Ethical Concerns of Generative AI and Mitigation Strategies: A Systematic Mapping Study

Yutan Huang[a], Chetan Arora[a], Wen Cheng Huong[a], Tanjila Kanij[b], Anuradha Madugalla[c], John Grundy[a]

[a]*Faculty of Information Technology, Monash University, Clayton, Victoria, Australia*
[b]*School of Science, Computing and Engineering Technologies, Swinburne University, Hawthorn, Victoria, Australia*
[c]*School of Information Technology, Deakin University, Burwood, Victoria, Australia*

Abstract

Generative AI technologies, particularly Large Language Models (LLMs), have transformed numerous domains by enhancing convenience and efficiency in information retrieval, content generation, and decision-making processes. However, deploying LLMs also presents diverse ethical challenges, and their mitigation strategies remain complex and domain-dependent. This paper aims to identify and categorise the key ethical concerns associated with using LLMs, examine existing mitigation strategies, and assess the outstanding challenges in implementing these strategies across various domains. We conducted a systematic mapping study, reviewing 39 studies that discuss ethical concerns and mitigation strategies related to LLMs. We analysed these ethical concerns using five ethical dimensions we extracted from various existing guidelines and frameworks, along with an analysis of mitigation strategies and implementation challenges. Our findings reveal that ethical concerns in LLMs are multi-dimensional and context-dependent. While proposed mitigation strategies address some of these concerns, significant challenges still remain. Our results highlight that ethical issues often hinder the practical implementation of mitigation strategies, particularly in high-stakes areas such as healthcare and public governance. Existing frameworks are often inflexible, failing to accommodate evolving societal expectations and diverse contexts.

Keywords:
Generative AI, AI Ethics, Large Language Models (LLMs), Systematic Mapping Study

1. Introduction

The evolution of Generative AI (GenAI), particularly Large Language Models (LLMs), has seen remarkable advancements since 2020 with the introduction of models like ChatGPT and Bard. LLMs have revolutionised tasks such as writing assistance, code generation, and customer support automation by leveraging vast amounts of data to generate coherent, contextually relevant natural language (NL) responses [1, 2]. As a subset of GenAI, LLMs go beyond traditional AI techniques, which focus primarily on analysing existing data. LLMs are capable of generating text, images, and music that mimic human creativity based on prompts provided by humans [3, 4]. This capability is powered by advancements in neural network architectures, especially transformers, which enable LLMs to learn the nuances of human language and produce semantically accurate content [5]. LLMs, such as OpenAI's GPT-4 and Google's Gemini, utilise transformer architecture to understand and generate human-like text. GPT-4, for instance, employs a transformer-decoder architecture with billions of parameters, enabling it to generate detailed, contextually relevant text across various topics [6].

LLMs have been used to address numerous challenges, offering innovative solutions across sectors such as healthcare, education, and finance. It has the potential to bring efficiency into these areas [7]. However, as a rapidly emerging technology, LLMs currently operate under limited regulation and oversight, raising significant ethical concerns [8]. Without mature, robust guidelines, these potent tools risk misuse or improper application. For instance, there have been cases where AI-generated content has perpetuated biases or disseminated misinformation, underscoring the critical need for comprehensive ethical guidelines [9–12]. Although ethics is a broad concept that varies across contexts and fields, understanding the ethics of LLMs is becoming increasingly important for guiding their responsible use and development.

To effectively mitigate these risks and harness the full potential of LLMs, it is crucial to understand the existing ethical concerns and the strategies proposed to address them. To this end, we present our systematic mapping study (SMS) aimed at identifying primary studies that have investigated the key ethical challenges associated with LLM use and analysing these studies to provide insights for developing a more comprehensive ethical framework to guide their responsible use. We observed that all the articles provided their understanding of the ethical concerns and categorised these concerns into specific ethical dimensions. The ethical dimensions were categorised based on existing ethical guidelines and regulatory frameworks. These ethical dimensions refer to topics that group similar ethical issues arising within the field, e.g., privacy and bias. We found that, while all the articles proposed differ-

ent strategies to address these ethical concerns, most were conceptual and lacked evaluation of their effectiveness in resolving ethical issues arising from generative AI.

We followed Kitchenham et al.'s [13] six-step protocol for systematic reviews: defining research questions, developing a search strategy, selecting studies with clear inclusion and exclusion criteria, assessing study quality, extracting data, and synthesizing results for performing systematic reviews. In parallel, we adopted Peterson et al.'s [14] systematic mapping guidelines, which emphasise designing a classification schema, mapping publications into that and identifying research trends and gaps.

