	✓	✓	✓
Legality	✓	-	-	-	-	-
Out-of-Distribution	-	-	✓	-	-	-
Uncertainty	-	✓	✓	✓	-	-

Scientific research poses distinct challenges for each of these guardrail aspects. To better understand these challenges, we explore critical dimensions that are particularly relevant to scientific inquiry, such as the dangers of hallucination, the importance of fairness and robustness, and the need for privacy.

1) Hallucination

Hallucinations refer to instances where the LLM generates factually incorrect or nonsensical outputs. This issue is particularly concerning in sensitive applications, such as scientific research, where accuracy is crucial. For example, [25] notes that in clinical settings even minor hallucinations – such as adding symptoms like ”fever” inaccurately to a patient summary – can reinforce diagnostic biases and misguide clinical decisions. Similarly, Huang et al. [26] examine the broader landscape of hallucinations in LLMs, presenting a detailed taxonomy of contributing factors such as inherent pre-training biases and flawed decoding strategies that can affect reliability in real-world applications.Researchers have explored various approaches to safeguard LLMs against these pitfalls. Techniques like formal verification and adversarial training have been proposed to embed safety checks within the model's architecture, ensuring that outputs adhere to rigorous standards of factual accuracy [27] [6]. Additionally, methods such as ChainPoll [28], SelfCheckGPT [29], GPTScore [30], and G-Eval [31] have been developed to detect and mitigate hallucinations effectively.

2) Fairness

Fairness in LLMs involves ensuring that the outputs are unbiased and do not perpetuate harmful stereotypes or discrimination. This is particularly challenging given that LLMs are often trained on extensive and diverse datasets encompassing multiple languages, cultures, and ideologies, which can inadvertently embed biases [19]. The limitations shown in clinical LLM studies, such as those highlighted in [32], demonstrate how these biases can manifest in medical decision-making, potentially impacting vulnerable patient groups. Addressing these challenges requires robust bias mitigation strategies during both training and deployment, including techniques like counterfactual data augmentation, debiasing modules [33], and advanced bias detection and correction mechanisms [34].

3) Privacy

Privacy concerns are particularly relevant in the context of LLMs, as these models are often trained on large datasets that may include sensitive or personally identifiable information (PII). Research has shown that LLMs can inadvertently memorize and reproduce such information, posing risks of data leakage and privacy violations [35]–[38]. Differential privacy and watermarking techniques are commonly employed to mitigate these risks. Tools like ProPILE [39] have demonstrated how privacy probes can assess the extent of PII exposure, underscoring the importance of robust privacy-preserving measures in training and deployment to safeguard sensitive data effectively.

4) Robustness

Robustness refers to the model's ability to maintain performance even when faced with challenging or adversarial inputs. This is critical for preventing the model from being easily manipulated or misled. For instance, LLMs used in chemical compound discovery were found to be susceptible to adversarial attacks that caused the models to suggest invalid molecular structures, compromising the reliability of scientific outputs [40]. Similarly, the incorrect handling of rare scientific terms in biological research led to inconsistent conclusions, highlighting weaknesses in the model's understanding of specialized terminology [41]. Adversarial training [34], [42] and robustness testing [43], [44] are common approaches to enhance the resilience of LLMs.

5) Toxicity and Legality

LLMs must be safeguarded against generating toxic or illegal content, such as hate speech or misinformation [4]. For example, in medical research, there have been cases where LLMs generated misleading treatment suggestions, potentially causing harm to patients if used unchecked [45]. In legal contexts, LLMs have also produced outputs that inadvertently violated data protection regulations, leading to privacy breaches and regulatory issues [38]. [46] examines such challenges of ensuring LLMs comply with data protection laws, highlighting instances where models have memorized and leaked sensitive information. Content moderation techniques, including the use of toxicity classifiers and red-teaming exercises, help to ensure that LLM outputs remain within acceptable ethical and legal boundaries [47]–[49].

6) Uncertainty and Explainability

Uncertainty quantification [50], [51] is crucial for assessing the confidence of LLM outputs, particularly in scientific and decision-making contexts. For example, in a real-world application involving LLMs for drug discovery, the lack of explainability led to misunderstandings about the efficacy of proposed compounds, resulting in wasted research efforts and resources [52], [53]. Similarly, in climate modeling, unexplained discrepancies in LLM-generated forecasts led to incorrect conclusions, affecting policy decisions [54]–[56]. Techniques like self-consistency [57] and chain-of-thought (CoT) prompting [58] are used to improve the transparency and reliability of LLM reasoning processes [59]. Tools such as LIME [60] and SHAP [61] have also been adopted to help explain model outputs, providing a deeper understanding of the decision-making pathways of LLMs in scientific contexts.

Building on these foundations, Dong et al. (2024) [27] discuss the challenges in constructing comprehensive guardrails, particularly when balancing multiple, sometimes conflicting requirements—such as fairness, privacy, and robustness. These challenges are compounded by the fact that the requirements for these guardrails often interact in complex ways, as highlighted by Raji and Buolamwini (2019) [62] in their discussion on the impact of biased datasets on AI outcomes, and Carlini et al. [35], [36], who demonstrated how LLMs could inadvertently memorize and expose private information. [63], [64] also emphasize the trade-offs between maintaining privacy and achieving fairness in machine learning models. They highlight the difficulty of establishing these guardrails due to the diverse nature of these requirements across different applications. A multidisciplinary approach, integrating both symbolic and learning-based methods, is proposed to address these challenges, ensuring that guardrails can adapt to the evolving capabilities of LLMs while maintaining rigorous safety standards. Furthermore, [65]–[67] provide insights into privacy vulnerabilities through membership inference attacks, which underscore the need for robust privacy guardrails in scientific contexts.

While existing LLM guardrail frameworks provide essential safety measures, the scientific domain presents distinct requirements that extend beyond conventional safeguards. Scientific applications demand additional layers of verification, reproducibility, and methodological rigor that current frameworks do not fully address. The following section examines these domain-specific challenges and introduces novel guardrail dimensions crucial for maintaining scientific integrity in LLM-assisted research.

B. Challenges of Implementing LLM Guardrails for Science

While LLMs present significant advantages across various domains, their deployment within scientific contexts introduces a unique set of challenges that must be addressed to maintain scientific integrity. These challenges primarily revolve around issues of scientific integrity, reliability, ethical considerations, and legal compliance. Scientific research demands precise and trustworthy information, making the mitigation of errors, biases, and unethical outputs crucial.

The importance of enhancing LLMs to prevent harmful outputs is further underscored by Tang et al. (2024) [68]. They examine the specific vulnerabilities of LLM-based agents in scientific domains, emphasizing the need for a triadic framework involving human regulation, agent alignment, and environmental feedback. This framework aims to mitigate risks such as factual errors, jailbreak attacks, and the misuse of scientific information. The authors argue that safeguarding efforts should prioritize risk control over the autonomous capabilities of LLMs, particularly in high-stakes scientific environments. In the context of scientific research, biases can differ significantly from those typically encountered in general-purpose LLMs. While general-domain biases often involve demographic-based issues, such as gender, race, or cultural stereotypes, biases in the scientific domain tend to focus on inaccuracies related to research quality, methodological errors, over-representation of certain hypotheses, and institutional or publication and positivity biases. These scientific biases can result in the disproportionate representation of particular theories, flawed conclusions due to biased training data, or even reinforcing historically dominant research paradigms.

Addressing these challenges requires implementing domain-specific solutions on top of general guardrails to ensure that LLMs support the advancement of reliable and ethical scientific research. Some of these specific challenges include:

1) Hallucination Mitigation

One of the most significant challenges is preventing hallucinations — instances where LLMs generate factually incorrect or misleading outputs. This is an issue in scientific contexts where the accuracy of information is critical. These hallucinations can either be open-domain, where models make false general claims easily verifiable against reliable sources, or closed-domain where the models deviate from the provided context or reference text [28]. Such inaccuracies in scientific contexts can have severe consequences, undermining the integrity and credibility of research outputs. Current approaches to hallucination prevention, such as formal verification and adversarial training, may not fully integrate with domain-specific knowledge bases [69]–[72], which are crucial for ensuring scientific accuracy.

2) Temporal Relevancy

Another critical challenge is ensuring that LLM-generated content remains up-to-date with the latest scientific findings. Scientific research is a dynamic field, and outdated information can lead to erroneous conclusions, negatively impacting the credibility of the research. Existing frameworks may lack mechanisms to ensure the time relevancy of LLM outputs, limiting their applicability in fast-evolving scientific domains.

