IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 2, April 2026, pp. 1071~1086
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i2.pp1071-1086

The role of prompt engineering in enhancing LLMs: a systematic review of applications and ethical implications

Izzul Fatawi[1], Muhammad Roil Bilad[2], Muhammad Asy'ari[3]
[1]Faculty of Education and Teacher Training, Universitas Terbuka, Banten, Indonesia
[2]Faculty of Integrated Technologies, Universiti Brunei Darussalam, Bandar Seri Begawan, Brunei Darussalam
[3]Faculty of Applied Science and Engineering, Universitas Pendidikan Madalika, Mataram, Indonesia

Article Info

Article history:

Received Oct 13, 2024
Revised Jul 5, 2025
Accepted Feb 6, 2026

Keywords:

Bias mitigation
Chain-of-thought prompting
Data privacy
Ethical artificial intelligence
Large language models
Prompt engineering
Retrieval-augmented generation

ABSTRACT

Large language models (LLMs) have transformed natural language processing (NLP), demonstrating exceptional proficiency in tasks such as text generation, translation, and summarization. However, LLMs are prone to generating biased, inaccurate, or contextually irrelevant outputs, posing significant risks in high-stakes domains such as healthcare, legal reasoning, and engineering. This paper systematically investigates the role of prompt engineering as a solution to these challenges. By strategically designing inputs, prompt engineering enhances LLM performance, yielding more accurate, contextually relevant, and ethically aligned outputs. Advanced techniques, including chain-of-thought (CoT) prompting and retrieval-augmented generation (RAG), are examined for their ability to improve reasoning capabilities, reduce errors, and mitigate bias. CoT prompting facilitates structured, stepwise reasoning, while RAG incorporates real-time data, ensuring output accuracy in rapidly evolving fields. In addition, we present a novel comparative perspective on these techniques, highlighting their distinct strengths and limitations across specialized applications such as healthcare diagnostics and scientific data extraction. The findings demonstrate that sophisticated prompt engineering significantly elevates the reliability and precision of LLM outputs, while addressing critical ethical concerns such as data privacy, bias, and hallucination. These insights underscore the necessity of advanced prompt design in optimizing LLMs for high-impact applications, ensuring both performance and ethical integrity.

Corresponding Author:

Muhammad Asy'ari
Faculty of Applied Science and Engineering, Universitas Pendidikan Mandalika
St. Pemuda No. 59A, Mataram-83125, Indonesia
Email: muhammadasyari@undikma.ac.id

1. INTRODUCTION

Large language models (LLMs) have rapidly become foundational in various academic and industrial sectors, primarily due to their remarkable proficiency in understanding and generating human language. These models, such as generative pre-trained transformer-4 (GPT-4) and bidirectional encoder representations from transformers (BERT), have revolutionized natural language processing (NLP) by leveraging vast datasets, allowing them to perform complex linguistic tasks with unparalleled accuracy [1], [2]. LLMs have demonstrated their ability to generalize across multiple tasks, excelling in settings that require minimal fine-tuning [3], [4]. However, maximizing their potential requires careful input structuring through prompt engineering—a practice that strategically crafts inputs to ensure optimal model performance [5]. Despite their strengths, LLMs present challenges such as inaccuracies, biases, and

inefficiencies in generating context-specific outputs. For instance, in critical fields like healthcare or education, a poorly structured prompt can lead to outputs that are irrelevant or even harmful [6]. This issue is exacerbated by the LLMs' inherent tendency toward generating hallucinations—factually incorrect information—which could undermine trust in AI applications, particularly in sensitive domains like medical diagnostics or legal reasoning [7]. Consequently, the problem lies in effectively guiding LLMs to produce accurate, relevant, and ethically sound outputs, particularly in high-stakes fields where precision is crucial.

Prompt engineering has emerged as a key solution to this problem. By designing precise, structured inputs, engineers can control and optimize the outputs of LLMs, making them more relevant and aligned with the user's needs [8]. This technique has evolved significantly, with methods such as chain-of-thought (CoT) prompting and retrieval-augmented generation (RAG) being developed to enhance LLM performance in complex reasoning tasks [9]. These advancements have expanded the utility of LLMs beyond simple text generation, enabling their application in domains requiring nuanced problem-solving, such as scientific research, engineering, and medical diagnostics [10], [11]. The existing literature underscores the critical role of prompt engineering in enhancing LLM capabilities across various fields. In healthcare, for example, prompt engineering has been used to refine LLMs for diagnostic support, enabling more accurate and reliable outputs by structuring prompts to simulate clinical decision-making processes [12]. In educational settings, LLMs powered by prompt engineering have been used to generate personalized learning materials, tailoring content to meet the specific needs of students [13]. Furthermore, in technical documentation, prompt engineering allows LLMs to process and generate complex reports more efficiently, thereby improving workflow and information dissemination [14].

This study proposes that sophisticated prompt engineering techniques—particularly CoT prompting and RAG—can significantly mitigate the challenges of hallucinations and bias in LLMs. CoT prompting, which guides LLMs through a step-by-step reasoning process, has proven effective in reducing errors by breaking down complex tasks into manageable steps [6]. This method is especially valuable in fields like medical diagnostics, where the cost of errors is high. Similarly, RAG enhances LLM performance by incorporating external databases into the generation process, allowing LLMs to access up-to-date information and improve the accuracy of their outputs [10]. These innovations represent a shift from general text generation to specialized, high-accuracy applications, underscoring the transformative potential of prompt engineering. The novelty of this study lies in its comprehensive approach to integrating these advanced prompt engineering techniques into diverse research and industrial applications. While previous research has focused on the general capabilities of LLMs, this study provides a detailed examination of how CoT and RAG techniques can be applied to improve LLM outputs in specific, high-stakes environments. Moreover, this research addresses the ethical implications of LLM deployment, particularly the need for transparency, bias mitigation, and data privacy [7]. By combining technical advancements with ethical considerations, this study offers a more holistic framework for the responsible use of LLMs in sensitive fields.

The proposed approach goes beyond existing literature by not only focusing on performance optimization but also addressing the ethical challenges that have become increasingly prominent as LLMs are integrated into critical decision-making processes. In healthcare, for example, CoT prompting and RAG can be used to improve diagnostic accuracy while simultaneously ensuring that patient data is protected and that biases in decision-making are minimized [15]. This dual focus on technical and ethical optimization sets the current study apart from prior research, which has often treated these issues separately. Prompt engineering represents a crucial advancement in the deployment of LLMs across various sectors. By strategically crafting inputs, engineers can guide LLMs to generate more accurate, relevant, and ethically sound outputs. Techniques such as CoT prompting and RAG are particularly promising, offering solutions to the persistent issues of hallucination and bias. Moreover, this paper emphasizes a comparative exploration of different prompting strategies, elucidating how domain-specific challenges—such as scientific data extraction or educational material development—can benefit uniquely from CoT, RAG, or contextual prompting approaches. This study's contribution lies in it is integration of these advanced techniques into a comprehensive framework that addresses both performance and ethical considerations, marking a significant step forward in the responsible and effective use of LLMs.

2. METHOD

This study employed a systematic literature review (SLR) to explore the role of prompt engineering in enhancing LLMs across scientific and engineering applications, with a particular emphasis on the ethical implications of LLM deployment. The review followed a rigorous and structured approach to ensure comprehensive coverage of the most relevant and high-quality research. Additionally, a cross-comparative analysis of different prompt engineering methods was incorporated to highlight domain-specific strengths and weaknesses, thereby extending the scope beyond standard SLR practices. The study aimed to address the

following research questions: i) how does prompt engineering improve the performance of LLMs in scientific and engineering domains?; ii) what are the key techniques in prompt engineering, such as CoT prompting and RAG, that enhance the accuracy and relevance of LLM outputs?; and iii) what are the ethical challenges associated with LLM deployment in science and engineering, particularly concerning bias, hallucination, and data privacy?

2.1. Database selection

The Scopus database was selected as the primary source for this review due to its extensive repository of peer-reviewed journals, conference proceedings, and authoritative reports [16]. Scopus is highly regarded for its comprehensive academic coverage. It is suitable for capturing studies that discuss both the technical aspects of prompt engineering and its applications in science and engineering [17].

2.2. Search strategy

A targeted search string was developed to capture the intersection of prompt engineering, LLMs, and their applications in research and industrial settings. The following search string was used in Scopus, with a focus on title, abstract, and keywords: TITLE-ABS-KEY ((“Prompt Engineering” OR “Prompt Design”) AND (“Large Language Models” OR ”LLMs”) AND (“Enhancing” OR ”Improving” OR “Optimizing”) AND (“Applications” OR “Use Cases”) AND (“Ethical Implications” OR “Ethics” OR “Responsible AI”) AND (“Science,” OR “Engineering”). This search string was designed to retrieve studies that focus specifically on the role of prompt engineering in improving LLMs in the context of scientific and engineering research. The inclusion of terms such as “science,” and “engineering” ensured the relevance of the articles to the study’s objectives. The flow of documents identification included in the current study presented in Figure 1.