We note that several recent reviews have explored AI ethics in different domains. However, none of them cover a cross-domain, methodologically mapped approach that includes both peer-reviewed studies and industry/governmental guidelines, as well as the empirical evaluation status of the proposed mitigation strategy and implementation challenges. For example, Li et al. [15] systematically mapped ethical concerns and mitigation strategies for AI in healthcare, highlighting privacy, transparency, and accountability issues in clinical decision support systems; however, their scope was confined to medical settings. On the other hand, Sánchez et al. [11] examined ethical dilemmas in AI-driven urban planning, focusing only on governance and public safety. Morley et al. [16] reviewed publicly available AI ethics tools and methods for translating high-level principles into practice, yet they did not evaluate their effectiveness in real-world settings. Atlam et al. [17] have provided a high-level mapping of 127 ethics studies; they did not assess the empirical evaluation status or the implementation challenges of the studies they identified. By contrast, our study represents a cross-domain mapping that (a) involves both peer-reviewed studies and industry/governmental guidelines, (b) assesses the empirical evaluation outcomes of each mitigation strategy, and (c) identifies the implementation challenges of the strategies in real-world settings. We answer the research questions (RQs) noted below in our study. In our RQs and throughout the paper, we refer to the term *ethical dimensions*. We use the term ethical dimensions to refer to broad, higher-order constructs that group closely related ethical concerns. An ethical dimension is therefore a conceptual lens, rather than a single measurable variable, through which multiple concrete issues can be examined and compared. Specifically, we consider Safety, Privacy, Transparency, Bias and Accountability as the core dimensions based on our review of academic studies and existing guidelines.

RQ1. What are the ethical dimensions defined in the use of generative AI across various fields? In RQ1, we wanted to identify key ethical dimensions associated with deploying and creating LLMs by analysing relevant literature. These ethical dimensions were then mapped against existing ethical guidelines and regulatory frameworks to assess whether they are acknowledged within these standards.

RQ2. What strategies are used or proposed to address the ethical concerns of using generative AI across different fields? Having identified specific ethical dimensions and issues in RQ1, we then examined mitigation strategies mentioned in the analysed primary studies used to address these ethical concerns across different dimensions.

RQ3. What are the challenges when implementing the strategies? We reviewed any reported challenges associated with implementing the identified mitigation strategies, detailing specific limitations and obstacles encountered in practice.

The main contributions of this mapping study include:

- We identified 39 primary studies that address the ethical concerns of using generative AI. All the studies identified various ethical concerns across specific ethical dimensions and provided strategies to address them.
- Of these 39 papers, 13 conducted some form of empirical evaluation to either investigate or validate a proposed strategy. The remaining 26 papers are conceptual papers that propose strategies but do not conduct empirical evaluations, or in some cases, empirical evaluation is not feasible.
- We identified the most prominent ethical dimensions where ethical concerns are concentrated, as well as those that require more focused attention.
- We identified a set of key research directions to address ethical issues arising from the use of LLMs in practice.

The rest of the paper is structured as follows: Section 2 provides the background and related work on ethical concerns of the use of generative AI. Section 3 details our methodology for conducting a literature search and selection process on ethical considerations of generative AI. Section 4 reports the results from the selected primary studies. Section 5 discusses key results and summaries. Section 6 addresses threats to validity. Section 7 concludes our study.

2. Background

2.1. Terminology

Ethics is the discipline (or philosophy) that deals with conduct and questions of morality, exploring what is right and wrong, as well as key principles that govern human behaviors [18]. This exploration includes normative ethics, which seeks to

establish general principles for how people should act; applied ethics, which addresses concrete ethical issues in various fields; and metaethics, which involves analysing the nature of moral judgments and the underlying assumptions of ethical theories [19, 20].

AI ethics is a multidisciplinary field that examines the ethical, legal, and social implications of AI systems [21]. It involves developing frameworks, principles, and guidelines to ensure that AI technologies are designed, implemented, and used to align with societal values, human rights, and ethical standards [22]. AI ethics is heavily grounded in normative ethics, as it seeks to establish principles and guidelines for designing and using AI. Normative questions in AI ethics include what constitutes fair AI, ensuring AI respects human rights, and developers' and users' responsibility in mitigating AI's potential harms [23]. AI ethics is also primarily a form of applied ethics, as it directly addresses practical ethical issues in real-world contexts, including the ethical deployment of AI in areas such as healthcare and law enforcement, which involve specific scenarios and case studies [24]. Metaethical questions are also addressed in the context of AI ethics, including the debate over whether AI systems can be considered moral agents and what ethical responsibilities should be assigned to them [25]. In other words, AI ethics applies ethical (moral) frameworks to considerations such as the responsibilities of developers and deployers toward affected stakeholders, the harms and benefits that should be accounted for in algorithmic decision making, and the values in practice [26].

Ethical dimensions are the ethical considerations that arise when developing, deploying, and utilising LLMs, informed by existing ethical guidelines and regulatory frameworks. When using LLMs, various ethical dimensions may need to be addressed. Mitigation strategies are action-oriented, specific, and actionable approaches designed or proposed by the authors to address and resolve ethical issues directly; they can be evaluated. Recommendations are guidance-oriented, broader suggestions or guidelines proposed by authors that aim to influence future actions or policies regarding generative AI ethics; they are not usually subject to empirical testing or short-term evaluation. In our analysis of primary studies in this paper, we identify whether the strategies mentioned in the studies have been evaluated in practice. A strategy is evaluated if it has undergone some form of empirical testing, validation, or assessment to determine its effectiveness. If the strategy is theoretical or conceptual, based on logical reasoning, best practices, or expert opinion without verification of actual impact, it is not evaluated.