3) Conflict Identification and Resolutiuon

Handling conflicting research results is a complex issue that general LLM approaches may not adequately address. In fields where research findings frequently evolve or contradict previous studies, LLMs must be capable of reconciling these conflicts to provide reliable and coherent outputs. However, current guardrails may not possess the necessary sophistication to manage such complexities effectively

4) Consistency

Maintaining consistency across outputs is crucial in scientific research. Inconsistent outputs can undermine the validity of research findings, leading to a lack of coherence and reliability in the generated content. Ensuring consistency also helps in building a reliable knowledge base that other researchers can trust and build upon. Moreover, maintaining consistency reduces the likelihood of conflicting interpretations, thereby safeguarding the integrity of the overall scientific inquiry.

5) Attribution

Ensuring proper attribution and citation is equally important in scientific research. Inadequate citation practices can lead to plagiarism and intellectual dishonesty, which can severely compromise the integrity of the research. Inconsistent outputs can undermine the validity of research, while inadequate citation practices can lead to plagiarism and intellectual dishonesty. General LLM guardrails may not enforce the rigorous standards required by scientific journals, posing risks to the integrity of the research.

6) Explainability and Transparency

In scientific research, the explainability of LLM outputs is essential for ensuring transparency and traceability. Researchers need to understand the reasoning process behind LLM-generated content to verify its alignment with established scientific principles. However, general-purpose guardrails often provide only superficial reasoning chains, lacking the depth required for rigorous scientific validation.

7) Ethical and Legal Considerations

The ethical and legal challenges associated with LLMs in scientific research are significant. These include ensuring fairness, mitigating biases, protecting privacy, and adhering to legal standards such as intellectual property rights. General LLM guardrails may not fully address these challenges, particularly in contexts like clinical research, where the stakes and requirements are high.

To address these challenges, He et al. [73] proposepossible enhancements to these guardrails that could be effective in scientific domains. They propose SciGuard, a system specifically designed to control the misuse risks of AI models in science. SciGuard serves as a mediator between users and AI models, implementing ethical and safety standards customizable to different domains. The growing prevalence of AI applications in science brings escalating concerns about their potential misapplication. It highlights the urgency for systems like SciGuard that can prevent unintended harm even when AI is used with good intentions. They further highlight that the dynamic nature of risks in scientific AI necessitates continuous monitoring and reassessment to ensure the effectiveness and relevance of risk mitigation strategies, making it imperative that future development focuses on building safeguarded scientific AI systems that align with both ethical standards and scientific integrity.

While prior studies have explored general-purpose guardrails for LLMs [6], there remains a significant gap in addressing the unique challenges specific to scientific research. To bridge this gap, we reformulate the guardrails within a different framework that extends traditional guardrail dimensions, incorporating aspects relevant to science such as time sensitivity, knowledge contextualization, and conflict resolution. Our contributions are two-fold:

- **Expansion of LLM guardrail dimensions to support applications in the scientific domain**: We introduce a categorization framework that expands the traditional guardrail dimensions to include those uniquely pertinent to scientific research.
- **Implementation strategies**: We provide actionable strategies for implementing these guardrails using a hybrid of white-box, black-box, and gray-box approaches, motivated by a similar approach taken by Dong et al [6] for preventing LLM attacks.

In the next section, we propose specific dimensions for implementing effective LLM guardrails tailored to scientific research.

III. Dimensions for Guardrails for Science

In the previous section, we identified various gaps in applying general LLM guardrails to the scientific domain, such as the lack of domain-specific adaptation, insufficient temporal relevance, and challenges in handling multidisciplinary data. Here we propose a set of tailored areas that address and encompass those properties within specific scientific needs. These dimensions are categorized under four main areas: *Trustworthiness*, *Ethics & Bias*, *Safety*, and *Legal*. Each category is further subdivided into critical areas for improvement, offering a roadmap for applying LLM guardrails more effectively within scientific contexts.

The dimensions are represented in a structured framework, illustrated in Figure 1, and categorized by the level of adaptation required. We use a color-coding scheme – blue, orange, red, and uncolored – to indicate the extent of modification needed for existing LLM guardrails to be applied to scientific research. Blue boxes signify areas where minimal adaptation is necessary, orange boxes denote dimensions that require some refinement, red boxes represent areas where significant development is required, and uncolored boxes represent a categorization that indicates areas already addressed by LLMs across multiple domains, rather than specific, well-defined guardrails.

We expand each category, highlighting key dimensions and explaining their importance in scientific research. Each of these dimensions has been chosen for detailed exploration based on its critical relevance to the unique challenges presented in scientific contexts. While LLMs can contribute broadly across various domains, these particular dimensions require specialized adaptations to meet the high standards of scientific research. For example, *Compliance* ensures that research adheres to institutional, legal, and ethical guidelines [74]–[76], while *Attribution* guarantees proper credit is given to original sources, a critical aspect of maintaining scientific integrity [77]–[79]. *Time Sensitivity* is important in fields where the timeliness of information can affect research outcomes [80], and *Knowledge Contextualization* ensures that the vast amount of information processed by LLMs is relevant and accurate within the specific scientific context it is applied [81]. We deliberately chose not to expand on every dimension in detail, as some are well-understood and already have robust guardrails in place that apply broadly across domains. By narrowing our focus to these specific dimensions, we aim to provide deeper insights into areas where LLMs must be particularly adapted or enhanced to meaningfully contribute to advancing scientific knowledge while maintaining the highest standards of reliability and integrity.

A. Blue Boxes

The blue boxes represent the dimensions of LLM guardrails that can be adapted for scientific use with minimal modification. These are established best practices that can transition smoothly into the scientific context, requiring only slight adjustments. The goal here is to leverage these existing guardrails to enhance LLMs for the specific needs of scientific research, ensuring their outputs are dependable and contextually appropriate. The following dimensions outline the key areas where LLM guardrails must be implemented to ensure they are suitable for scientific research. These dimensions aim to enhance the reliability, safety, and ethical compliance of LLM outputs

- **Verification:** Ensuring that LLM outputs can be verified against known, established sources.
- **Uncertainty Identification:** Mechanisms for identifying uncertainty in LLM outputs.
- **Consistency:** Maintaining consistency in information, particularly concerning scientific data.
- **Fairness:** Addressing fairness concerns to prevent biased or prejudiced outputs.