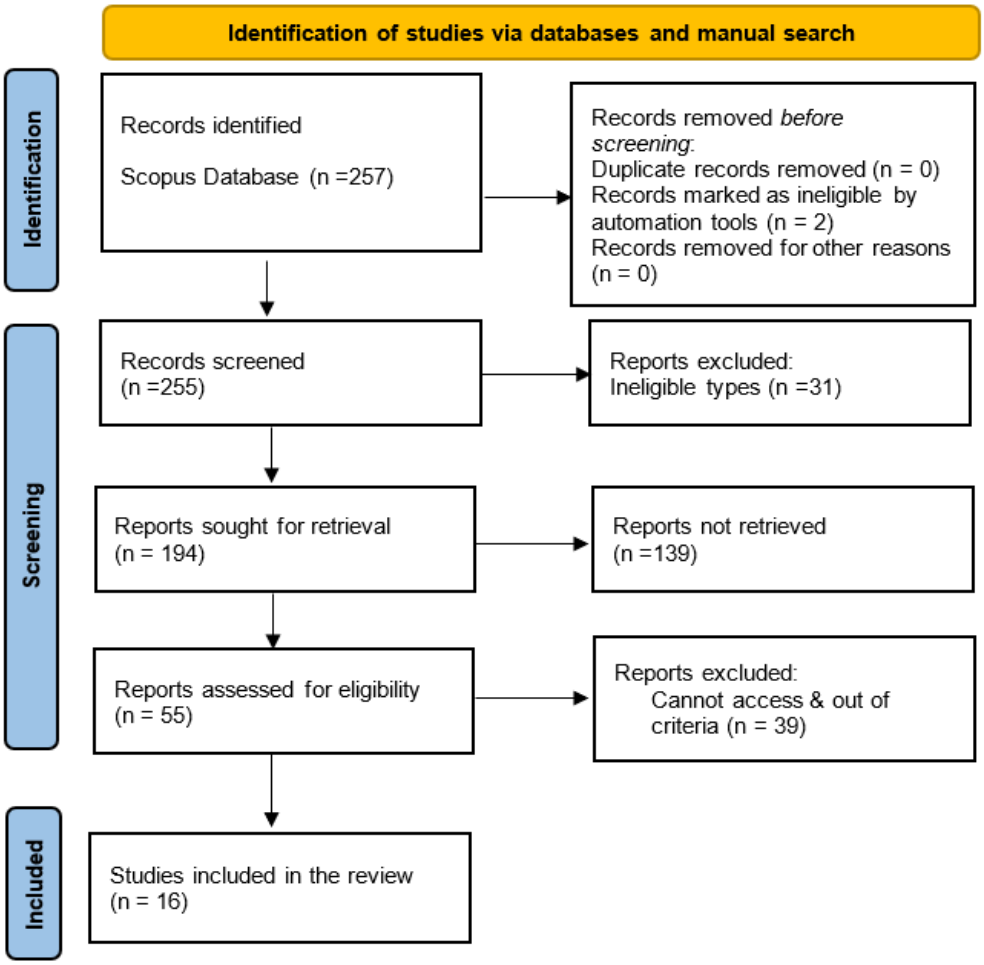

Figure 1. PRISMA flow

2.3. Inclusion and exclusion criteria

Table 1 presents the inclusion and exclusion criteria applied in the current study. The criteria were designed to ensure the relevance of the studies included in the review. They also help maintain the overall quality of the review.

Table 1. Inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
– Studies published between 2022 and 2024, reflecting the most recent advancements in LLMs and prompt engineering. – Peer-reviewed journal articles, conference papers, and authoritative reports that focus on prompt engineering within the context of science and engineering research. – Articles discussing the ethical implications of LLM usage, including concerns around bias, hallucination, and data privacy.	– Articles that do not explicitly address prompt engineering or its application in LLMs. – Publications focused on LLMs in general, but without a clear focus on scientific or engineering domains. – Non-peer-reviewed materials, such as editorials, opinion pieces, or studies without empirical evidence.

2.4. Study selection process

The initial search using the predefined string yielded a substantial number of studies. The first screening involved reviewing the titles and abstracts to remove irrelevant papers. In the second phase, a full-text review of the remaining studies was conducted to ensure that each paper met the inclusion criteria and addressed at least one of the research questions. Throughout the selection process, particular attention was paid to the rigor and relevance of each study's methodology. Studies were evaluated based on their empirical approach, the clarity of their findings, and the significance of their contributions to understanding prompt engineering in LLMs. Articles that lacked sufficient methodological detail or were tangential to the core research questions were excluded.

2.5. Data extraction and synthesis

For each selected study, key data were extracted, including: i) the specific prompt engineering techniques employed (e.g., CoT and RAG); ii) the domain of application (e.g., healthcare, scientific data extraction, and technical documentation); iii) the outcomes of implementing prompt engineering in LLMs, with a focus on improvements in accuracy, efficiency, and problem-solving capabilities; and iv) ethical considerations discussed by the authors, particularly related to bias, data privacy, and the problem of hallucination in LLM-generated content. Once data were extracted, the studies were synthesized to identify common themes and trends. During this synthesis, we also performed a domain-specific comparative review to assess the relative efficacy of each technique (e.g., CoT vs. RAG) within different high-stakes applications, such as scientific data extraction or engineering design. This approach provided a more nuanced understanding of how particular prompt engineering methods excel or face limitations in distinct contexts. Furthermore, the ethical concerns raised in these studies were grouped into recurring themes, such as model transparency, bias mitigation, and privacy protection.

2.6. Limitations

While this systematic review provides a comprehensive synthesis of the role of prompt engineering in enhancing LLMs, it is important to acknowledge several limitations. First, the review focused exclusively on studies from the Scopus database, which may have excluded relevant studies indexed in other databases such as IEEE Xplore or Google Scholar. Second, the search string, while targeted, may have missed studies that used alternative terminology for prompt engineering or LLMs. Third, the review was restricted to studies published in English, which may have led to the exclusion of important research conducted in other languages.

3. RESULTS AND DISCUSSION

3.1. Summary of findings

Prompt engineering has emerged as a key methodology for optimizing the capabilities of LLMs in various domains, improving both the efficiency and effectiveness of LLM outputs. The current systematic review identifies several notable applications of prompt engineering (16 studies included for review) that administered in Table 2. To provide a more thorough comparative perspective, we have expanded Table 2 to include additional domain-specific details on each application, highlighting not only the ethical implications but also which prompt engineering techniques (e.g., CoT and RAG) were employed. This enrichment aims to more explicit demonstrations of how prompt engineering strategies impact LLM performance across diverse fields.