2.2. Evolution and advancements of Generative AI and LLMs

AI ethics have been discussed since the 1950s, and the discussion related to LLMs began in 2018, with various scholars and institutions proposing frameworks to ad-

dress the ethical implications of AI technologies [27]. However, it was not until after 2020 that authorities began implementing concrete regulatory and policy frameworks to govern these rapidly emerging and adopted AI technologies. The increasing complexity and capabilities of LLMs, as well as incidents of data breaches and misuse, underscore the urgent need for robust ethical guidelines and regulatory oversight [28]. For instance, the mass data leakage from Chinese applications in 2020 highlighted significant privacy concerns, prompting a global call for stronger data protection measures [29]. Below, we review the evolution of generative AI and LLMs, the ethical concerns they raise, and the regulatory responses from major jurisdictions and industry leaders.

Initially, AI research focused on rule-based systems and basic machine learning algorithms, but the development of Generative Adversarial Networks (GANs) marked a breakthrough [30]. GANs enabled AI to create realistic images, videos, and text, laying the foundation for Generative AI and opening the way for further advancements. The introduction of the Transformer model in 2017 revolutionised Natural Language Processing (NLP), leading to the creation of powerful LLM models like GPT, which advanced the generative AI field to a large extent [31].

2.2.1. OpenAI's GPT Series

These models leverage large amounts of data to perform complex language processing tasks with proficiency. GPT-3, introduced in 2020, comprised 175 billion parameters, was well known for its size and scope, and has been used across a wide range of applications, from creative writing to code generation and translation, question answering with fine-tuning, demonstrating its learning capabilities where the model can learn from a few examples of a given task [32].

The successor to GPT-3, GPT-4, advances these capabilities by incorporating more parameters to improve its accuracy, fluency, and versatility across tasks. It is evident in areas requiring context-sensitive information, such as legal document drafting, medical diagnosis support, and scientific research summarization [33]. There are several other LLMs. However, the GPT series of LLMs is the most widely known; hence, we only mention them, and specific LLMs are beyond the scope of this paper.

2.3. Ethical Implications of LLMs

The ethical concerns surrounding LLMs have been extensively discussed in the literature, with bias and privacy as the primary concerns. Bender et al. [34] highlighted the risks of bias and misinformation inherent in LLMs, as those trained on large, diverse datasets from the internet often reflect and amplify societal biases, potentially leading to harmful stereotypes and discrimination. This issue is compounded by the "black box" nature of LLMs, where decision-making processes are

not easily interpretable and opaque, making them difficult to identify and correct biases [35]. Additionally, LLMs' ability to generate realistic text can be misused to spread misinformation and conduct social engineering attacks.

On the other hand, the extensive data requirements for training LLMs raise serious privacy issues. In 2017, Shokri et al. [36] demonstrated that AI models could be vulnerable to membership inference attacks, where an attacker can determine whether a specific individual's data was included in the training set. This has underscored the need for robust privacy-preserving techniques in the development and deployment of LLMs. In 2021, Facebook experienced a significant data breach involving several popular applications with embedded AI components. The incident exposed the personal information of over 500 million users. The database containing user details, such as real names, usernames, gender, location, and phone numbers, was exposed and offered for sale [37]. Since LLMs are trained on datasets and user-generated content from social media platforms, such breaches raise significant concerns about privacy and data security in AI training processes, particularly problematic if the leaked data includes personal identifiers or proprietary details [38].

2.4. Regulatory and Policy Frameworks

The ethical and privacy concerns associated with LLMs have prompted various regulatory responses worldwide. The European Union's proposed Artificial Intelligence Act aims to establish a comprehensive legal framework to ensure the ethical use of AI technologies, focusing on transparency, accountability, and human oversight. This legislation addresses the need for AI systems to disclose AI-generated content and provide explanations for AI-driven decisions, particularly in high-risk applications such as healthcare and law enforcement [39]. In the United States, the Federal Trade Commission (FTC) has issued guidelines emphasising fairness, transparency, and accountability in AI systems [40].

Additionally, the National Institute of Standards and Technology (NIST) is developing a framework for AI risk management to provide organisations with practical tools for managing AI-related risks [41]. China's Interim Measures for the Management of Generative AI Services, introduced in 2024, require clear labelling of AI-generated content and enforce strict data collection and usage guidelines to protect user privacy [42]. These regulatory efforts aim to balance innovation with ethical considerations, ensuring the responsible development and deployment of Generative AI technologies. Furthermore, industry-specific guidelines, such as those from IEEE, provide a framework for ethical AI design, highlighting principles such as transparency, accountability, and data privacy [43].

Tech companies have also established their own ethical guidelines. For instance,

Table 1: Timeline of AI-Ethics Guidelines and Frameworks

Framework / Guideline	**Year**
AI Principles (Google)	2018
OpenAI Charter (OpenAI)	2018
Ethics Guidelines for Trustworthy AI (EU)	2019
Ethically Aligned Design (IEEE)	2019
FTC Guidelines	2021
Artificial Intelligence Act (EU)	2021
AI Risk Management Framework (NIST)	2023
Interim Measures for Generative AI Services (China)	2024

Google's AI principles emphasise fairness, interpretability, privacy, and safety, aiming to prevent misuse of AI technologies [44]. OpenAI has similar guidelines focusing on long-term safety, robustness, and compliance with ethical standards [45].

Despite these regulatory advancements, challenges remain. Overly descriptive regulations might slow down innovation and make compliance burdensome for companies, particularly smaller enterprises that need to use LLMs. Additionally, the global nature of AI development necessitates international cooperation to integrate standards and ensure effective enforcement across jurisdictions. The varying regulatory landscapes across different countries can lead to inconsistencies and gaps in ethical standards and protections [27, 46]. The timeline of the development of regulatory and policy frameworks is shown in Table 1.