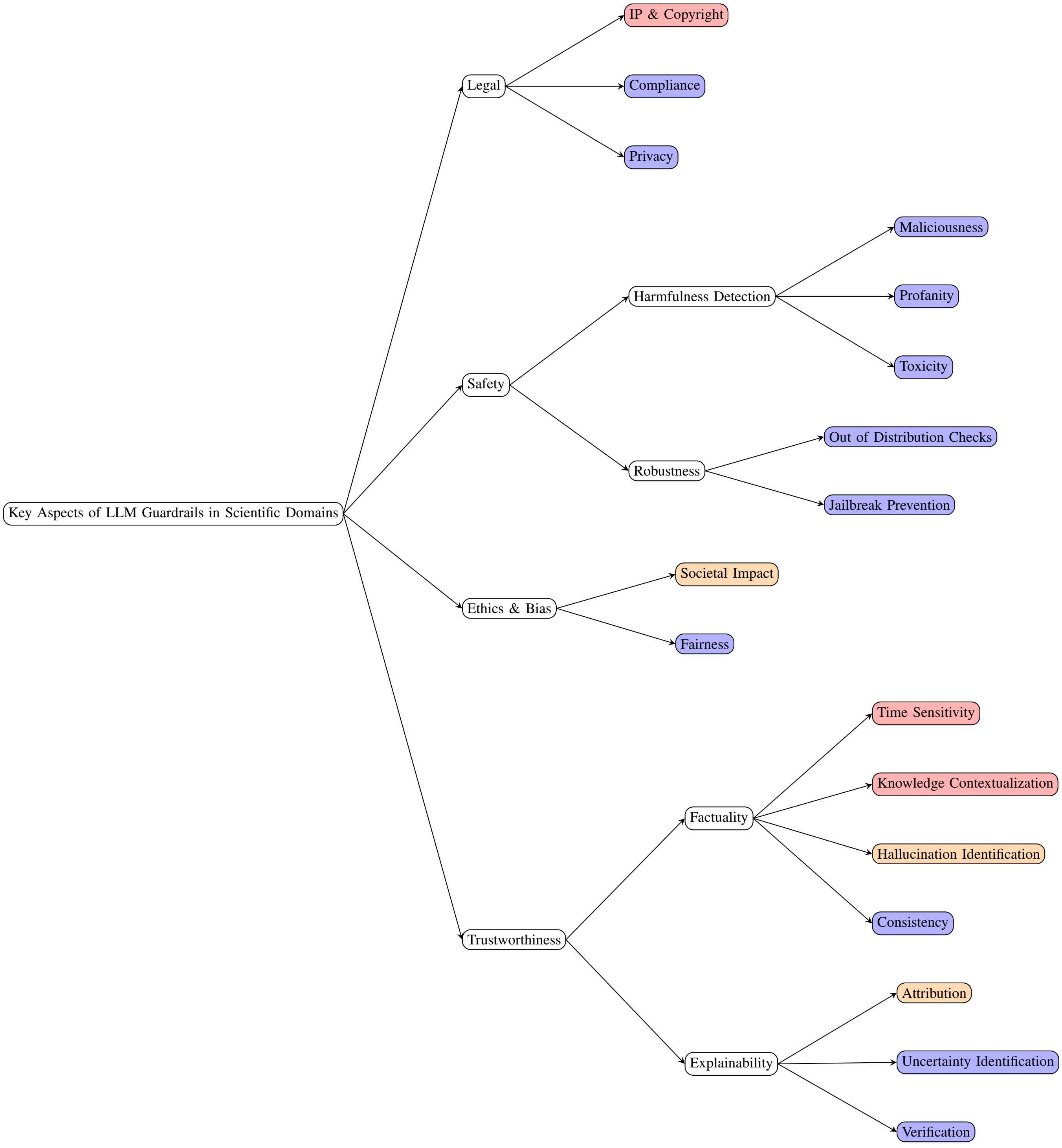

Fig. 1: Key Aspects of LLM Guardrails in Scientific Domains

- **Robustness:** Strengthening LLMs against potential failures, including jailbreak prevention and out-of-distribution checks.
- **Harmfulness Detection:** Detecting harmful outputs related to toxicity, profanity, and maliciousness.
- **Privacy:** Ensuring LLMs comply with privacy laws and safeguard personal information.
- **Compliance:** Ensuring that LLMs adhere to legal, ethical, and institutional guidelines.

1) Compliance

The **Compliance** dimension ensures that the identification of similar research work and relevant grants adhere to legal, ethical, and institutional guidelines, while also promoting trustworthiness in scientific research. This functionality is critical for optimizing resource allocation, avoiding research duplication, and fostering collaboration. This also focuses on developing sophisticated tools that can automatically search and compare existing literature, grant opportunities, and ongoing research projects, thereby aiding researchers in navigating the vast and ever-expanding body of scientific knowledge. For instance, an LLM could analyze a researcher's proposal and automatically suggest related studies or previous work in the same area, helping to ensure that the new research builds on existing knowledge rather than inadvertently replicating it. No guardrail dimensions exist yet that address this problem. While there has been some work [82] on matching grants in a specific domain, there has yet to be proper discipline on the implementation in a wider scientific domain.

Similarly, it also ensures that the works follow publication guidelines and scientific values. Identifying related grants is particularly valuable for researchers seeking funding opportunities that align with their projects. By scanning grant databases and matching them with the researcher's area of study, LLMs can suggest relevant funding opportunities, streamlining the grant application process and increasing the chances of securing support. This functionality saves time and opens up potential collaborations by identifying research groups working on similar problems.

The development of these capabilities would require LLMs to perform comprehensive searches and analyses, cross-referencing multiple sources of academic literature and funding databases. Advanced natural language processing (NLP) techniques would be necessary to accurately parse and understand the nuanced details of research proposals and grant descriptions, ensuring that the recommendations [83] made are highly relevant and useful to the researchers.

Incorporating these search and comparison tools into LLMs would empower researchers to make more informed decisions, enhance the efficiency of the research process, and foster a more collaborative scientific community. These tools represent a significant step toward integrating AI more deeply into the research lifecycle, supporting the broader goals of innovation and knowledge advancement in the scientific community. [84]

B. Orange boxes

The orange boxes represent existing dimensions but require significant refinement to meet the unique demands of scientific integrity, accountability, and precision. These dimensions are crucial in ensuring that LLMs provide accurate outputs, uphold ethical standards, and promote contextual understanding within scientific disciplines.

- **Attribution:** Ensuring proper credit to original sources to maintain academic integrity.
- **Hallucination Identification:** The ability to recognize when LLMs provide inaccurate or fabricated information.
- **Societal Impact:** Considering the broader societal effects of LLM-generated outputs.

1) Attribution

Proper attribution is essential in scientific research to acknowledge the original sources, avoid plagiarism, and maintain academic integrity. LLMs may inadvertently generate content that closely mirrors existing literature without appropriate citations, raising concerns about intellectual dishonesty and transparency of LLM-generated content. To address this, LLMs must be incorporated with mechanisms for accurate attribution [79], such as generating citations in the correct academic formats, implementing source-tracking capabilities, and integrating plagiarism detection algorithms. [85] [86]

Challenges include the complexity of source materials, knowledge cutoffs [87] leading to outdated citations, and variability in citation styles across disciplines. Strategies to improve attribution involve integrating LLMs with citation databases like PubMed or CrossRef, employing retrieval-augmented generation techniques [88] to access the latest publications, and standardizing citation practices within the models. By prioritizing accurate attribution, LLMs can enhance the trustworthiness of their outputs, uphold ethical standards, and facilitate verification of information, thereby fostering a more transparent and reliable scientific community.

C. Red boxes

The red boxes represent dimensions where current LLM guardrails are either underdeveloped to meet scientific standards or non-existent in the general literature. These dimensions remain underdeveloped precisely because they address challenges unique to scientific discourse—such as truthfulness, accuracy, and reproducibility—which rarely arise in general-purpose applications. For scientific research, certain attributes require a significant focus due to the unique challenges they present. For example, *Time Sensitivity* is essential because scientific knowledge evolves rapidly, and LLMs must provide information that reflects the latest developments to remain relevant. *Knowledge Contextualization* is equally crucial, as scientific recommendations must be tailored to specific fields, regions, or study conditions to be genuinely useful. By expanding on these dimensions, we aim to ensure that LLMs are better equipped to handle the complexity and dynamism of scientific research. These areas are crucial for the future development of LLMs tailored to scientific research.

1) Time Sensitivity

LLMs are trained on static datasets with a specific knowledge cutoff date and lack mechanisms for real-time knowledge updates [87] [80]. This limitation presents several critical issues with relying on LLMs for scientific applications:

- **Static Knowledge and Rapid Obsolescence**: LLMs are ”frozen” at their last point of training, unable to incorporate or reflect the most recent advancements unless explicitly retrained. [80], [87] In rapidly evolving scientific fields like epidemiology, biotechnology, or climate science, research findings and policy changes occur frequently, rendering information obsolete within months or weeks. This static nature is especially problematic in domains where timely, accurate

information is essential for decision-making, such as public health, finance, or crisis management [89]. LLMs cannot respond to breaking news or incorporate real-time data about stock market fluctuations, emerging viral outbreaks, or sudden regulatory changes, limiting their utility in environments that demand agility and current knowledge.
- **Consequences of Outdated Information and False Confidence**: LLMs often present outdated information with high confidence, lacking built-in mechanisms to indicate that their knowledge may be outdated. This makes it difficult for users to discern whether the information provided is still accurate or relevant. For example, a scientist seeking advice on the latest vaccine technologies may receive authoritative-sounding responses that reflect research no longer considered cutting-edge. Relying on outdated information can lead to significant consequences, including the spread of misinformation, misallocation of resources, and ineffective strategies [90]. In fields like public health, this could result in promoting vaccines less effective against new viral strains, leading to increased infection rates and unnecessary strain on healthcare systems. The model's high confidence in outdated data exacerbates the risk of misguided conclusions or recommendations, as users may not be alerted to the possibility that newer research has changed the consensus.

Therefore, as scientific knowledge is dynamic and constantly evolving, LLMs must provide information that is not only accurate but also temporally relevant. The challenge is that many LLMs are trained on datasets that may become outdated [87] [80], leading to the generation of content that no longer reflects the current state of scientific understanding. To address this, models need to be equipped with mechanisms to verify the temporal validity of information by cross-referencing against real-time data sources. This could involve integrating LLMs with updated knowledge bases or using retrieval-augmented generation (RAG) [88] techniques to pull in current information as part of the response generation process.

2) Knowledge Contextualization

Similar to time sensitivity, LLM responses must adapt their responses to the specific context in which information is applied. In scientific research, the same data can produce varying conclusions depending on the region, field of study, or unique circumstances of the user's inquiry. **Knowledge contextualization** ensures that LLMs take these variations into account, allowing for more nuanced, tailored responses. This means recognizing disciplinary differences, adjusting recommendations based on the user's level of expertise [81], and considering geographical or environmental factors that may influence how a particular piece of knowledge should be applied. By understanding the context, LLMs can offer outputs that are not only factually correct but also pragmatically relevant to the specific research or situation at hand.