Table 2. Notable applications of prompt engineering based on synthesis results

Role and application of LLMs/AI	Prompt engineering techniques	Ethical implications	Ref
Role of LLMs in feedback and education: – LLMs (e.g., GPT-3.5-turbo) used for generating personalized student feedback, improving revision performance (effect size d=0.19) and motivation (d=0.36). – Employed as educational assistants for tailored feedback. – Higher AI literacy identified as crucial for effective prompt design and improved learning outcomes. – LLMs enhance adaptive, personalized, and self-directed learning.	– Iterative prompting (refining prompts for improved student feedback) – Contextual prompting (incorporating domain-specific educational goals)	– Misinformation risk: LLMs can generate inaccurate feedback if not verified by educators. – Accessibility and Bias: over-reliance on AI outputs may exacerbate bias, especially with limited AI literacy. – Unequal access: disparities in AI literacy can widen educational and professional opportunity gaps. – Privacy concerns: using commercial LLM services for sensitive student data may violate privacy.	[18] – [21]
Prompt engineering for task-specific optimization: – Iterative prompt refinement (e.g., Gaussian process expected improvement (GPEI) methodology) to optimize LLM responses. – Role-playing and CoT prompting for adaptability in tasks like patient interaction and cybersecurity. – Educational chatbots refined for responsiveness and adaptability.	– GPEI for refining prompts – Role-playing prompts (to simulate real-world interactions) – CoT (stepwise reasoning for task complexity)	– Privacy concerns: using commercial LLMs in healthcare may expose sensitive data to third parties. – Misalignment in AI outputs: role-playing or CoT prompting can introduce biases if not carefully controlled. – Bias and quality issues: despite refinements, prompt-engineered outputs can still exhibit inconsistencies and biases.	[19], [22], [23]
LLMs for enhancing educator efficiency and financial analysis: – Automating administrative tasks (e.g., clinical notes) and analyzing financial data. – Generating synthetic data for rebalancing training sets. – Analyzing financial and economic trends for investment and risk decisions	– Contextual prompting (domain-specific prompts for financial or administrative tasks) – Iterative refinement (improving synthetic data generation accuracy)	– Security and privacy risks: sensitive clinical or financial data could be mishandled. – Bias in generated data: synthetic data might embed or amplify existing biases. – Data security: confidential info leakage (e.g., ChatGPT exposure). – Bias in financial analysis: propagation of biased datasets may distort decision-making.	[23] – [26]
LLMs for personalized interactions and chatbots: – Used in healthcare diagnostics, multilingual public services, and education. – "Llama 2" deployed locally to extract structured medical data. – Chatbots for patient self-diagnosis and healthcare support.	– Contextual prompting (adapting prompts to medical or multilingual settings) – Task-Specific Role Definition (e.g., patient vs. doctor roles)	– Hallucination and misinformation: plausible yet false outputs pose risks in healthcare and education. – Bias and ageism: technology misuse may marginalize older adults with limited digital literacy. – Privacy concerns: cloud-based LLMs must comply with regulations (e.g., GDPR). – Risk of misinformation: misleading medical info may harm patients if unvalidated.	[23], [24], [27], [28]
LLMs for dataset Jailbreaking and adversarial attacks: – Genetic algorithms automate jailbreaking prompts to bypass alignment techniques. – Creation of universal jailbreak prompts across multiple LLMs. – Adversarial prompts undermine LLM safety.	– Adversarial prompt engineering (exploiting system vulnerabilities) – Genetic algorithm-based prompt optimization (iteratively refining malicious prompts)	– Security vulnerabilities: jailbreaking exposes flaws, letting adversarial prompts bypass safety. – Ethical misuse: universal jailbreak prompts pose risks of malicious exploitation. – Hallucinations in code: erroneous outputs become more frequent or less detectable under adversarial conditions.	[29] – [31]
Role of AI literacy in prompt engineering: – AI literacy is vital for non-experts to use LLMs effectively. – Prompt engineering skills directly affect output quality (iterative improvement). – Anthropomorphizing AI can lead to misunderstanding LLM capabilities. – AI literacy influences search strategies for clinical questions.	– Iterative prompting (user-driven feedback loops) – Contextual prompting (guiding non-experts to frame questions effectively)	– Bias and misalignment: non-experts may misinterpret AI outputs, worsening bias and reliance on inaccuracies. – Privacy concerns: in educational contexts, insecure AI tools risk data exposure. – Risk of misinformation: retrieving false or outdated literature from LLMs can undermine research validity.	[20], [21], [30], [32]
LLMs in human evaluation and public services: – GPT-4 vs. Google AI comparison for translation, creativity, accuracy. – Automating tasks (e.g., drafting clinical notes) for administrative efficiency. – Testing bias in phishing detection (F1-score 97.29% fine-tuned vs. 92.74% prompt-engineered).	– Prompt-based vs. fine-tuned approaches (comparing effectiveness in tasks like phishing detection). – Contextual prompting (e.g., domain-tailored for public services).	– Privacy and misinformation risks: both GPT-4 and Google AI show biases and hallucination issues, especially in critical areas like healthcare. – Ethical use: unvalidated LLMs may yield harmful advice in sensitive fields. – Transparency: need clarity and interpretability in AI outputs, especially in cybersecurity applications (e.g., phishing detection).	[23], [28], [33]
LLMs for code generation and debugging: – ChatGPT used to generate/refine code (e.g., constructing LDA topic models). – Streamlining and reviewing existing code for more efficient development.	– Iterative code prompting (repeatedly refining and debugging code) – Potential CoT (stepwise explanation during code generation)	– Hallucinations in code: silent errors may appear correct but function incorrectly. – Legal concerns: copyright issues arise if LLMs were trained on protected repositories. – Ethical attribution: proper acknowledgment of AI-generated code is necessary to maintain transparency and intellectual honesty.	[31]

Prompt engineering has become a crucial tool in enhancing the performance of LLMs across various fields. In education, LLMs support personalized learning by generating feedback tailored to individual student needs, thereby improving learning outcomes and engagement. The iterative refinement of prompts, as seen in the feedback Copilot, showcases the importance of well-designed prompts in automating feedback processes [18], [30]. In healthcare, prompt engineering has optimized LLMs to deliver task-specific guidance, improving clinical interactions and automating administrative tasks [23], [27]. Data analysis has also benefited from prompt-driven synthetic data generation, improving dataset balance and model reliability [25]. Adversarial applications of prompt engineering, such as exploiting vulnerabilities in LLMs, have been further highlighted in Table 2 to illustrate the range of ethical concerns—from privacy risks and bias to the potential for AI misuse [29], [33]. Furthermore, disparities in AI literacy and access to prompt engineering skills highlight a growing need for equitable interventions to ensure fair use and benefit distribution of LLM technologies [21].

The growing role of prompt engineering in enhancing LLMs highlights the need for continued research and development to optimize both performance and ethical safety. Standardized guidelines for prompt engineering across various fields can help mitigate biases and improve the reliability of model outputs. Furthermore, integrating AI literacy, particularly prompt engineering, into educational curricula will empower educators and students alike, reducing the digital divide. Prompt-aligned gradient tuning emerges as a powerful strategy, allowing for the refinement of prompts to better align with the desired outputs. This mechanism highlights the nuanced control that prompt engineering affords, bridging the gap between human intention and AI execution, and underscores the importance of understanding how prompts catalyze specific responses from LLMs. Figure 2 showcase explicit examples of ineffective vs. effective prompts, each accompanied by representative outputs. In sectors like healthcare, prioritizing privacy-preserving AI approaches, such as locally deployed LLMs, is essential to address the risks associated with cloud-based models. Finally, establishing ethical frameworks for managing adversarial prompt use is critical to prevent misuse of these technologies. As prompt engineering continues to advance, ensuring its responsible application is key to maximizing its potential across different domains, while mitigating risks related to privacy, security, and bias.

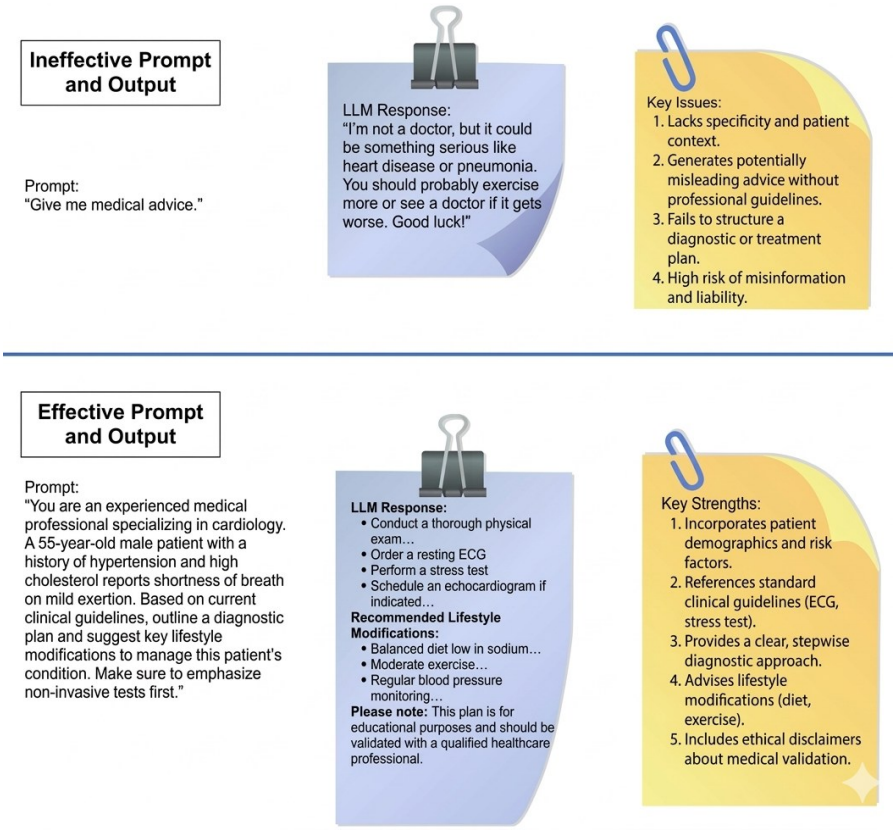

Figure 2. Simplified illustration of how effective prompting aids in obtaining the desired response in LLMs showing ineffective and effective promptings

3.2. Research question 1: performance of LLMs in scientific and engineering domains?

Prompt engineering is critical to improving the functionality of LLMs in specialized areas such as science and engineering. By carefully structuring inputs—known as "prompts"—researchers can direct LLMs to generate more accurate, contextually relevant outputs and tailor them to the specific technical demands of these fields. In scientific and engineering disciplines, where precision and technical understanding are vital, prompt engineering ensures that LLMs perform optimally, meeting the rigorous standards required. These insights also enable a comparative understanding of how different prompt strategies—such as CoT or iterative refinement—impact performance in high-stakes settings. These advancements underscore the versatility and expanding capabilities of LLMs, driven by cutting-edge research in prompt engineering. The most popular prompt techniques are summarized in Figure 3.

Figure 3. Popular prompting techniques to enhance LLM's capability in providing desired response

3.2.1. Enhancing domain-specific understanding and contextualization

A key advantage of prompt engineering is it is ability to refine LLM performance by incorporating domain-specific language and context into prompts. Scientific and engineering fields are marked by highly specialized jargon and concepts. LLMs trained on general data sets may misinterpret or generate outputs not aligned with such complex topics. Prompt engineering mitigates this by embedding detailed, relevant information within prompts, thus guiding LLMs to deliver outputs that meet the demands of specialized tasks. For instance, in materials science, researchers use specific prompts to help LLMs extract critical information from complex research papers, such as material properties, synthesis methods, and experimental data [14]. LLMs are better equipped to analyze and interpret technical literature by employing domain-specific terminology in prompts, reducing errors in subsequent research applications. This level of accuracy is crucial, as minor misinterpretations in scientific data can lead to significant consequences in research and development.