2.5. Industry and governmental guidelines selected

As mentioned in the previous section, multiple frameworks and guidelines have been developed to aid in identifying and addressing AI ethics and governance issues. We selected four guidelines/frameworks due to their availability, authority, and significant impact on shaping the ethical use of AI technologies.

The IEEE Ethically Aligned Design Document aims to establish ethical principles for Autonomous and Intelligent Systems (A/IS) that advance human beneficence, i.e., prioritise benefits to humanity and the environment, and mitigate risks associated with these technologies. It emphasises that prioritising human well-being must not conflict with environmental sustainability [47].

National Institute of Standards and Technology (US)'s Artificial Intelligence Risk Management Framework (NIST) outlines characteristics neces-

sary for building trustworthy AI systems, recognising both the transformative potential and unique risks of AI. The document emphasises the importance of balancing various trustworthiness attributes to minimise potential harms [48].

Microsoft Responsible AI Standard aims to operationalise ethical principles across Microsoft's AI development and deployment processes by providing concrete, actionable guidelines [49]. The goals are organised into six key areas, each designed to support responsible and trustworthy AI.

European Union (EU)'s AI Act introduces a comprehensive regulatory framework to manage the development and deployment of AI across the EU, using a risk-based approach to tailor requirements for various AI applications [39].

2.6. Existing Mitigation Strategies for Ethical Concerns

Several mitigation strategies have been proposed and implemented to address the ethical concerns associated with LLMs. One key approach is the development of bias detection and mitigation techniques. Researchers have created tools such as the Large Language Model Bias Index (LLMBI) to quantify and address biases in LLM outputs [50]. These tools use advanced NLP techniques to detect and correct biases related to race, gender, and other sensitive attributes.

Another strategy involves enhancing transparency and interpretability through techniques like SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), which provide insights into how models make decisions and allow users to understand and trust AI outputs [51]. Privacy-preserving techniques such as differential privacy and federated learning are also being used to protect individual data during LLM training. Differential privacy introduces noise to the data to prevent the identification of specific individuals, while federated learning allows models to be trained across multiple decentralised devices without sharing raw data, thus enhancing privacy [52].

Collaboration between AI developers, policymakers, and other stakeholders is crucial in creating ethical guidelines and regulatory frameworks that are both effective and adaptable. Industry-wide initiatives, such as the Partnership on AI, bring together diverse perspectives to address ethical challenges and promote best practices in AI development and deployment [53]. However, these strategies also have limitations. Bias detection tools can only address known biases and may miss subtle or emerging issues, potentially leading to ongoing ethical challenges [50]. Transparency techniques like SHAP and LIME can help interpret decisions but do not eliminate the fundamental complexity of LLMs, and their effectiveness relies on users correctly applying and understanding these methods [51]. Privacy techniques like differential privacy and federated learning require careful implementation to balance privacy

with model performance, and there is limited empirical evaluation of their effectiveness in large-scale deployments [52]. Moreover, the current mitigation strategies often lack comprehensive evaluation, and their long-term impacts on AI ethics and performance remain uncertain [53].

Overall, while these mitigation strategies are essential for addressing ethical concerns, they must be continuously refined and adapted to keep pace with the rapid advancements in AI technology. Collaborative efforts and ongoing research are vital to realising the benefits of Generative AI while minimising potential harms.

2.7. Related Systematic Reviews

The ethical implications of AI systems have garnered increasing attention in recent years, resulting in several systematic reviews and mapping studies. Li et al. (2023) conducted a systematic review focusing on the ethical concerns and related strategies for AI in healthcare, identifying 45 relevant studies. This review highlighted the need for transparency, privacy, and accountability in AI design and implementation, but it was primarily centred on healthcare applications, leaving a gap in understanding how these ethical concerns translate across other domains [15]. Atlam et al. (2024) mapped AI ethics research over the past seven years and identified 127 primary studies. Their findings revealed a concentration of research on fairness, transparency, and accountability while pointing out a lack of empirical evidence supporting AI ethics principles [17]. However, the study was more of a high-level categorisation, lacking in-depth analysis of specific methodologies used to address these ethical concerns. A systematic literature review by the International Journal of Data Science and Analytics (2023) identified 66 papers focusing on developing objective metrics to assess the ethical compliance of AI systems [54]. While this review underscored the need for standardised, quantifiable metrics to evaluate AI ethics, it was limited by its exclusion of frameworks that require human intervention, which might be necessary for a comprehensive understanding of ethical AI [54]. Although these studies have provided valuable insights into the current development of AI ethics, their limitations highlight gaps in the literature. There is a need for a more comprehensive analysis that covers various domains of AI application when the focus is on healthcare, provides high-level categorisations, and excludes human-centred frameworks. This systematic mapping study aims to fill these gaps by providing a broader, more detailed exploration of AI ethics across diverse sectors, identifying strategies for addressing ethical concerns, and laying a foundation for the development of comprehensive, cross-domain ethical guidelines and frameworks.