For example, in agricultural research, a general recommendation on farming best practices may be valid for a temperate region but completely unsuitable for arid climates. A lack of contextualization could result in crop damage or environmental harm. LLMs need the ability to adapt information dynamically to reflect these unique local conditions, which often requires an understanding of diverse disciplines, localized expertise, and complex interdependencies within the context of a specific inquiry. Below are key challenges that emerge when LLMs fail to properly contextualize knowledge:

- **Overgeneralization:** LLMs often provide generalized responses that fail to account for local variations or specific situational needs. This can lead to recommendations that are ineffective or even harmful when applied in contexts with unique environmental, demographic, or infrastructural factors. Overgeneralization risks poor decision-making and unintended negative consequences, particularly in domains like agriculture and healthcare, where local conditions are critical for successful outcomes.
- **Failure to Integrate Multidisciplinary Knowledge:** LLMs often struggle to synthesize information from multiple disciplines, providing narrow, discipline-specific responses. This can lead to incomplete or unbalanced recommendations, particularly in complex fields like sustainable farming, where solutions require integrating knowledge from climatology, ecology, economics, and soil science. The lack of multidisciplinary integration may result in decisions that optimize one aspect while neglecting others, leading to suboptimal or harmful outcomes, even if the query to LLM might seem benign. One potential mitigation strategy could be to adapt LLMs to use tools that incorporate these disciplines. [91]
- **Lack of Adaptive Reasoning:** LLMs often lack the ability to adjust their responses to the specific context of a problem, such as local economic, environmental, or social factors. Adaptive reasoning [92] [93] requires weighing multiple, sometimes conflicting variables, which LLMs struggle to do effectively. Without this ability, they may offer one-size-fits-all recommendations that overlook critical nuances, leading to poor decision-making and undesirable outcomes in areas like public policy, healthcare, or disaster management.
- **Positivity Bias**: LLMs tend to exhibit a positivity bias [94], [95], often providing overly optimistic responses that may not accurately reflect the underlying data or context. This bias can lead to an unrealistic portrayal of outcomes or underestimation of risks, particularly in fields like healthcare or environmental management. Proper contextualization of facts and more nuanced evaluations are necessary to ensure that responses are balanced and reflect the true range of potential outcomes.

Additionally, it is common to encounter studies with differing conclusions involving the synthesis of conflicting data and perspectives. Current LLMs may struggle to handle these conflicts effectively [96], leading to oversimplified or biased outputs. LLMs need to be able to identify and appropriately manage these conflicts, rather than presenting conflicting results as equally valid without distinction. Therefore, it is necessary to develop frameworks that allow LLMs to recognize, reconcile, and accurately represent conflicting findings in scientific literature. This would involve integrating advanced reasoning algorithms like chain-of-thoughts [58], chain-of-verification [97], verify-and-edit [98] that can weigh evidence, evaluate the credibility of sources, and present balanced conclusions. Furthermore, solutions could involve developing models that can prioritize information based on the consensus within the scientific community or highlight the divergence in findings to guide users in making informed decisions. Conflict resolution mechanisms within LLMs could be employed to manage these situations effectively [99] [100].

3) IP & Copyright

Intellectual property (IP) and copyright play a fundamental role in scientific research, where the originality and ownership of ideas, data, and innovations are crucial [101]. This dimension focuses on developing mechanisms within LLMs that can accurately identify, track, and manage intellectual property and copyrighted content, ensuring that the outputs generated respect the proprietary rights of researchers, institutions, and other stakeholders. By integrating these capabilities, LLMs can uphold legal and ethical standards, maintaining the integrity of the scientific process while protecting the rights of individuals and organizations.

As LLMs generate content, it is crucial to recognize and label portions of text that may be novel, patentable, or otherwise protected under intellectual property laws [101]. This capability is essential not only for protecting IP rights but also for guiding researchers and organizations in managing and utilizing these outputs. For instance, if an LLM generates an innovative idea or a unique piece of content, the system should be able to flag this as potentially patentable, alerting users to the need for further steps to secure IP protection [102].

Moreover, ensuring that attribution of sources is a vital component of IP and copyright management. LLMs should be equipped with citation and attribution mechanisms to accurately acknowledge the origins of data, ideas, or prior work referenced in generated content. This preserves academic integrity and mitigates legal risks related to IP infringement or unauthorized use of proprietary information. By doing so, the system can identify instances where the output may require licensing or other legal compliance before it can be used commercially [103].

This proposed categorization of LLM guardrails addresses the complex challenges of using LLMs in scientific applications that emphasize trust, ethical responsibility, safety, and legal compliance. We focused primarily on dimensions highlighted by the red boxes, which represent areas where current LLM guardrails are insufficient for the scientific domain. In the following section, we outline implementation strategies for enhancing these guardrails, categorized as white-box, black-box, and gray-box approaches to meet the unique demands of the scientific domain.

IV. Implementation Strategies

To ensure that the proposed framework for LLM guardrails in the scientific domain is effectively implemented, we adapt the categorization of guardrail-attack types from Dong et al. [6] into our approach for guardrail implementation: **White-box**, **Black-box**, and **Gray-box**. While Dong et al. originally used these categories to classify attack vectors based on model access levels, we extend this paradigm to systematically organize guardrail implementation strategies. In their work, white-box attacks assume full visibility into the model's parameters, black-box attacks restrict access to observing model outputs, and gray-box attacks offer partial access, such as to training data. We adapted this framework to describe corresponding strategies for enforcing guardrails based on the level of access and intervention available within the model. These strategies form a comprehensive framework to address challenges related to trustworthiness, safety, ethics, and legal compliance within the specific context of scientific research, in various levels of effort.

A. White-Box Approaches

White-box approaches leverage direct access to and modification of LLM internal architecture and its parameters. This allows for precise control over the model's behavior, providing the ability to enforce guardrails from within the system itself. Techniques such as fine-tuning, model optimization, and formal verification allow developers to adjust and improve the model's outputs in a controlled and trustworthy manner.

Model Fine-Tuning: This approach involves adjusting the weights and parameters of the LLM through additional training data, allowing developers to refine the model's behavior. In the context of scientific research, it helps reduce biases and enhance the accuracy of the model's output, especially for domain-specific tasks like handling complex datasets or ensuring factual correctness. [104] [105]

Architectural Modifications: This approach involves modifying the underlying architecture of the LLM [106], such as introducing new layers or mechanisms (e.g. memory systems [107], mixture-of-experts [108]), enables more robust internal guardrails. These modifications can optimize the model for key scientific challenges such as consistency, uncertainty quantification, and knowledge contextualization. [109]

Bias Mitigation Techniques: These techniques involve embedding fairness and bias mitigation [33] [34] directly within the model, ensuring that the LLM produces equitable outcomes across different datasets. This is crucial for scientific domains where unbiased and reliable results are essential.

Formal Verification: This technique involves proving that the model's behavior conforms to predefined standards or rules, ensuring trustworthiness by verifying specific properties of the system. For example, formal verification can ensure that an LLM does not produce harmful or misleading results. [110]

Adversarial Training and Robustness: In scientific domains where accuracy and robustness are critical, adversarial training helps the model learn to resist potential attacks or manipulations, ensuring that outputs remain reliable even under challenging input conditions. [34] [43]

White-box strategies are particularly effective for scenarios requiring high levels of control and precision, making them ideal for working with sensitive scientific data that demands rigorous internal safety and trustworthiness mechanisms. While these methods allow for superior control by enabling direct modification of the LLM's internal workings, they are often harder to implement due to their complexity, requiring substantial resources and expertise. However, they introduce significant complexity, largely due to the resource-intensive nature of training and fine-tuning, the need for domain-specific expertise, and the challenge of acquiring large, high-quality datasets. For instance, reinforcement learning from human feedback (RLHF) [104] fine-tunes [111] the LLM based on human feedback to align the model with specific scientific goals, improving its trustworthiness. Yet, RLHF requires substantial resources, curated datasets, and continuous human oversight, adding to the complexity of implementation. Moreover, scientific data's diversity and complexity, along with the need to regularly update models with new findings, complicate the maintenance of these internal guardrails. Thus, while white-box methods provide better control and precision, their resource demands and complexity may limit practical downstream implementation, especially for smaller organizations with constrained resources.

B. Black-Box Approaches

In contrast to white-box methods, black-box approaches rely on external mechanisms to enforce guardrails without modifying the LLM's internal structure. These approaches monitor, filter, and adjust the outputs of the model from the outside, making them effective when direct access to the model is not possible or when the model's internal workings must remain unchanged.