In engineering design, particularly in civil or mechanical engineering fields, prompt engineering ensures that LLMs can understand detailed specifications, design constraints, and performance metrics. Prompts tailored to include technical terms related to load-bearing capacities, structural integrity, and thermal expansion allow LLMs to generate useful design alternatives that align with specific requirements [11]. This minimizes the need for extensive manual corrections, improving workflow efficiency and accuracy. In chemical engineering, prompts can also emphasize specific variables such as reaction mechanisms and catalyst efficiencies. This helps LLMs focus on critical aspects of scientific analysis, enhancing their ability to interpret complex data sets with precision [34]. In such fields, where reliability is crucial, prompt engineering provides a significant advantage by steering LLMs toward more accurate and reliable predictions.

3.2.2. Improved reasoning through chain-of-thought prompting

One of the main challenges in using LLMs for scientific and engineering applications is their difficulty in managing multi-step reasoning tasks. These tasks often require a logical sequence of steps, and without explicit guidance, LLMs may struggle to maintain coherence. CoT prompting addresses this issue by guiding LLMs through each step of the reasoning process. By structuring prompts to mirror logical progression, CoT prompting helps LLMs maintain consistency and reduce errors.

In scientific research, CoT prompting improves LLM capabilities in hypothesis generation and experimental design by structuring the prompts according to the stages of scientific inquiry—from hypothesis formulation to data analysis [6]. This allows LLMs to assist in planning and interpreting experiments more effectively. In engineering tasks, such as circuit design and structural analysis, CoT prompting ensures that all critical variables, such as voltage, resistance, and structural stresses, are taken into account at each step of the process [35], [36]. This reduces the likelihood of the model skipping important steps, enhancing the reliability of the generated outputs.

3.2.3. Real-time data integration with retrieval-augmented generation

One limitation of traditional LLMs is their reliance on static datasets, which can quickly become outdated in fast-evolving fields like biomedical research and environmental science. RAG overcomes this by

enabling LLMs to access real-time data from external sources, integrating it into the generation process. We highlight that RAG can be configured to query domain-tailored repositories—such as specialized medical knowledge graphs or real-time environmental sensors—for more accurate and context-relevant outputs. This capability is particularly valuable in dynamic fields, where up-to-date information is crucial for accuracy. RAG combines LLM generation with a retrieval mechanism that allows access to real-time databases, improving the accuracy and relevance of outputs. For instance, in biomedical research, LLMs enhanced with RAG can access the latest clinical guidelines, ensuring that their recommendations are current and evidence-based [10]. Similarly, in environmental science, RAG allows LLMs to retrieve real-time climate data or environmental impact reports, ensuring that models are grounded in the most recent information [13]. This real-time integration is essential for decision-making in fields that rely on constantly updated data.

3.2.4. Enhancing interpretability and reducing errors

In high-stakes fields like aerospace engineering or pharmaceutical research, the interpretability and reliability of LLM outputs are critical. Structured prompting, where prompts are designed to elicit detailed explanations of the reasoning process, enhances the transparency and trustworthiness of LLM outputs. For example, in mechanical engineering, prompts can guide LLMs to not only provide solutions but also explain how they arrived at those solutions. This allows engineers to review the logic behind the output, making it easier to identify and correct errors before implementation [37]. In pharmaceutical research, where even minor errors can have serious consequences, structured prompting reduces the risk of hallucinations—incorrect but plausible-sounding information—by guiding the LLM to focus on verified data and key parameters [38]. This improves the reliability of outputs and ensures that generated content aligns with established scientific principles.

3.2.5. Automation and efficiency gains through prompt engineering

Prompt engineering also enhances automation by allowing LLMs to take over routine tasks, such as data extraction, literature review, and technical report generation. We note that iterative prompting often outperforms single-pass approach in tasks like SLR, as repeated interactions can capture nuances missed in one-step queries. This frees up researchers and engineers to focus on more complex and creative tasks. In bioinformatics, for example, prompt engineering can automate extraction of relevant data from large research databases, speeding up the analysis process [11]. Similarly, in engineering, LLMs can be guided by prompts to generate technical reports that adhere to industry standards, saving time and reducing human error [14].

3.2.6. Cross-domain applications and interdisciplinary research

One of the most exciting developments in prompt engineering is its ability to facilitate interdisciplinary research. In modern research, complex problems often require insights from multiple fields. Prompt engineering allows LLMs to function across various domains by guiding them to integrate knowledge from different disciplines. For instance, in environmental engineering, prompts can be structured to draw on climatology, ecology, and civil engineering, providing comprehensive solutions to environmental problems [13]. In biomedical engineering, prompts can guide LLMs to synthesize knowledge from biology, medicine, and mechanical engineering to develop advanced medical devices [39]. This cross-domain applicability fosters innovation by encouraging the synthesis of ideas from diverse fields.

3.3. Research question 2: key techniques in prompt engineering

Prompt engineering is essential for optimizing the outputs of LLMs, particularly in specialized fields such as science and engineering. Key techniques, including CoT prompting and RAG, have emerged as powerful methods for overcoming limitations like static datasets and complex reasoning challenges. The RAG framework operates by translating input prompts into targeted queries, which are then used to fetch relevant information from an array of sources such as search engines or knowledge graphs. CoT excels at multi-step logical reasoning while RAG ensures access to current, domain-specific data—highlighting how each technique can be chosen or combined depending on the task requirements in scientific and engineering contexts. Figure 4 illustrates this process, showcasing how queries extracted from prompts can be augmented with information retrieved from external databases to synthesize more informed and context-rich responses [40]. This section explores how these and other techniques, such as contextual prompting and active learning, contribute to refining LLM performance in generating context-specific, accurate, and reliable outputs.

3.3.1. Chain-of-thought prompting

CoT prompting enhances the ability of LLMs to handle multi-step reasoning tasks. In scientific and engineering contexts, such tasks frequently involve carefully considering multiple factors, requiring the model to follow a logical sequence. CoT prompting helps LLMs manage this complexity by breaking down problems into smaller, more manageable steps. We underscore comparing CoT with other techniques like

contextual prompting: while CoT excels at stepwise logical flow, contextual prompting ensures domain-specific accuracy. CoT prompting guides LLMs through a process similar to "thinking aloud," where the model is encouraged to articulate each step in the problem-solving process. This method improves transparency and reduces errors, especially in fields that require meticulous reasoning, such as mechanical engineering and mathematics. For example, in a calculation task, CoT prompting may ask the LLM to explain each phase of a design or analysis, significantly reducing the likelihood of errors or oversights [41]. This approach is equally valuable in symbolic reasoning tasks, such as algebraic equations or circuit design, where accuracy in each step is critical [37].

CoT prompting has proven particularly effective in hypothesis generation, experimental design, and data interpretation in scientific research. For instance, researchers working in material science may use CoT prompting to help an LLM systematically generate a hypothesis, plan an experiment, and interpret the results. By structuring the prompt to reflect the logical flow of scientific inquiry, LLMs can better provide consistent and reliable outputs grounded in scientific principles [33]. Moreover, CoT prompting addresses one of the major pitfalls of LLMs—hallucination, where the model generates plausible but incorrect information. By requiring the LLM to explain its reasoning process at every step, CoT prompting helps prevent hallucinations. This stepwise structure ensures that each conclusion is based on logical reasoning, making it easier for users to review and verify outputs. This is particularly important in high-stakes applications like medical diagnostics or aerospace engineering, where even small mistakes can have serious consequences [35].