3. Methodology

We conducted an SMS on the emerging ethical concerns surrounding the use of GenAI, particularly LLMs. We conducted our study on the scientific literature, existing frameworks, and guidelines, e.g., industry guidelines and governmental guidelines for AI ethics. Given the rapidly evolving nature of the field, we recognised that a comprehensive understanding of the ethical concerns surrounding LLMs cannot be derived solely from scientific literature. Therefore, we expanded our study to include industry frameworks, governmental guidelines, and other authoritative sources, ensuring a more holistic and up-to-date perspective on the ethical challenges in this dynamic area. This SMS has been carried out in accordance with the guidelines for systematic mapping studies as outlined by Peterson et al. [14].

Our mapping study was conducted in three stages: planning, conducting, and reporting the review. The planning phase involved writing a protocol, formulating research questions (RQs), and reviewing it for alignment with the main aim of our study. The protocol included a search strategy. During the conducting stage, the first author developed search strings that the other authors reviewed. A librarian from our university was also consulted at this stage to ensure alignment with the best practices for conducting systematic reviews. The search strings went through an iterative process, and some that were not specific or only marginally relevant were removed. The first author then led our search process (illustrated in Figure 1) with the search strings in various databases and selected ACM Digital Library, IEEE Xplore, Proquest, Wiley, Web of Science, and Science Direct. The first author searched all selected databases and removed duplicates using EndNote in the initial paper screening. We identified relevant keywords and search strings to exhaustively explore existing empirical evidence. We note that, in accordance with the guidelines for this stage, we used the search process to identify relevant scientific literature. We identified industry and governmental guidelines using references in the scientific literature and snowball sampling. Following Wohlin's guidelines [55], we applied forward snowballing to the 37 studies remaining. Two researchers independently screened the citing articles against our predefined inclusion and exclusion criteria, resulting in 2 additional eligible studies. In total, our search process led to 39 scientific studies and six industry and governmental guidelines.

3.1. Research Questions

We formulated three high-level research questions (RQs):

RQ1 - What are the ethical dimensions defined in the use of generative AI across various fields?

RQ2 - What strategies are used or proposed to address the ethical concerns of using

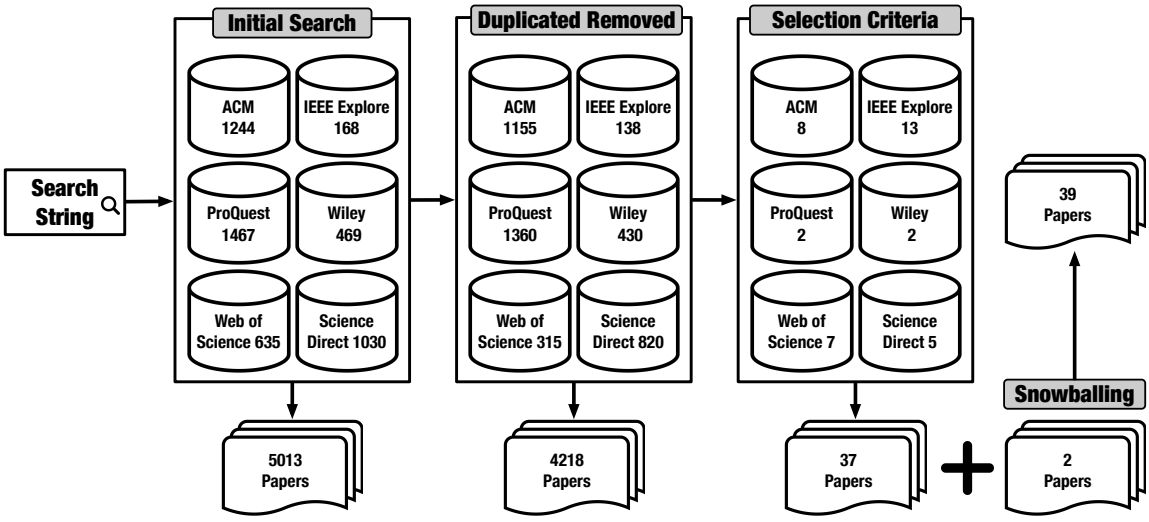

Figure 1: Study Search Process

generative AI across various fields?
RQ3 - What are the challenges when implementing the strategies?
We use these RQs to identify studies on the ethical dimensions of Generative AI use and the strategies to address them. In our study, an ethical dimension refers to specific ethical considerations arising from the use of GenAI or LLMs [19, 20, 22]. These dimensions encompass a range of ethical concerns, principles, and guidelines that are crucial to the responsible development and use of LLMs.

3.2. Search Strategy

3.2.1. Search String Formulation

We formulated a search string to query online databases using key and alternative terms across four main areas: large language models, guidelines, development, and ethics. However, we acknowledge that these areas may sometimes overlap with related concepts. Thus, we included a comprehensive set of terms in the search string to capture all relevant literature, shown in 9. The search string was tested across six online databases: IEEE Xplore, ACM Digital Library, ProQuest, Web of Science, Wiley Online Library, and ScienceDirect. These databases were chosen because they are comprehensive and widely recognised sources of high-quality research in computer science, engineering, and technology. In addition to these databases, we conducted a parallel search on Arxiv, finding three preprint studies that met our inclusion criteria. We included these papers to capture the most recent and relevant studies. We included the Arxiv papers as the topic of Ethical AI is relatively new,

and Arxiv provides us with studies and research that may not yet have gone through peer review, which allows us to explore emerging ideas, diverse perspectives, and early-stage research that might not yet be available in journals and conferences.