Output Filtering: This approach uses post-processing tools [20] [24] to monitor and filter the outputs of the LLM. It ensures that generated content aligns with ethical guidelines, scientific standards, and legal requirements. Output filtering is particularly effective for managing issues such as toxicity and hallucination identification in a scientific context, ensuring the content remains reliable. [47] [112]

Rule-Based Post-Processing: This method applies predefined rules to the output and identifies and controls any content that violates those rules [21]. For example, rules could be set to ensure that the model's predictions adhere to specific scientific norms or avoid controversial statements. This helps in ensuring that the model's outputs align with ethical standards such as fairness and societal impact.

External Fact-Checking: Scientific contexts often require high factual accuracy. By integrating external fact-checking systems [21] [113], the black-box approach allows for real-time validation of model outputs [29], ensuring that incorrect or misleading information is flagged or corrected before being disseminated. [114] [3]

Content Moderation: In domains where sensitive information is involved, content moderation systems [115] [21] can help prevent the dissemination of harmful, inappropriate, or legally problematic outputs, such as biased or inaccurate scientific data. This ensures that the LLM maintains a high standard of safety and reliability. [116]

Adversarial Input Detection: Black-box approaches can also involve detecting and mitigating adversarial inputs that could exploit vulnerabilities in the LLM. This approach ensures that the model maintains robustness even in challenging scenarios by flagging these inputs before they affect the output. [117]

Black-box strategies are ideal for integrating external guardrails with pre-existing LLMs, especially in scientific research where ethical, legal, and factual considerations are critical. A common example of this approach is output filtering, where the LLM processes user input, and the generated output is passed through external guardrails, such as content filters to remove harmful language, PII redaction for privacy compliance, word filters to block specific terms, and denied topics to prevent restricted content. Additionally, a contextual grounding check ensures factual accuracy and relevance. Encoder-based models [118], [119], often used for classification tasks, effectively classify the generated outputs to ensure they adhere to ethical and factual standards [20]. They can be used to label outputs for compliance, detect sensitive topics, or evaluate whether the generation aligns with pre-defined guidelines, thereby supporting responsible LLM outputs. These strategies enforce responsible AI policies without modifying the model's internal structure, making them adaptable, cost-effective, and easy to implement [120]. However, their effectiveness depends heavily on the quality of external rules and filters, which can lead to challenges in managing complex or nuanced scenarios. Furthermore, black-box methods can introduce latency in processing and limit control over deeper model behavior, particularly in cases where more precise intervention is required [27].

C. Gray-Box Approaches

Gray-box approaches offer a middle ground that combines the control and precision of white-box methods with the adaptability of black-box methods, providing a hybrid solution that allows for both internal modifications and external interventions. These strategies provide flexibility, making them particularly useful for managing dynamic, evolving challenges in scientific research where internal model adjustments and external verifications are both necessary.

Type	Strategy	Trustworthiness							Ethics & Bias		Safety		Legal
		Verification	Uncertainty Identification	Attribution	Consistency	Hallucination Identification	Knowledge Contextualization	Time Sensitivity	Fairness	Societal Impact	Robustness	Harmfulness Detection	IP & Copyright
White-box	Model Fine-Tuning			[104], [105]	[104], [105]	[104], [105]	[104], [105]	[104], [105]	[104], [105]				
	Architectural Modifications	[106]–[108], [121]			[106]–[108], [121]	[106]–[108], [121]	[106]–[108], [121]	[106]–[108], [121]					
	Bias Mitigation Techniques								[33], [34]	[33], [34]			
	Knowledge Base Integration			[121]		[121]	[121]	[121]					
	Regularization Methods				[51]								
	Adversarial Training	[34], [36]	[34], [36]			[34], [36]			[34], [36]		[34], [36]	[34], [36]	
	Confidence Scoring		[59]										
	Formal Verification	[110]											
	Customized Loss Functions				[34]	[34]							
Gray-box	Ensemble Modeling		[122]		[122]	[122]							
	Human-in-the-Loop Systems	[123]		[123]	[123]	[123]	[123]	[123]	[123]	[123]		[123]	
	Retrieval-Augmented Generation			[88]		[88]	[88]	[88]					
	Hybrid Feedback Mechanisms	[43]	[43]	[43]	[43]	[43]							
	Semi-Supervised Learning			[124], [125]		[124], [125]			[124], [125]				
	External Knowledge Integration					[121]	[121]	[121]					
	Dynamic Parameter Adjustment				[51]								
	Selective Response Generation		[59]										
	Post-Hoc Explainability Tools	[106]	[106]										
	Policy Gradient Fine-Tuning	[111]							[111]	[111]	[111]	[111]	
Black-box	Output Filtering				[20], [24]	[20], [24]				[20], [24]		[20], [24]	
	Rule-Based Post-Processing	[21]		[21]	[21]	[21]	[21]	[21]	[21]	[21]	[21]	[21]	[21]
	External Fact-Checking	[29], [88]				[29], [88]							
	Content Moderation								[47], [48]	[47], [48]	[47], [48]	[47], [48]	
	Adversarial Input Detection	[44]	[44]								[44]	[44]	
	Plagiarism Detection			[35]									[35]

TABLE II: Implementation Strategies for LLM Guardrails for Science

Human-in-the-Loop Systems: These systems incorporate human oversight [111] into the LLM's decision-making process. Human experts can verify outputs in real time, ensuring that the model adheres to scientific standards and ethical guidelines. This approach is particularly effective for managing complex or novel situations where human expertise is essential. [123]

Ensemble Modeling: Ensemble approaches [122] combine multiple models or techniques to increase the robustness and reliability of LLM outputs [126]. In the scientific domain, this can be used to cross-check outputs, ensuring greater consistency and accuracy.

Hybrid Feedback Mechanisms: These approaches integrate feedback from both internal model performance and external monitoring systems. For example, a gray-box method could combine formal verification (internal) with rule-based post-processing (external) to ensure that the model's output is not only consistent but also ethical and legally compliant.

Retrieval-Augmented Generation: RAG augments the generative capabilities of LLMs by retrieving relevant information from trusted sources, ensuring that the outputs are factual and reliable. This helps the model produce accurate, context-aware responses in scientific contexts by integrating up-to-date information from peer-reviewed research or scientific databases. [88] [127]

Semi-Supervised Learning: These approaches leverage labeled and unlabeled data, enabling the model to learn from real-world data while still receiving feedback on performance. This hybrid learning method ensures the LLM can continuously improve its outputs over time. [124] [125]

Gray-box strategies are well-suited for complex scientific domains where models need to be both internally trustworthy and externally validated. By combining internal adjustments with external interventions, these hybrid approaches offer a balanced solution to managing bias, societal risks, and evolving research demands. Techniques like chain-of-thought (CoT) prompting [58], which enhances the model's reasoning by guiding it to generate more logical, step-by-step responses, improve precision without requiring deep modifications. Additionally, retrieval-augmented generation (RAG) [88] allows the model to pull in real-time, trusted external data during the generation process, ensuring outputs are grounded in verified, up-to-date scientific knowledge. These techniques make gray-box methods highly adaptable in handling dynamic scientific challenges. However, they can be resource-intensive, requiring some internal modifications and external interventions. While more flexible than white-box strategies, they may lack the same level of fine-tuned control over the model's internal processes.

To ensure the effectiveness of these LLM guardrails across diverse applications in the scientific domain, table II outlines a variety of approaches categorized into white-box, black-box, and gray-box methods. These approaches align specific techniques to key dimensions such as consistency, uncertainty identification, fairness, and attribution, ensuring that each method can address the unique challenges posed by the LLM in scientific contexts. White-box approaches, such as model architecture modification and fine-tuning, enhance internal mechanisms like bias mitigation, formal verification, and uncertainty quantification. These methods are well-suited for tasks requiring high levels of control and trustworthiness, allowing LLMs to handle sensitive scientific data with greater precision.

On the other hand, black-box approaches offer external mechanisms such as rule-based enforcement and API-based moderation, making them ideal for addressing ethical concerns like toxicity, factuality, and societal impact without altering the internal structure of the model. Gray-box methods provide a flexible combination of internal and external techniques, leveraging hybrid strategies such as human-in-the-loop verification and cross-validation with trusted datasets. These approaches, mapped to specific dimensions in the table, provide a balanced solution for managing complex scenarios, such as intellectual property compliance, conflict identification, and time-sensitive outputs. By aligning the appropriate guardrail strategy with the relevant dimension, the framework ensures that LLMs can be governed effectively in the evolving and dynamic landscape of scientific research.