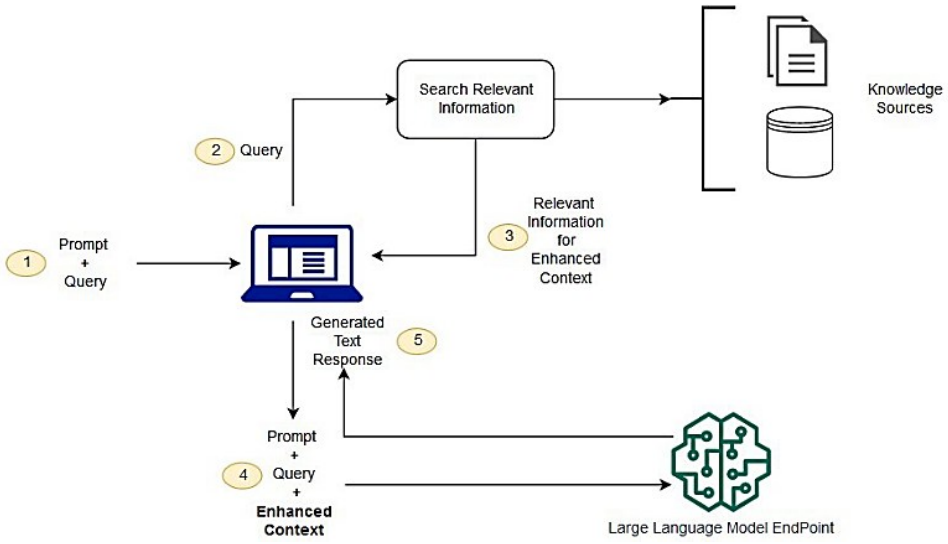

Figure 4. Illustration of using a knowledge graph as a retrieval mechanism in conjunction with LLMs to enhance response with structured external knowledge [40]

3.3.2. Retrieval-augmented generation

While CoT prompting improves reasoning capabilities, RAG addresses another fundamental limitation of LLMs: their dependence on static, pre-trained datasets. LLMs are typically trained on large, fixed datasets, which may become outdated, especially in fields with rapidly evolving knowledge bases like biomedical research and environmental science. We note that RAG excels in scenarios requiring live updates or domain-specific queries—a key advantage over CoT's focus on multi-step reasoning. RAG enables LLMs to query real-time external databases during the generation process, improving the accuracy and relevance of their outputs. RAG combines LLMs' generative abilities with a retrieval mechanism that accesses up-to-date information from external sources. For example, in biomedical applications, LLMs equipped with RAG can retrieve the latest clinical guidelines or research findings, ensuring that outputs remain current and reflective of the most recent knowledge [10]. This capability is invaluable in medical diagnostics, where outdated information could lead to incorrect treatment recommendations.

RAG allows LLMs to pull relevant data on demand in research and development fields. For instance, when summarizing research in neuroscience, a RAG-enabled LLM could access the latest studies from scientific databases, providing a more comprehensive and up-to-date analysis than a model relying solely on pre-trained knowledge [11]. This makes RAG particularly beneficial in fast-moving fields where new discoveries and data are continuously emerging. Integrating real-time data through RAG also enhances LLM outputs in engineering domains. For example, LLMs must often reference current building codes or material

specifications in civil or mechanical engineering to produce reliable design recommendations. RAG enables LLMs to access this up-to-date information, ensuring that outputs adhere to industry standards and regulations [13]. Moreover, in environmental modeling or market analysis, RAG allows LLMs to incorporate real-time data into predictive models, adapting to changing conditions and producing more relevant outputs [42].

RAG effectively bridges knowledge gaps in dynamic or highly specialized fields. As LLMs are limited by the static nature of their training datasets, their outputs may become obsolete. By incorporating real-time data retrieval, RAG ensures that LLMs can generate outputs that are both accurate and relevant, particularly in rapidly advancing areas like genomics or climate science [43], [44].

3.3.3. Contextual prompting

Contextual prompting improves LLM performance by embedding domain-specific knowledge into prompts, while CoT and RAG focus on reasoning and real-time data integration. Contextual prompts are designed to include the relevant technical language and context necessary for LLMs to interpret queries more accurately. We clarify that contextual prompting can also work in tandem with active learning—introducing domain-specific constraints over multiple iterations—to correct misunderstandings in evolving queries. This is particularly useful in fields where terminology may have different meanings depending on the context. For example, in legal research, prompts that include specific legal definitions or references to case law can help LLMs generate more accurate interpretations of statutes or contracts [45]. This reduces the risk of misinterpretation due to the model's lack of domain-specific understanding. Similarly, in biomedical engineering, contextual prompts can assist LLMs in navigating complex medical terminology, leading to more accurate diagnostic support or research outputs [7], [46]–[48].

In physics and engineering, where terms like "force" or "energy" may have multiple interpretations, contextual prompting ensures that LLMs provide the correct responses based on the specific field of application. For instance, in thermodynamics, embedding clues in the prompt about whether "heat" refers to energy transfer or a colloquial concept helps refine the LLM's output [39]. Combining contextual prompting with CoT and RAG techniques ensures that LLMs can produce highly relevant, context-sensitive outputs that align with the intricacies of specialized fields [10].

3.3.4. Active learning prompts and iterative refinement

Active learning prompts focus on iterative refinement, engaging LLMs in continuously generating and refining outputs based on user feedback. This method is valuable in areas requiring precision and continuous improvement, such as software development or aerospace engineering. It is worth noting that active learning can be combined with CoT and RAG to provide both stepwise clarity and up-to-date data during iterative cycles—an approach particularly beneficial in complex tasks like multi-phase engineering projects. In software development, for example, active learning prompts allow LLMs to generate code, which is then iteratively refined based on performance feedback. This leads to outputs that are tailored to specific functional or security requirements [49].

In aerospace design, active learning prompts help optimize design variables by prompting LLMs to adjust recommendations based on feedback from engineers or real-time data. This iterative refinement results in more reliable and optimized outputs, which are crucial in fields where precision is paramount [50]. Active learning prompts are also useful in exploratory research. By guiding LLMs to revisit and refine their responses, researchers can explore alternative hypotheses or solutions, leading to more comprehensive and well-rounded outputs [14], [51], [52].

3.4. Research question 3: ethical challenges

The integration of LLMs into science and engineering is transforming these fields by enhancing innovation and streamlining processes. However, alongside these benefits come ethical challenges that require careful attention, especially in high-stakes applications like medical diagnostics, engineering design, and legal analysis. The primary ethical concerns include bias, hallucination, and data privacy—issues that can significantly affect the quality and safety of LLM outputs. We also underscore the role of prompt engineering in mitigating these issues—by embedding fairness prompts, enforcing stepwise logical checks, or integrating secure data retrieval routines—to ensure responsible LLM deployment and prevent unintended consequences in sensitive domains.

3.4.1. Bias in LLM outputs

Bias in LLM outputs represents a critical ethical concern. LLMs are trained on vast amounts of data, often sourced from the internet and other publicly available datasets, which inherently contain biases related to gender, race, culture, and socioeconomic status. These biases can influence LLM outputs, leading to unfair or inaccurate results in fields ranging from healthcare to engineering and legal research. The root cause of bias lies in the data used to train these models. Since LLMs learn by identifying patterns in their training

data, any biases present in the data can be reflected in the model's outputs. We emphasize comparing multiple prompt strategies—for instance, fairness prompts vs. contextual prompts—to assess which approach more effectively mitigates domain-specific bias. For example, if an LLM is trained on historical medical research, it may inherit gender or ethnic biases, which could affect its diagnostic recommendations [15]. This is particularly concerning in healthcare, where biased outputs could disproportionately affect underrepresented groups [53], [54]. For instance, an LLM trained predominantly on data from male patients may provide less accurate diagnostic suggestions for female patients, especially in areas like cardiology, where symptoms can differ significantly by gender [7].

Bias is not limited to healthcare; it also extends to engineering and urban planning. In civil engineering, for example, LLMs might be used to analyze urban development projects. If the model's training data includes biased historical planning decisions, the LLM's outputs could perpetuate those biases, leading to decisions that disproportionately affect minority or disadvantaged communities [13]. Such biases can result in inequitable designs or predictions that fail to meet the needs of all groups, undermining the fairness of engineering decisions. To mitigate bias, efforts focus on curating diverse and representative datasets, although this is challenging given the scale and variability of data used to train LLMs [38]. Additionally, prompt engineering techniques are being developed to reduce bias. Fairness prompts, for instance, guide LLMs to produce more balanced outputs by explicitly instructing them to consider gender, racial, and ethnic diversity. In healthcare, prompts can be designed to ensure that the model accounts for demographic diversity when generating diagnostic recommendations [45]. However, while these strategies reduce bias, they do not eliminate it entirely, underscoring the need for ongoing research to develop more effective bias mitigation techniques.

3.4.2. Hallucination in LLM outputs

Hallucination, where an LLM generates factually incorrect or nonsensical information, is another major ethical concern. This issue is particularly problematic in fields like scientific research, engineering, and medicine, where accuracy is critical. LLMs do not "understand" the content they generate; they rely on patterns in their training data, which can lead to outputs that appear plausible but are factually incorrect or entirely fabricated. In scientific research, hallucinations can mislead users by producing inaccurate summaries of findings or even inventing non-existent studies. For instance, an LLM tasked with summarizing biomedical research might incorrectly interpret data or cite fictional studies, potentially leading researchers astray [6], [48], [54].

In medicine, hallucinated outputs can have severe consequences, such as recommending incorrect treatments or diagnoses based on fabricated clinical guidelines, jeopardizing patient safety. Similarly, hallucinations can lead to flawed technical analyses or incorrect design recommendations in engineering. For example, an LLM used in structural engineering might fabricate safety calculations or suggest untested design methods, leading to safety risks or costly design errors [14]. These risks are particularly high in industries where the reliability of outputs is critical to safety and performance, such as aerospace or civil engineering.