The ACM Digital Library and IEEE Xplore are particularly relevant as they contain a vast collection of research papers and conference proceedings focused on AI and machine learning. ProQuest, Wiley, Web of Science, and ScienceDirect were included for their extensive coverage of multidisciplinary studies, including regulatory guidelines and ethical considerations in technology. Additionally, we included some of these databases to capture grey literature, e.g., the industry and governmental guidelines. Given that AI ethics is a relatively new field, grey literature (not part of the main 39 papers) provides valuable insights and complements our understanding.

The search string was refined over several iterations to maximise the relevance of the results. For instance, the first author would randomly select a sample of 8-10 papers, review them for relevance, and further refine the search string. In constructing search strategies for various databases, we used a range of logical connectors to ensure comprehensive and relevant retrieval of literature on large language models and related topics. Each database required specific adjustments to optimise the search strategy, using various connectors and operators tailored to the database.

In IEEE, Web of Science, Wiley Online Library, and ScienceDirect, the search strategies used the "AND" connector to combine key concepts, ensuring that search results included multiple relevant topics, such as governance, ethics, transparency, and accountability, in relation to large language models (LLMs). The "OR" operator was used to include alternate expressions of similar concepts (e.g.,"ethic" OR "moral", "development" OR "design"). Additionally, symbols like the asterisk (*) were used to capture word variations. For instance, "Large Language Model*" ensures that both "Large Language Model" and "Large Language Models" are retrieved. This strategy helps retrieve documents that mention different term forms, maximising the search's comprehensiveness. In ACM Digital Library, the search strategy used the "AND" operator and "OR" operator to combine multiple concepts and capture alternate terms. "AND" was used to ensure that search results contained all the specified concepts, such as "large language model*" AND "ethics" AND "governance." This ensures that the retrieved documents address all selected topics. "OR" was used to include variations or synonyms of a concept, such as "ethics" OR "moral", broadening the search to include documents that may use different terms for the same idea.

In ProQuest, we used a more refined approach by employing the "noft" operator, which stands for "not in full text". This operator excludes terms that appear only in the body of the document and not in more critical fields such as the title, abstract, or subject headings. The use of "noft" is beneficial in ProQuest because this database

often contains a large volume of full-text content, including dissertations, reports, and articles that may mention key terms incidentally without focusing on them in depth. By using "noft", we can filter out documents in which a term like "governance" is only briefly mentioned and not central to the research focus. For example, using "noft(governance)" ensures that we retrieve documents where "governance" is emphasised in key fields, like the title or abstract, indicating that the topic is a primary focus of the work rather than a tangential mention. This increases the relevance and specificity of the search results, ensuring that only documents that deal substantially with "governance" are included. Additionally, the "AND" operator and "OR" operator were also used in ProQuest to combine different thematic elements and provide alternate terms for key concepts, ensuring comprehensive coverage of topics while maintaining the precision of the search.

To include grey literature, we conducted targeted searches of the websites of major AI ethics bodies (IEEE, NIST, Microsoft, and the European Commission). Because these documents are not indexed in academic databases, we did not apply a Boolean search string. Instead, we navigated each site and used its built-in search tools with keyword filtering (e.g., "AI ethics" "responsible AI"). We identified four guidelines/frameworks for the paper, which are presented in Section 2.5.

Table 2: Inclusion and Exclusion Criteria

Criteria ID	**Criterion**
Inclusion Criteria	
I01	Papers discussing ethical concerns in the use of Generative AI
I02	Full text of the article is available.
I03	Peer-reviewed studies, sector-specific studies and grey literature (Arxiv, Guidelines and Frameworks).
I04	Papers written in English language.
Exclusion Criteria	
E01	Papers about GenAI that do not discuss ethical concerns
E02	Papers about ethical concerns that are not in the field of GenAI
E03	Papers that are less than four pages in length.
E04	Conference or workshop papers if an extended journal version of the same paper exists.
E05	Papers with inadequate information to extract relevant data.
E06	Vision papers, books (chapters), posters, discussions, opinions, keynotes, magazine articles, experience, and comparison papers.

3.2.2. Automated Search and Filtering

Our search was conducted on databases between April and May 2024. The final search string was applied to the selected database online search engines, and an initial pool of 5013 papers was extracted. The list of papers was downloaded and exported to EndNote. Despite the large number of database results obtained by our search string (5013 papers), many papers were irrelevant or duplicates.

We first removed all duplicates (remaining papers = 4218) and applied a set of inclusion and exclusion criteria to filter out pertinent studies as part of our SMS protocol, shown in Table 2). For this mapping study, we considered only English-language studies that directly discuss the ethical dimensions or concerns. Given that the topic is relatively new, we included short papers of more than four pages to capture emerging ideas. However, we excluded posters, keynotes, opinion papers, and magazine articles. We required full text availability (I02) to enable thorough analysis of ethical dimensions, mitigation strategies, and implementation challenges, which are often absent from abstracts. We excluded book chapters (E06) to maintain consistency in peer-review standards, focusing on recent conference and journal papers that undergo more comparable review processes. The selection criteria were applied to all studies to identify the most relevant ones, with discussions among all authors during the study filtering process. This left us with 39 papers remaining.