V. Conclusion

The scientific domain presents both significant challenges as well as opportunities in the deployment of large language models. These challenges are not just technical but also involve addressing broader concerns related to trust, ethical use, and societal impact. While LLMs can accelerate discovery, enhance workflows, and aid in knowledge synthesis, they also pose significant risks related to reliability, ethics, and legal compliance. Mitigating these risks is crucial to harnessing the full potential of LLMs without compromising the core values of scientific inquiry. The comprehensive guardrails proposed in this paper are essential for addressing these concerns, ensuring that LLMs provide accurate, fair, and legally compliant outputs. By focusing on key dimensions – such as hallucination prevention, knowledge contextualization, and conflict identification – the proposed framework aims to enhance the trustworthiness, ethical responsibility, and legal integrity of LLMs in scientific domains. This will ultimately contribute to a more reliable and transparent use of AI in science, ensuring that its integration supports, rather than undermines, the scientific process.

As the scientific landscape continues to evolve, these guardrails must be adaptable and responsive to emerging developments. Flexibility in updating and evolving these guardrails is key to their effectiveness, particularly in the face of rapidly changing scientific knowledge and technologies. By incorporating tailored strategies such as white-box, black-box, and gray-box approaches, we aim to provide a structured framework for understanding the challenges of guardrails and proposing mitigation steps, required efforts, and future directions to address the evolving needs of the scientific community. Each of these approaches brings distinct advantages, and their combined use offers a holistic mechanism to ensure safety, fairness, and accountability in LLM deployments. This proposed framework serves as a foundation for advancing the safe, effective, and ethical use of LLMs in science, fostering innovation while upholding the integrity and rigor essential to scientific progress. Ongoing collaboration among AI developers, domain experts, and policymakers will be vital to refine and implement these guardrails, ensuring that LLMs continue to serve the best interests of the scientific community.

References

[1] T. B. Brown, “Language models are few-shot learners,” *arXiv preprint arXiv:2005.14165*, 2020.

[2] A. Radford, “Improving language understanding by generative pre-training,” 2018.

[3] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, “Finetuned language models are zero-shot learners,” *arXiv preprint arXiv:2109.01652*, 2021.

[4] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?” in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 610–623.

[5] J. Xu, D. Ju, M. Li, Y.-L. Boureau, J. Weston, and E. Dinan, “Bot-adversarial dialogue for safe conversational agents,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 2950–2968.

[6] Y. Dong, R. Mu, Y. Zhang, S. Sun, T. Zhang, C. Wu, G. Jin, Y. Qi, J. Hu, J. Meng, S. Bensalem, and X. Huang, “Safeguarding large language models: A survey,” *arXiv preprint arXiv: 2406.02622*, 2024.

[7] D. Hovy and S. L. Spruit, “The social impact of natural language processing,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2016, pp. 591–598.

[8] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, “Dissecting racial bias in an algorithm used to manage the health of populations,” *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[9] H. H. Thorp, “Chatgpt is fun, but not an author,” pp. 313–313, 2023.

[10] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.

[11] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, “The llama 3 herd of models,” *arXiv preprint arXiv:2407.21783*, 2024.

[12] Anthropic, “Introducing claude 3.5 sonnet,” Jun 2024. [Online]. Available: https://www.anthropic.com/news/claude-3-5-sonnet

[13] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand *et al.*, “Mixtral of experts,” *arXiv preprint arXiv:2401.04088*, 2024.

[14] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth *et al.*, “Gemini: a family of highly capable multimodal models,” *arXiv preprint arXiv:2312.11805*, 2023.

[15] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, “On the opportunities and risks of foundation models,” *arXiv preprint arXiv:2108.07258*, 2021.

[16] M. R. AI4Science and M. A. Quantum, “The impact of large language models on scientific discovery: a preliminary study using gpt-4,” *arXiv preprint arXiv:2311.07361*, 2023.

[17] J. Boyko, J. Cohen, N. Fox, M. H. Veiga, J. I.-H. Li, J. Liu, B. Modenesi, A. H. Rauch, K. N. Reid, S. Tribedi, A. Visheratina, and X. Xie, “An interdisciplinary outlook on large language models for scientific research,” *arXiv preprint arXiv: 2311.04929*, 2023.

[18] J. Haltaufderheide and R. Ranisch, “The ethics of chatgpt in medicine and healthcare: a systematic review on large language models (llms),” *npj Digital Medicine*, vol. 7, no. 1, Jul. 2024. [Online]. Available: http://dx.doi.org/10.1038/s41746-024-01157-x

[19] J. Zhang, X. Ji, Z. Zhao, X. Hei, and K.-K. R. Choo, “Ethical considerations and policy implications for large language models: Guiding responsible development and deployment,” *arXiv preprint arXiv: 2308.02678*, 2023.

[20] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, D. Testuggine, and M. Khabsa, “Llama guard: Llm-based input-output safeguard for human-ai conversations,” *arXiv preprint arXiv: 2312.06674*, 2023.

[21] T. Rebedea, R. Dinu, M. Sreedhar, C. Parisien, and J. Cohen, “Nemo guardrails: A toolkit for controllable and safe llm applications with programmable rails,” *arXiv preprint arXiv: 2310.10501*, 2023.

[22] L. Beurer-Kellner, M. Fischer, and M. Vechev, “Prompting is programming: A query language for large language models,” *Proceedings of the ACM on Programming Languages*, vol. 7, no. PLDI, pp. 1946–1969, 2023.

[23] S. Slundberg and contributors, “Guidance: A language for controlling large language models,” https://github.com/guidance-ai/guidance, 2024, accessed: 2024-09-30.

[24] S. Rajpal and contributors, “Guardrails: Adding guardrails to large language models,” https://github.com/guardrails-ai/guardrails, 2024, accessed: 2024-09-30.

[25] K. E. Goodman, H. Y. Paul, and D. J. Morgan, “Ai-generated clinical summaries require more than accuracy,” *JAMA*, 2024.

[26] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin *et al.*, “A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions,” *arXiv preprint arXiv:2311.05232*, 2023.

[27] Y. Dong, R. Mu, G. Jin, Y. Qi, J. Hu, X. Zhao, J. Meng, W. Ruan, and X. Huang, “Building guardrails for large language models,” *arXiv preprint arXiv: 2402.01822*, 2024.

[28] R. Friel and A. Sanyal, “Chainpoll: A high efficacy method for llm hallucination detection,” *arXiv preprint arXiv:2310.18344*, 2023.

[29] P. Manakul, A. Liusie, and M. J. Gales, “Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models,” *arXiv preprint arXiv:2303.08896*, 2023.

[30] J. Fu, S.-K. Ng, Z. Jiang, and P. Liu, “Gptscore: Evaluate as you desire,” *arXiv preprint arXiv:2302.04166*, 2023.

[31] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, “G-eval: Nlg evaluation using gpt-4 with better human alignment,” *arXiv preprint arXiv:2303.16634*, 2023.

[32] P. Hager, F. Jungmann, R. Holland, K. Bhagat, I. Hubrecht, M. Knauer, J. Vielhauer, M. Makowski, R. Braren, G. Kaissis *et al.*, “Evaluation and mitigation of the limitations of large language models in clinical decision-making,” *Nature medicine*, vol. 30, no. 9, pp. 2613–2622, 2024.

[33] N. Meade, E. Poole-Dayan, and S. Reddy, “An empirical survey of the effectiveness of debiasing techniques for pre-trained language models,” *arXiv preprint arXiv:2110.08527*, 2021.

[34] B. H. Zhang, B. Lemoine, and M. Mitchell, “Mitigating unwanted biases with adversarial learning,” in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.

[35] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson *et al.*, “Extracting training data from large language models,” in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 2633–2650.

[36] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, “The secret sharer: Evaluating and testing unintended memorization in neural networks,” in *28th USENIX security symposium (USENIX security 19)*, 2019, pp. 267–284.

[37] S. Neel and P. Chang, “Privacy issues in large language models: A survey,” *arXiv preprint arXiv:2312.06717*, 2023.

[38] R. Staab, M. Vero, M. Balunovi’c, and M. T. Vechev, “Beyond memorization: Violating privacy via inference with large language models,” *International Conference on Learning Representations*, 2023.

[39] S. Kim, S. Yun, H. Lee, M. Gubri, S. Yoon, and S. J. Oh, “Propile: Probing privacy leakage in large language models,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[40] A. Wong, H. Cao, Z. Liu, and Y. Li, “Smiles-prompting: A novel approach to llm jailbreak attacks in chemical synthesis,” *arXiv preprint arXiv: 2410.15641*, 2024.