Techniques to mitigate hallucinations include RAG, which enables LLMs to access real-time, verified data from external databases. By grounding outputs in current, factual information, RAG significantly reduces the risk of hallucination [10]. In biomedical applications, for example, RAG allows LLMs to retrieve the latest clinical guidelines from trusted medical databases, ensuring that the model's outputs are accurate and up to date [6]. Another effective method is CoT prompting, which enhances the logical reasoning process of LLMs. CoT prompting requires LLMs to explain each step of their reasoning, making it easier for users to follow the logic behind the outputs and verify their accuracy [41], [55], [56]. In medical diagnostics, for instance, CoT prompting can guide an LLM through a structured clinical reasoning process, ensuring that each recommendation is based on sound evidence and logic [35].

3.4.3. Data privacy and security concerns

Data privacy and security are significant ethical concerns when deploying LLMs in sensitive fields like healthcare, legal research, and engineering. LLMs are trained on vast datasets, which may include personal or proprietary information, and their deployment often requires access to sensitive data such as patient health records or confidential engineering designs. This raises the risk of data breaches or misuse, which could have serious consequences for privacy and confidentiality. In healthcare, for example, LLMs used for diagnostic support or clinical decision-making often require access to patient data. This raises the possibility that sensitive health information could be inadvertently exposed, violating privacy regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States [57]. Such breaches could undermine trust in AI systems and lead to legal liabilities for healthcare providers. Similarly, in legal contexts, LLMs used to analyze case files or draft legal documents could expose confidential information, risking breaches of client-attorney privilege or intellectual property theft [38].

Several strategies are being developed to address data privacy concerns. Data anonymization is a common approach, where personally identifiable information is removed before being processed by LLMs. While this reduces the risk of exposing sensitive information, it is not foolproof, as anonymized data can sometimes be re-identified when combined with other datasets [7]. Therefore, anonymization must be complemented by strict data handling protocols, such as encryption and access controls, to ensure that only authorized users can access sensitive data. We emphasize that secure prompt engineering—where prompts are carefully designed to exclude identifying details—can further minimize privacy risks in high-stakes fields. In addition, secure deployment environments for LLMs, particularly in fields like healthcare and engineering, are critical. For example, biomedical research applications can use locally deployed LLMs in controlled environments where data privacy regulations are strictly enforced [6]. Auditing mechanisms that track how data is accessed and used by LLMs can also help ensure accountability and transparency, quickly identifying any misuse of data [10].

3.4.4. Accountability and ethical transparency

Beyond the specific challenges of bias, hallucination, and data privacy, the broader issue of accountability and transparency in LLM deployment also presents an ethical challenge. As LLMs are increasingly integrated into decision-making processes, determining who is responsible for the outcomes they generate becomes more complex. This is particularly important in high-stakes fields like medical diagnostics or engineering, where errors can lead to serious consequences.

One challenge is the black-box nature of LLMs, which operate through complex neural networks that make it difficult to explain how they arrive at certain decisions. This lack of transparency complicates accountability, as it may be unclear who is responsible for an incorrect decision—whether it is the developers, users, or the model itself [39]. In fields like aerospace engineering, where safety is paramount, the inability to fully understand the decision-making process of an LLM could undermine trust in it is outputs.

To address these concerns, there is a growing need for ethical guidelines and frameworks to govern the responsible use of LLMs. These guidelines should establish transparency, accountability, and fairness standards in LLM deployment. We also recommend the adoption of explainable prompt engineering (XPE), wherein prompts themselves include justification requests (e.g., "explain how you derived this") to augment explainable artificial intelligence (XAI) approaches and make the decision pathway more explicit. Moreover, developing XAI techniques is essential to making LLM decision-making processes more understandable. XAI tools would enable users to audit and verify LLM outputs, ensuring that the results generated by these models can be trusted and validated, particularly in critical contexts like engineering and healthcare [13].

3.5. Future research paradigm

Integrating LLMs into scientific and engineering domains holds vast potential, but key ethical challenges must be addressed to ensure their responsible use. Future research should focus on mitigating issues such as bias, hallucinations, and data privacy, while improving transparency and accountability. We emphasize that prompt engineering—including fairness prompts, CoT, and retrieval-augmented methods—can serve as a central scaffold for addressing these challenges by guiding LLMs toward more ethically aligned and context-specific outputs.

3.5.1. Addressing bias in LLMs

Bias remains a persistent challenge in LLM deployment due to the inherent biases in the datasets used for training. Although fairness-focused prompt engineering has shown some success, more robust methods are needed. Future research should also explore comparative analyses of CoT vs. contextual vs. fairness prompts in mitigating bias across domains, thereby identifying domain-specific best practices. Future research should develop comprehensive bias detection and correction frameworks, targeting the data preprocessing and model training stages [15]. Bias auditing tools could automatically identify and quantify biases in LLM outputs across domains, helping refine training data or prompts for more equitable results. Reinforcement learning could be combined with contextual prompting to allow LLMs to evolve toward generating less biased outputs over time, further enhancing fairness [58].

3.5.2. Mitigating hallucinations

Hallucination, where LLMs generate incorrect or nonsensical information, poses significant risks, especially in high-stakes fields like medical research. Techniques like CoT prompting and RAG have helped reduce hallucinations by grounding outputs in logical steps or real-time data [59], [60]. Future research could explore iterative prompt engineering strategies that combine active learning with real-time retrieval, enabling LLMs to continuously self-check outputs for accuracy and consistency. Additionally, future work should focus on advancing these techniques to make them scalable and more robust. A promising area for research is hybrid models that combine LLMs with symbolic reasoning systems for real-time fact-checking.

Such models could cross-reference outputs with trusted knowledge bases, reducing the likelihood of hallucinations [6]. Enhancing explainability through XAI frameworks is also crucial. These frameworks would allow users to visualize LLMs' decision-making processes, increasing transparency, particularly in fields where auditability is essential [35].

3.5.3. Enhancing data privacy and security

Data privacy is a critical concern in fields such as healthcare and legal research, where LLMs handle sensitive information. Future research must focus on developing privacy-preserving AI technologies like federated learning, which trains LLMs on decentralized data without transferring sensitive information to a central repository [42]. Differential privacy mechanisms should also be integrated into LLM workflows to ensure that outputs do not reveal details about individuals whose data was used for training [61]. Data anonymization techniques must be refined to prevent re-identification risks, particularly in sensitive fields like biomedical research. Additionally, encrypted LLM systems and secure data-handling protocols must be developed to ensure the confidentiality of interactions, particularly in legal and medical contexts [45].

3.5.4. Accountability and ethical frameworks

The black-box nature of LLMs presents challenges in determining accountability, especially when outputs lead to harmful consequences, such as faulty engineering designs or incorrect medical diagnoses. Future research on XPE could supplement existing XAI methods by requiring LLMs to generate structured rationale for outputs—thus making accountability more transparent. Future research should focus on developing ethical frameworks that clarify responsibility when LLMs are involved in decision-making [39]. These frameworks should address transparency, accountability, and fairness in LLM deployment, alongside regulatory guidelines that enforce adherence to ethical principles [62]. XAI tools should also be created to help users audit LLM decision-making processes, ensuring outputs can be verified in high-stakes environments like medical diagnostics and engineering design [35].

3.5.5. Interdisciplinary and cross-domain research

LLMs are increasingly used across multiple domains, but interdisciplinary collaboration is essential for handling complex global challenges, such as climate change. Future research should aim to improve LLMs' ability to synthesize information from diverse fields like ecology, engineering, and economics, providing holistic solutions [13]. Refining interdisciplinary prompt engineering techniques will ensure that LLMs maintain contextual relevance and avoid introducing biases when integrating data from various domains [11]. These efforts will make LLMs more effective in interdisciplinary research, enabling breakthroughs in fields like sustainable engineering and genomics [63].

4. CONCLUSION

This review has explored the transformative potential of LLMs in enhancing scientific and engineering applications, focusing on the ethical challenges and the key techniques that can improve their accuracy, relevance, and usability. Techniques such as CoT prompting and RAG have been highlighted as pivotal in reducing errors and enhancing the reasoning capabilities of LLMs, particularly in domains that require complex problem-solving and real-time data integration. Additionally, contextual and active learning prompts have shown significant promise in tailoring LLMs for domain-specific tasks, improving their precision and adaptability. However, despite these advancements, LLM deployment is not without its challenges. Bias remains a critical issue, particularly in healthcare, legal research, and engineering, where biased outputs can lead to unfair or inaccurate results. Similarly, the risk of hallucination, where models generate incorrect or misleading information, poses significant concerns in high-stakes fields such as medical diagnostics and engineering design. Data privacy and security issues, especially when handling sensitive information, further complicate the responsible use of LLMs in these domains. The future research paradigm must focus on developing more robust bias mitigation strategies, improving the explainability of LLM outputs, and advancing privacy-preserving AI techniques. There is also a pressing need for clearer accountability frameworks to ensure that LLM-generated decisions are transparent and justifiable, especially in critical sectors. Additionally, the development of interdisciplinary frameworks and cross-domain knowledge integration techniques will be essential for unlocking the full potential of LLMs in addressing complex, global challenges. We further highlight that prompt engineering—particularly methods like fairness prompts, explainable prompts, and iterative refinement—can serve as cornerstones for ethical innovation, guiding LLMs to operate responsibly across diverse scientific and engineering contexts. In conclusion, while LLMs have the potential to revolutionize scientific and engineering research, their success depends on continuous innovation, ethical oversight, and collaboration across disciplines. By addressing these challenges, the scientific and engineering communities can ensure that LLMs are used responsibly and effectively, driving innovation while maintaining the highest ethical standards.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Izzul Fatawi	✓	✓		✓			✓			✓		✓	✓	✓
Muhammad Roil Bilad		✓	✓			✓		✓	✓	✓	✓	✓		
Muhammad Asy'ari	✓		✓	✓	✓	✓		✓	✓	✓	✓		✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The authors confirm that the data supporting the findings of this study are available within the article.