3.2.3. Data Analysis Process

After completing our automated search and filtering, we conducted a structured analysis of the 39 selected primary studies. Our data analysis process is shown in Figure 2. First, the first author created an article matrix to capture the studies' aims, methodologies, and key results; this matrix was then shared among all authors to establish an overview of the studies. The first and third authors then systematically identified and extracted data from each study that were relevant to our three RQs. The other authors independently reviewed these data extractions to verify their accuracy and comprehensiveness. The first and third authors then organised the extracted data into preliminary categories aligned with our RQs. These categories were refined through an iterative process of condensing and open coding. During extraction, we labelled each proposed approach as either a *mitigation strategy* (actionable and potentially evaluable) or a *recommendation* (broader guidance, not necessarily designed for short-term empirical testing), following our definitions in Section2.1. For example, "apply differential privacy during training to reduce membership inference risk" is treated as a mitigation strategy, whereas "organisations should adopt privacy-by-design policies" is treated as a recommendation. We

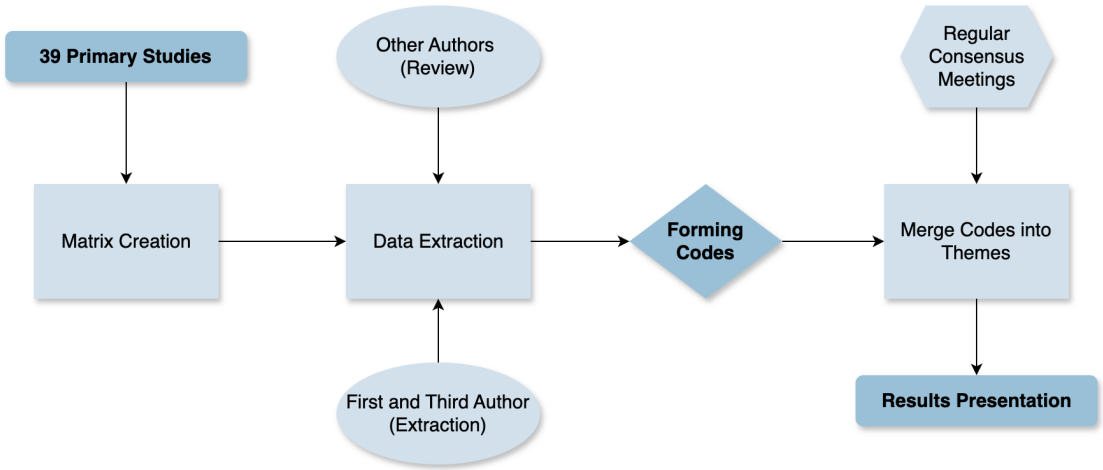

Figure 2: Data Analysis Process

highlighted the relevant information for each RQ from the extracted data. We coded them according to the key information, and similar codes were then combined into themes. Throughout this phase, all authors met regularly to discuss coding decisions, reconcile discrepancies and reach consensus.

4. Results

4.1. Selected Studies

After filtering, we selected 39 primary studies on ethical concerns in LLMs, published from 2020 to 2024 (Table 3A). Publications were limited from 2020 to 2022, with only five studies total. However, in 2023, the number increased sharply to 27 studies, reflecting significant growth in research interest. The lower number in 2024 (7 studies) likely results from our June 2024 search cutoff. Table 3B shows venue distribution: 61.5% in journals, 30.8% in conferences, and 7.7% on Arxiv. We included Arxiv papers to capture recent research in this fast-changing field, allowing us to examine early-stage ideas together with established peer-reviewed publications.

Our selected primary studies span multiple application domains. Each domain brings its own set of challenges and ethical considerations, reflecting the diverse applications and potential impacts of LLMs. This section categorises the studies by the domains mentioned in the primary studies, offering a detailed exploration of how ethical concerns manifest across different contexts.

Table 3: Distribution of Publications by Year and Venue

(A) Number of Publications Per Year	
Year	**Number of Publications**
2020	1
2021	1
2022	3
2023	27
2024	7
Total	**39**

(B) Publication Venue Distribution	
Venue Type	**Number (Percentage)**
Journal	24 (61.5%)
Conference	12 (30.8%)
Arxiv	3 (7.7%)
Total	**39 (100%)**

- *General AI*: Papers categorised under "General AI" address AI's broad concepts and applications without focusing on a specific domain. They discuss relevant ethical challenges across multiple sectors where AI is deployed.

- *Healthcare*: Healthcare is a domain that frequently raises ethical concerns, such as privacy, safety, and bias, especially as AI and LLMs are increasingly used for medical applications. The papers in this domain discuss how AI-driven systems can both enhance and complicate healthcare practices, with patient confidentiality, data security, and algorithmic fairness being recurring themes. This domain is highly sensitive due to the direct impact on human well-being and medical decision-making.

- *Legal*: Papers in this domain highlight the intersection of AI and law, exploring issues such as accountability, transparency, and bias in legal AI systems. The

integration of AI in legal contexts raises concerns about fairness, potential racial biases in predictive policing tools, and the transparency of AI-driven legal decisions.

- *Education*: The educational domain explores the ethical use of AI in learning environments, with concerns including privacy, fairness, and the transparency of AI systems in assessing student performance and providing educational content.

- *Public safety*: This domain involves the use of AI in public safety systems, where the ethical dimensions centre on accountability, bias, and safety. AI systems used in policing, emergency response, and public surveillance must be scrutinised for fairness and transparency to avoid unintended harm to communities, especially marginalised groups.