[41] Q. Zhang, K. Ding, T. Lyv, X. Wang, Q. Yin, Y. Zhang, J. Yu, Y. Wang, X. Li, Z. Xiang, K. Feng, X. Zhuang, Z. Wang, M. Qin, M. Zhang, J. Zhang, J. Cui, T. Huang, P. Yan, R. Xu, H. Chen, X. Li, X. Fan, H. Xing, and H. Chen, “Scientific large language models: A survey on biological & chemical domains,” *arXiv preprint arXiv: 2401.14656*, 2024.

[42] N. Jain, A. Schwarzschild, Y. Wen, G. Somepalli, J. Kirchenbauer, P. yeh Chiang, M. Goldblum, A. Saha, J. Geiping, and T. Goldstein, “Baseline defenses for adversarial attacks against aligned language models,” *arXiv preprint arXiv: 2309.00614*, 2023.

[43] E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving, “Red teaming language models with language models,” *arXiv preprint arXiv:2202.03286*, 2022.

[44] J. Wang, X. Hu, W. Hou, H. Chen, R. Zheng, Y. Wang, L. Yang, H. Huang, W. Ye, X. Geng *et al.*, “On the robustness of chatgpt: An adversarial and out-of-distribution perspective,” *arXiv preprint arXiv:2302.12095*, 2023.

[45] C. T.-T. Chang, H. Farah, H. Gui, S. J. Rezaei, C. Bou-Khalil, Y.-J. Park, A. Swaminathan, J. A. Omiye, A. Kolluri, A. Chaurasia *et al.*, “Red teaming large language models in medicine: Real-world insights on model behavior,” *medRxiv*, pp. 2024–04, 2024.

[46] D. Zhang, P. Finckenberg-Broman, T. Hoang, S. Pan, Z. Xing, M. Staples, and X. Xu, “Right to be forgotten in the era of large language models: Implications, challenges, and solutions,” *AI and Ethics*, 2023.

[47] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, “Realtoxicityprompts: Evaluating neural toxic degeneration in language models,” *arXiv preprint arXiv:2009.11462*, 2020.

[48] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin *et al.*, “Opt: Open pre-trained transformer language models,” *arXiv preprint arXiv:2205.01068*, 2022.

[49] Q. Liao and J. Vaughan, “Ai transparency in the age of llms: A human-centered research roadmap,” *Special Issue 4: Grappling With the Generative AI Revolution*, 2023.

[50] C. Zhu, B. Xu, Q. Wang, Y. Zhang, and Z. Mao, “On the calibration of large language models and alignment,” *arXiv preprint arXiv: 2311.13240*, 2023.

[51] K. A. Sankararaman, S. Wang, and H. Fang, “Bayesformer: Transformer with uncertainty estimation,” *arXiv preprint arXiv:2206.00826*, 2022.

[52] M. Wysocka, O. Wysocki, M. Delmas, V. Mutel, and A. Freitas, “Large language models, scientific knowledge and factuality: A framework to streamline human expert evaluation,” *Journal of Biomedical Informatics*, 2023.

[53] Y. Zheng, H. Y. Koh, M. Yang, L. Li, L. T. May, G. I. Webb, S. Pan, and G. Church, “Large language models in drug discovery and development: From disease mechanisms to clinical trials,” *arXiv preprint arXiv: 2409.04481*, 2024.

[54] J. Bulian, M. S. Schäfer, A. Amini, H. Lam, M. Ciaramita, B. Gaiarin, M. C. Huebscher, C. Buck, N. G. Mede, M. Leippold *et al.*, “Assessing large language models on climate information,” *arXiv preprint arXiv:2310.02932*, 2023.

[55] M. Fore, S. Singh, C. Lee, A. Pandey, A. Anastasopoulos, and D. Stamoulis, “Unlearning climate misinformation in large language models,” in *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, D. Stammbach, J. Ni, T. Schimanski, K. Dutia, A. Singh, J. Bingler, C. Christiaen, N. Kushwaha, V. Muccione, S. A. Vaghefi, and M. Leippold, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 178–192. [Online]. Available: https://aclanthology.org/2024.climatenlp-1.14

[56] M. Kraus, J. A. Bingler, M. Leippold, T. Schimanski, C. C. Senni, D. Stammbach, S. A. Vaghefi, and N. Webersinke, “Enhancing large language models with climate resources,” *arXiv preprint arXiv:2304.00116*, 2023.

[57] T. Ahmed and P. Devanbu, “Better patching using llm prompting, via self-consistency,” in *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 2023, pp. 1742–1746.

[58] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[59] Z. Jiang, J. Araki, H. Ding, and G. Neubig, “How can we know when language models know? on the calibration of language models for question answering,” *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 962–977, 2021.

[60] M. T. Ribeiro, S. Singh, and C. Guestrin, “”why should I trust you?”: Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 2016, pp. 1135–1144.

[61] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[62] I. D. Raji and J. Buolamwini, "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 429–435.

[63] S. Mittal, K. Thakral, R. Singh, M. Vatsa, T. Glaser, C. C. Ferrer, and T. Hassner, "On responsible machine learning datasets with fairness, privacy, and regulatory norms," *arXiv preprint arXiv: 2310.15848*, 2023.

[64] H. Chen, T. Zhu, T. Zhang, W. Zhou, and P. S. Yu, "Privacy and fairness in federated learning: on the perspective of tradeoff," *ACM Computing Surveys*, vol. 56, no. 2, pp. 1–37, 2023.

[65] S. Zarifzadeh, P. Liu, and R. Shokri, "Low-cost high-power membership inference attacks," in *Forty-first International Conference on Machine Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=sT7UJh5CTc

[66] N. Manzonelli, W. Zhang, and S. Vadhan, "Membership inference attacks and privacy in topic modeling," *arXiv preprint arXiv:2403.04451*, 2024.

[67] R. Wen, Z. Li, M. Backes, and Y. Zhang, "Membership inference attacks against in-context learning," *arXiv preprint arXiv:2409.01380*, 2024.

[68] X. Tang, Q. Jin, K. Zhu, T. Yuan, Y. Zhang, W. Zhou, M. Qu, Y. Zhao, J. Tang, Z. Zhang, A. Cohan, Z. Lu, and M. Gerstein, "Prioritizing safeguarding over autonomy: Risks of llm agents for science," *arXiv preprint arXiv: 2402.04247*, 2024.

[69] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," *arXiv preprint arXiv:2004.10964*, 2020.

[70] J. Li, Y. Yuan, and Z. Zhang, "Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases," *arXiv preprint arXiv: 2403.10446*, 2024.

[71] L. Cao, "Learn to refuse: Making large language models more controllable and reliable through knowledge scope limitation and refusal mechanism," *arXiv preprint arXiv: 2311.01041*, 2023.

[72] S. M. T. I. Tonmoy, S. M. M. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, and A. Das, "A comprehensive survey of hallucination mitigation techniques in large language models," *arXiv preprint arXiv: 2401.01313*, 2024.

[73] J. He, W. Feng, Y. Min, J. Yi, K. Tang, S. Li, J. Zhang, K. Chen, W. Zhou, X. Xie, W. Zhang, N. Yu, and S. Zheng, "Control risk for potential misuse of artificial intelligence in science," *arXiv preprint arXiv: 2312.06632*, 2023.

[74] J. M. DuBois, J. T. Chibnall, and J. Gibbs, "Compliance disengagement in research: Development and validation of a new measure," *Science and engineering ethics*, vol. 22, pp. 965–988, 2016.

[75] R. Watkins, "Guidance for researchers and peer-reviewers on the ethical use of large language models (llms) in scientific research workflows," *AI and Ethics*, pp. 1–6, 2023.

[76] D. B. Resnik and M. Hosseini, "The ethics of using artificial intelligence in scientific research: new guidance needed for a new tool," *AI and Ethics*, pp. 1–23, 2024.

[77] F. L. Macrina, *Scientific integrity: Text and cases in responsible conduct of research*. John Wiley & Sons, 2014.

[78] N. A. of Sciences, Policy, G. Affairs, and C. on Responsible Science, *Fostering integrity in research*. National Academies Press, 2017.

[79] J. Huang and K. Chang, "Citation: A key to building responsible and accountable large language models," *NAACL-HLT*, 2023.

[80] S. M. Mousavi, S. Alghisi, and G. Riccardi, "Dyknow:dynamically verifying time-sensitive factual knowledge in llms," *arXiv preprint arXiv: 2404.08700*, 2024.

[81] L. Ning, L. Liu, J. Wu, N. Wu, D. Berlowitz, S. Prakash, B. Green, S. O'Banion, and J. Xie, "User-llm: Efficient llm contextualization with user embeddings," *arXiv preprint arXiv:2402.13598*, 2024.