REFERENCES

[1] A. Chowdhery *et al.*, "PaLM: scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 1, pp. 11324–11436, 2023.
[2] T. B. Brown *et al.*, "Language models are few-shot learners," *34th Conference on Neural Information Processing Systems (NeurIPS 2020),* Jul. 2020, pp. 1877–1901.
[3] A. Rajasekharan, Y. Zeng, P. Padalkar, and G. Gupta, "Reliable natural language understanding with large language models and answer set programming," *International Conference on Logic Programming*, pp. 274–287, Aug. 2023, doi: 10.4204/EPTCS.385.27.
[4] J. Wei, S. Kim, H. Jung, and Y.-H. Kim, "Leveraging large language models to power chatbots for collecting user self-reported data," in *Proceedings of the ACM on Human-Computer Interaction*, Apr. 2024, pp. 1–35, doi: 10.1145/3637364.
[5] K. S. Jasmine, "Unlocking the power of prompt engineering," in *Advances in Educational Technologies and Instructional Design*, 2024, pp. 411–432, doi: 10.4018/979-8-3693-1351-0.ch020.
[6] K. Shah *et al.*, "Large language model prompting techniques for advancement in clinical medicine," *Journal of Clinical Medicine*, vol. 13, no. 17, Aug. 2024, doi: 10.3390/jcm13175101.
[7] N. Marques, R. R. Silva, and J. Bernardino, "Using ChatGPT in software requirements engineering: a comprehensive review," *Future Internet*, vol. 16, no. 6, May 2024, doi: 10.3390/fi16060180.
[8] J. Gu, V. Tresp, and Y. Qin, "Are vision transformers robust to patch perturbations?," in *Computer Vision – ECCV 2022*, 2022, pp. 404–421, doi: 10.1007/978-3-031-19775-8_24.
[9] B. Zhu, Y. Niu, Y. Han, Y. Wu, and H. Zhang, "Prompt-aligned gradient for prompt tuning," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2023, pp. 15613–15623, doi: 10.1109/ICCV51070.2023.01435.
[10] J. Miao, C. Thongprayoon, S. Suppadungsuk, O. A. G. Valencia, and W. Cheungpasitporn, "Integrating retrieval-augmented generation with large language models in nephrology: advancing practical applications," *Medicina*, vol. 60, no. 3, Mar. 2024, doi: 10.3390/medicina60030445.
[11] M. P. Polak and D. Morgan, "Extracting accurate materials data from research papers with conversational language models and prompt engineering," *Nature Communications*, vol. 15, no. 1, Feb. 2024, doi: 10.1038/s41467-024-45914-8.
[12] S. Ali *et al.*, "General purpose large language models match human performance on gastroenterology board exam self-assessments." *medRxiv,* Sep. 25, 2023. doi: 10.1101/2023.09.21.23295918.
[13] D. Huang, C. Yan, Q. Li, and X. Peng, "From large language models to large multimodal models: a literature review," *Applied Sciences*, vol. 14, no. 12, Jun. 2024, doi: 10.3390/app14125068.
[14] K. S. Cheung, "Real estate insights unleashing the potential of ChatGPT in property valuation reports: the 'red book' compliance chain-of-thought (CoT) prompt engineering," *Journal of Property Investment & Finance*, vol. 42, no. 2, pp. 200–206, Apr. 2024, doi: 10.1108/JPIF-06-2023-0053.
[15] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of GPT-4 on medical challenge problems," Apr. 2023, *arXiv: 2303.13375*.
[16] M. D. H. Wirzal, N. S. A. Halim, N. A. H. Md Nordin, and M. A. Bustam, "Metacognition in science learning: bibliometric analysis of last two decades," *Jurnal Penelitian dan Pengkajian Ilmu Pendidikan: e-Saintika*, vol. 6, no. 1, pp. 43–60, Mar. 2022, doi: 10.36312/esaintika.v6i1.665.
[17] M. R. Bilad, L. N. Yaqin, and S. Zubaidah, "Recent progress in the use of artificial intelligence tools in education," *Jurnal Penelitian dan Pengkajian Ilmu Pendidikan: e-Saintika*, vol. 7, no. 3, pp. 279–315, Oct. 2023, doi: 10.36312/esaintika.v7i3.1377.
[18] J. Meyer *et al.*, "Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions," *Computers and Education: Artificial Intelligence*, vol. 6, Jun. 2024, doi: 10.1016/j.caeai.2023.100199.