[82] J. Zhu, B. G. Patra, H. Wu, and A. Yaseen, "A novel nih research grant recommender using bert," *PloS one*, vol. 18, no. 1, p. e0278636, 2023.

[83] L. Wu, Z. Zheng, Z. Qiu, H. Wang, H. Gu, T. Shen, C. Qin, C. Zhu, H. Zhu, Q. Liu, H. Xiong, and E. Chen, "A survey on large language models for recommendation," *World Wide Web*, vol. 27, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s11280-024-01291-2/fulltext.html

[84] L. Floridi and J. Cowls, "A unified framework of five principles for ai in society," *Machine learning and the city: Applications in architecture and urban design*, pp. 535–545, 2022.

[85] L. Gao, Z. Dai, P. Pasupat, A. Chen, A. T. Chaganty, Y. Fan, V. Y. Zhao, N. Lao, H. Lee, D.-C. Juan *et al.*, "Rarr: Researching and revising what language models say, using language models," *arXiv preprint arXiv:2210.08726*, 2022.

[86] B. Bohnet, V. Q. Tran, P. Verga, R. Aharoni, D. Andor, L. B. Soares, M. Ciaramita, J. Eisenstein, K. Ganchev, J. Herzig *et al.*, "Attributed question answering: Evaluation and modeling for attributed large language models," *arXiv preprint arXiv:2212.08037*, 2022.

[87] C. Zhu, N. Chen, Y. Gao, Y. Zhang, P. Tiwari, and B. Wang, "Is your llm outdated? evaluating llms at temporal generalization," *arXiv preprint arXiv: 2405.08460*, 2024.

[88] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[89] M. A. Roberto, C. Antonio *et al.*, "Updating knowledge in large language models: an empirical evaluation," in *Conference Proceedings: 2024 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, 2024.

[90] I. Vykopal, M. Pikuliak, I. Srba, R. Móro, D. Macko, and M. Bieliková, "Disinformation capabilities of large language models," *Annual Meeting of the Association for Computational Linguistics*, 2023.

[91] P. Lu, B. Peng, H. Cheng, M. Galley, K.-W. Chang, Y. N. Wu, S.-C. Zhu, and J. Gao, "Chameleon: Plug-and-play compositional reasoning with large language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[92] S. Chen and B. Li, "Toward adaptive reasoning in large language models with thought rollback," in *Forty-first International Conference on Machine Learning*, 2024.

[93] A. M. P. Aggarwal, Y. Yang, and Mausam, "Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with llms," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Association for Computational Linguistics, 2023, pp. 12 375–12 396. [Online]. Available: https://doi.org/10.18653/v1/2023.emnlp-main.761

[94] L. Tjuatja, V. Chen, T. Wu, A. Talwalkwar, and G. Neubig, "Do llms exhibit human-like response biases? a case study in survey design," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 1011–1026, 2024.

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

[96] R. Xu, Z. Qi, Z. Guo, C. Wang, H. Wang, Y. Zhang, and W. Xu, "Knowledge conflicts for llms: A survey," *arXiv preprint arXiv: 2403.08319*, 2024.

[97] S. Dhuliawala, M. Komeili, J. Xu, R. Raileanu, X. Li, A. Celikyilmaz, and J. Weston, "Chain-of-verification reduces hallucination in large language models," *arXiv preprint arXiv:2309.11495*, 2023.

[98] R. Zhao, X. Li, S. R. Joty, C. Qin, and L. Bing, "Verify-and-edit: A knowledge-enhanced chain-of-thought framework," *Annual Meeting of the Association for Computational Linguistics*, 2023.

[99] Y. Liu, Z. Yao, X. Lv, Y. Fan, S. Cao, J. Yu, L. Hou, and J. Li, "Untangle the KNOT: Interweaving conflicting knowledge and reasoning skills in large language models," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 17 186–17 204. [Online]. Available: https://aclanthology.org/2024.lrec-main.1493

[100] I.-C. Chern, S. Chern, S. Chen, W. Yuan, K. Feng, C. Zhou, J. He, G. Neubig, and P. Liu, “Factool: Factuality detection in generative ai - a tool augmented framework for multi-task and multi-domain scenarios,” *arXiv preprint arXiv: 2307.13528*, 2023.

[101] B. L. Sobel, “Artificial intelligence’s fair use crisis,” *Colum. JL & Arts*, vol. 41, p. 45, 2017.

[102] S. Chesterman, “Good models borrow, great models steal: intellectual property rights and generative ai,” *Policy and Society*, p. puae006, 2024.

[103] I. Abdikhakimov, “Unraveling the copyright conundrum: Exploring ai-generated content and its implications for intellectual property rights,” in *International Conference on Legal Sciences*, vol. 1, no. 5, 2023, pp. 18–32.

[104] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, “Training language models to follow instructions with human feedback,” *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[105] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, “Direct preference optimization: Your language model is secretly a reward model,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[106] E. Mitchell, C. Lin, A. Bosselut, C. Finn, and C. D. Manning, “Fast model editing at scale,” *arXiv preprint arXiv:2110.11309*, 2021.

[107] Y. Wu, M. N. Rabe, D. Hutchins, and C. Szegedy, “Memorizing transformers,” *arXiv preprint arXiv:2203.08913*, 2022.

[108] N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. W. Yu, O. Firat *et al.*, “Glam: Efficient scaling of language models with mixture-of-experts,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 5547–5569.

[109] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, D. De Las Casas, A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. Rae, E. Elsen, and L. Sifre, “Improving language models by retrieving from trillions of tokens,” in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 2206–2240. [Online]. Available: https://proceedings.mlr.press/v162/borgeaud22a.html

[110] X. Huang, W. Ruan, W. Huang, G. Jin, Y. Dong, C. Wu, S. Bensalem, R. Mu, Y. Qi, X. Zhao *et al.*, “A survey of safety and trustworthiness of large language models through the lens of verification and validation,” *Artificial Intelligence Review*, vol. 57, no. 7, p. 175, 2024.

[111] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan *et al.*, “Training a helpful and harmless assistant with reinforcement learning from human feedback,” *arXiv preprint arXiv:2204.05862*, 2022.

[112] J. Xu, D. Ju, M. Li, Y.-L. Boureau, J. Weston, and E. Dinan, “Recipes for safety in open-domain chatbots,” *arXiv preprint arXiv:2010.07079*, 2020.

[113] Google Fact Check Explorer, “Google fact check explorer - recent claims,” https://toolbox.google.com/factcheck/, 2024, accessed: 2024-10-01.

[114] I. Vykopal, M. Pikuliak, S. Ostermann, and M. Šimko, “Generative large language models in automated fact-checking: A survey,” 2024. [Online]. Available: https://arxiv.org/abs/2407.02351

[115] V. Lai, S. Carton, R. Bhatnagar, Q. V. Liao, Y. Zhang, and C. Tan, “Human-ai collaboration via conditional delegation: A case study of content moderation,” in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–18.

[116] D. Kumar, Y. A. AbuHashem, and Z. Durumeric, “Watch your language: Investigating content moderation with large language models,” in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 18, 2024, pp. 865–878.

[117] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing *et al.*, “Judging llm-as-a-judge with mt-bench and chatbot arena,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 46 595–46 623, 2023.

[118] J. Devlin, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[119] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv: 1907.11692*, 2019.

[120] J. Cheng, X. Liu, K. Zheng, P. Ke, H. Wang, Y. Dong, J. Tang, and M. Huang, “Black-box prompt optimization: Aligning large language models without model training,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 3201–3219. [Online]. Available: https://aclanthology.org/2024.acl-long.176

[121] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark *et al.*, “Improving language models by retrieving from trillions of tokens,” in *International conference on machine learning*. PMLR, 2022, pp. 2206–2240.

[122] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” *Advances in neural information processing systems*, vol. 30, 2017.

[123] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, “Deep reinforcement learning from human preferences,” *Advances in neural information processing systems*, vol. 30, 2017.

[124] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, “Big self-supervised models are strong semi-supervised learners,” *Advances in neural information processing systems*, vol. 33, pp. 22 243–22 255, 2020.

[125] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, “Fixmatch: Simplifying semi-supervised learning with consistency and confidence,” *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.

[126] D. Jiang, X. Ren, and B. Y. Lin, “Llm-blender: Ensembling large language models with pairwise ranking and generative fusion,” *arXiv preprint arXiv:2306.02561*, 2023.

[127] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson, “From local to global: A graph rag approach to query-focused summarization,” *arXiv preprint arXiv:2404.16130*, 2024.