[19] N. Alfirević, D. G. Praničević, and M. Mabić, "Custom-trained large language models as open educational resources: an exploratory research of a business management educational chatbot in Croatia and Bosnia and Herzegovina," *Sustainability*, vol. 16, no. 12, Jun. 2024, doi: 10.3390/su16124929.
[20] Y. Walter, "Embracing the future of artificial intelligence in the classroom: the relevance of ai literacy, prompt engineering, and critical thinking in modern education," *International Journal of Educational Technology in Higher Education*, vol. 21, no. 1, Feb. 2024, doi: 10.1186/s41239-024-00448-3.
[21] N. Knoth, A. Tolzin, A. Janson, and J. M. Leimeister, "AI literacy and its implications for prompt engineering strategies," *Computers and Education: Artificial Intelligence*, vol. 6, Jun. 2024, doi: 10.1016/j.caeai.2024.100225.
[22] J. D. V.-Henao, C. J. F.-Cardona, and L. C.-Higuita, "Prompt engineering: a methodology for optimizing interactions with AI-language models in the field of engineering," *DYNA*, vol. 90, no. 230, pp. 9–17, Nov. 2023, doi: 10.15446/dyna.v90n230.111700.
[23] T. F. Tan *et al.*, "Generative artificial intelligence through ChatGPT and other large language models in ophthalmology," *Ophthalmology Science*, vol. 3, no. 4, Dec. 2023, doi: 10.1016/j.xops.2023.100394.
[24] A. P.-Martín, C. G.-Mateo, L. D.-Fernández, and M. del C. L.-Pérez, "Ethical challenges in the development of virtual assistants powered by large language models," *Electronics*, vol. 12, no. 14, Jul. 2023, doi: 10.3390/electronics12143170.
[25] M. Kochanek *et al.*, "Improving training dataset balance with ChatGPT prompt engineering," *Electronics*, vol. 13, no. 12, Jun. 2024, doi: 10.3390/electronics13122255.
[26] I. H. Sarker, "LLM potentiality and awareness: a position paper from the perspective of trustworthy and responsible ai modeling," *Discover Artificial Intelligence*, vol. 4, no. 1, May 2024, doi: 10.1007/s44163-024-00129-0.
[27] I. C. Wiest *et al.*, "Privacy-preserving large language models for structured medical information retrieval," *npj Digital Medicine*, vol. 7, no. 1, Sep. 2024, doi: 10.1038/s41746-024-01233-2.
[28] F. Trad and A. Chehab, "Prompt engineering or fine-tuning? a case study on phishing detection with large language models," *Machine Learning and Knowledge Extraction*, vol. 6, no. 1, pp. 367–384, Feb. 2024, doi: 10.3390/make6010018.
[29] R. Lapid, R. Langberg, and M. Sipper, "Open sesame! universal black-box jailbreaking of large language models," *Applied Sciences*, vol. 14, no. 16, p. 7150, Aug. 2024, doi: 10.3390/app14167150.
[30] S. Pozdniakov *et al.*, "Large language models meet user interfaces: the case of provisioning feedback," *Computers and Education: Artificial Intelligence*, vol. 7, Dec. 2024, doi: 10.1016/j.caeai.2024.100289.
[31] C. F. Atkinson, "ChatGPT and computational-based research: benefits, drawbacks, and machine learning applications," *Discover Artificial Intelligence*, vol. 3, no. 1, Dec. 2023, doi: 10.1007/s44163-023-00091-3.
[32] X. Luo *et al.*, "Potential roles of large language models in the production of systematic reviews and meta-analyses," *Journal of Medical Internet Research*, vol. 26, Jun. 2024, doi: 10.2196/56780.
[33] I. A. Zahid *et al.*, "Unmasking large language models by means of OpenAI GPT-4 and Google AI: a deep instruction-based analysis," *Intelligent Systems with Applications*, vol. 23, Sep. 2024, doi: 10.1016/j.iswa.2024.200431.
[34] B. Lund, "The prompt engineering librarian," *Library Hi Tech News*, vol. 40, no. 8, pp. 6–8, Nov. 2023, doi: 10.1108/LHTN-10-2023-0189.
[35] B. Zhang, "Prompt engineers or librarians? an exploration," *Medical Reference Services Quarterly*, vol. 42, no. 4, pp. 381–386, Oct. 2023, doi: 10.1080/02763869.2023.2250680.
[36] H. Wang, J. Li, H. Wu, E. Hovy, and Y. Sun, "Pre-trained language models and their applications," *Engineering*, vol. 25, pp. 51–65, Jun. 2023, doi: 10.1016/j.eng.2022.04.024.
[37] N. Ranade, M. Saravia, and A. Johri, "Using rhetorical strategies to design prompts: a human-in-the-loop approach to make AI useful," *AI and SOCIETY*, vol. 40, no. 2, pp. 711–732, Feb. 2025, doi: 10.1007/s00146-024-01905-3.
[38] B. Schulte, "Considerations for prompting large language models," *JAMA Oncology*, vol. 10, no. 4, Apr. 2024, doi: 10.1001/jamaoncol.2023.6963.
[39] M. Johnson, A. P. Ribeiro, T. M. Drew, and P. N. R. Pereira, "Generative ai use in dental education: efficient exam item writing," *Journal of Dental Education*, vol. 87, no. S3, pp. 1865–1866, Dec. 2023, doi: 10.1002/jdd.13294.
[40] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying large language models and knowledge graphs: a roadmap," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 7, pp. 3580–3599, Jul. 2024, doi: 10.1109/TKDE.2024.3352100.
[41] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems*, Nov. 2022, pp. 22199-22213, doi: 10.5555/3600270.3601883.
[42] E. Chen, R. Huang, H.-S. Chen, Y.-H. Tseng, and L.-Y. Li, "GPTutor: a ChatGPT-powered programming tool for code explanation," 2023, pp. 321–327. doi: 10.1007/978-3-031-36336-8_50.
[43] C. Njeh, H. Nakouri, and F. Jaafar, "Enhancing rag-retrieval to improve llms robustness and resilience to hallucinations," *Hybrid Artificial Intelligent Systems* pp. 201–213, 2025, doi: 10.1007/978-3-031-74186-9_17.
[44] F. Stella, C. D. Santina, and J. Hughes, "How can llms transform the robotic design process?," *Nature Machine Intelligence*, vol. 5, no. 6, pp. 561–564, Jun. 2023, doi: 10.1038/s42256-023-00669-7.
[45] J. J. Nay *et al.*, "Large language models as tax attorneys: a case study in legal capabilities emergence," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 382, no. 2270, Apr. 2024, doi: 10.1098/rsta.2023.0159.
[46] J. Zaghir, M. Naguib, M. Bjelogrlic, A. Névéol, X. Tannier, and C. Lovis, "Prompt engineering paradigms for medical applications: scoping review," *Journal of Medical Internet Research*, vol. 26, Sep. 2024, doi: 10.2196/60501.
[47] T. Savage, A. Nayak, R. Gallo, E. Rangan, and J. H. Chen, "Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine," *npj Digital Medicine*, vol. 7, no. 1, Jan. 2024, doi: 10.1038/s41746-024-01010-1.
[48] X. Meng *et al.*, "The application of large language models in medicine: a scoping review," *iScience*, vol. 27, no. 5, May 2024, doi: 10.1016/j.isci.2024.109713.
[49] Y. Chen, G. Yang, D. Wang, and D. Li, "Eliciting knowledge from language models with automatically generated continuous prompts," *Expert Systems with Applications*, vol. 239, Apr. 2024, doi: 10.1016/j.eswa.2023.122327.
[50] R. Wahid, J. Mero, and P. Ritala, "Editorial: written by ChatGPT, illustrated by Midjourney: generative AI for content marketing," *Asia Pacific Journal of Marketing and Logistics*, vol. 35, no. 8, pp. 1813–1822, Nov. 2023, doi: 10.1108/APJML-10-2023-994.
[51] Y. Li *et al.*, "Can large language models write reflectively," *Computers and Education: Artificial Intelligence*, vol. 4, 2023, doi: 10.1016/j.caeai.2023.100140.
[52] D. K. Kanbach, L. Heiduk, G. Blueher, M. Schreiter, and A. Lahmann, "The GenAI is out of the bottle: generative artificial intelligence from a business model innovation perspective," *Review of Managerial Science*, vol. 18, no. 4, pp. 1189–1220, Apr. 2024, doi: 10.1007/s11846-023-00696-z.
[53] A. A.-alrazaq *et al.*, "Large language models in medical education: opportunities, challenges, and future directions," *JMIR Medical Education*, vol. 9, Jun. 2023, doi: 10.2196/48291.

[54] C. Peng *et al.*, "A study of generative large language model for medical research and healthcare," *npj Digital Medicine*, vol. 6, no. 1, Nov. 2023, doi: 10.1038/s41746-023-00958-w.
[55] Y. K. Dwivedi *et al.*, "Opinion paper: 'so what if ChatGPT wrote it?' multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy," *International Journal of Information Management*, vol. 71, Aug. 2023, doi: 10.1016/j.ijinfomgt.2023.102642.
[56] K. Singhal *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, Aug. 2023, doi: 10.1038/s41586-023-06291-2.
[57] S. Chen, G. K. Savova, and D. S. Bitterman, "Considerations for prompting large language models—reply," *JAMA Oncology*, vol. 10, no. 4, Apr. 2024, doi: 10.1001/jamaoncol.2023.6966.
[58] X. Li, S. Wang, S. Zeng, Y. Wu, and Y. Yang, "A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges," *Vicinagearth*, vol. 1, no. 1, Oct. 2024, doi: 10.1007/s44336-024-00009-2.
[59] Y. Zhao, H. Cao, X. Zhao, and Z. Ou, "An empirical study of retrieval augmented generation with chain-of-thought," *ACM Transactions on Software Engineering and Methodology*, vol. 34, no. 7, Aug. 2025, doi: 10.1145/3717061.
[60] M. Alkhalaf, P. Yu, M. Yin, and C. Deng, "Applying generative ai with retrieval augmented generation to summarize and extract key clinical information from electronic health records," *Journal of Biomedical Informatics*, vol. 156, Aug. 2024, doi: 10.1016/j.jbi.2024.104662.
[61] Y. Wang, Q. Wang, L. Zhao, and C. Wang, "Differential privacy in deep learning: privacy and beyond," *Future Generation Computer Systems*, vol. 148, pp. 408–424, Nov. 2023, doi: 10.1016/j.future.2023.06.010.
[62] N. Balasubramaniam, M. Kauppinen, A. Rannisto, K. Hiekkanen, and S. Kujala, "Transparency and explainability of AI systems: from ethical guidelines to requirements," *Information and Software Technology*, vol. 159, Jul. 2023, doi: 10.1016/j.infsof.2023.107197.
[63] G. Marvin, N. Hellen, D. Jjingo, and J. Nakatumba-Nabende, "Prompt engineering in large language models," in *Data Intelligence and Cognitive Informatics*, 2024, pp. 387–402. doi: 10.1007/978-981-99-7962-2_30.

BIOGRAPHIES OF AUTHORS

Izzul Fatawi is currently a lecturer at Universitas Terbuka, Indonesia, within the Faculty of Education and Teacher Training. He has also served as a lecturer at the Islamic Institute of Nurul Hakim. He holds a Ph.D. in Education from the State University of Malang and has a master's degree from the Islamic Institute of Sunan Ampel. His research areas span online learning environments, learning analytics, e-learning systems, and educational technologies. He has authored multiple publications, including works on fuzzy logic, student engagement, and learning outcomes, with a focus on improving education through technology. He is especially interested in using learning analytics to enhance interaction and personalized learning experiences. He can be contacted at email: izzul.official@ecampus.ut.ac.id.

Muhammad Roil Bilad is a highly accomplished researcher and academic, holding a Ph.D. in Bioprocess Technology from Leuven University, Belgium. He also has a master's degree in Chemical Engineering from Universiti Teknologi PETRONAS, Malaysia, and a bachelor's in Chemical Engineering from Institut Teknologi Bandung, Indonesia. He has extensive teaching and research experience, currently serving as associate professor at Universiti Brunei Darussalam. He has also worked at leading institutions in Malaysia, Singapore, and the UAE. His prolific research career includes 307 publications, an impressive h-index of 45, and over 7,000 citations, focusing on chemical engineering and related fields. His contributions are well-recognized across various international platforms. He can be contacted at email: roil.bilad@ubd.edu.bn.

Muhammad Asy'ari is a lecturer at Universitas Pendidikan Mandalika in Mataram, Indonesia, specializing in science education. He earned his master's degree in Science Education from Universitas Negeri Surabaya and a bachelor's degree in Physics Education from Institut Keguruan dan Ilmu Pendidikan Mataram. His research focuses on enhancing critical thinking, metacognitive skills, and science process skills in prospective teachers. He has published extensively on topics related to inquiry-based learning, metacognition, and the development of science teaching materials. He has a strong interest in integrating innovative learning models to improve student outcomes, particularly in science education. He can be contacted at email: muhammadasyari@undikma.ac.id.