- *Societal Impact*: Papers in this domain examine how LLMs interact with broader societal issues, such as their role in shaping public discourse, policy, and societal norms. Ethical concerns focus on the responsibility of those deploying LLMs to consider their social impact, including how they influence public opinion, access to information, and equality.

- *Cybersecurity*: This domain primarily focuses on privacy concerns related to systems that use AI components or LLMs. Papers here focus on issues such as the protection of personal data, the risks of data leakage, and how AI systems can be designed to respect user privacy, especially focusing on data-hungry models like LLMs.

- *Economics*: Papers in this domain focus on the economic impact and ethical concerns surrounding AI and LLMs, particularly in relation to automation, job displacement, and the ethical use of AI in economic decision-making. The economic dimension of LLMs also raises issues of accessibility and fairness in how AI systems are deployed and who benefits from their use.

The specific papers associated with each domain and ethical dimensions can be seen in Table 4, organised into three columns. The first column "Domain" groups the selected studies by domains. The second column "Papers" lists each primary study in that domain. The third column, "Ethical Dimensions (n)," shows, for each domain, how many of its studies address each ethical dimension, along with the count. From the studies, it is worth noting that accountability is rarely addressed in the education and public safety domains, while both transparency and accountability

Table 4: Domains mentioned in papers and their ethical-dimension counts

Domain	Papers	Ethical Dimensions (n)
Cybersecurity	[P1, P23, P27, P29, P31, P33]	Safety (2), Privacy (6), Bias (1), Transparency (0), Accountability (0)
Education	[P14, P16, P26]	Safety (2), Privacy (3), Bias (2), Transparency (2), Accountability (1)
General AI	[P2, P4, P5,P15, P17, P20, P28, P30, P35, P37, P39]	Safety (4), Privacy (7), Bias (8), Transparency (5), Accountability (4)
Healthcare	[P1, P9, P10, P11, P13, P21, P22, P24, P32, P34]	Safety (7), Privacy (7), Bias (7), Transparency (7), Accountability (5)
Societal Impact	[P2, P25, P38]	Safety (2), Privacy (3), Bias (2), Transparency (2), Accountability (1)
Legal	[P3, P6, P8]	Safety (2), Privacy (1), Bias (3), Transparency (2), Accountability (1)
Public Safety	[P7, P18, P19, P27]	Safety (4), Privacy (1), Bias (2), Transparency (1), Accountability (0)

are completely absent from the cybersecurity literature. Transparency also receives only minimal attention in the education, societal impact, legal, and public safety papers. It is critical that future research focuses on these underrepresented ethical dimensions within their respective sectors.

We identified a diverse range of LLM models being used in the selected primary studies, shown in Table 5. A significant number of papers used versions of ChatGPT (P1, P7, P8, P11, P12, P14, P16, P17, P18, P19, P24, P25, P27, P28, P34, P35, P37, and P39), highlighting its versatile use across various domains such as healthcare, education, and general conversational interfaces. Conversational Agents (CA), another key area of focus, were discussed in multiple papers (P2, P5, P15, P29, P31, P36. These agents are known for their roles in interactive and supportive capacities, including customer service and educational tools.

General LLMs were featured in several papers (P4, P6, P20, P21, P30, P32, P33, P38). These papers emphasised the broad capabilities of LLMs in understanding and generating human-like text, highlighting their potential across various sectors. Additionally, LLM-based chatbots and virtual assistants were discussed in papers

Table 5: Generative AI Technologies Used in Primary Studies

LLM Models	Number Mentions
ChatGPT	18
Conversational Agents	6
DALL-E	1
EduLLMs	1
Gemini (Bard)	2
GPT-2	1
General LLMs	8
LLM-based Chatbots	1
LLM Virtual Assistants	1
RoBERTa	1
Transformers	1

such as P9, P10, and P13, indicating their growing relevance in enhancing user interaction and automating responses.

Models such as GPT-2 (P3) and transformers (P22) were also explored, with RoBERTa being mentioned alongside ChatGPT in P23 for the use of fine-tuning LLMs. Educational Large Language Models (EduLLMs) were specifically addressed in P26, showcasing their application in creating and facilitating educational content. Moreover, newer AI models such as Gemini (Bard) and DALL-E were discussed in the context of their capabilities and potential applications in papers P37 and P39.

4.2. RQ1 Results

We wanted to identify the various ethical issues related to AI that have been discussed across the selected papers. The purpose was to explore both the breadth and depth of these issues and how they align with key ethical dimensions. We operationalised five core ethical dimensions (Safety, Privacy, Bias, Transparency, and Accountability) because they are consistently emphasised across the four major frameworks reviewed in Section 2.5 and were also the most recurrent categories in our coding of the 39 primary studies. Related values (e.g., fairness, autonomy, consent) were captured within these dimensions (fairness under Bias; autonomy under Accountability; consent under Privacy).

Safety: reliability, robustness, and harm prevention in high-risk settings.
Accountability: responsibility assignment, governance, auditability, and liability.
Bias: fairness and non-discrimination in model behaviour and outcomes.
Transparency: explainability, disclosure, traceability, and audit support.
Privacy: data protection, confidentiality, minimisation, and user control.

We use these dimensions as top-level categories for coding ethical issues and grouping them into themes. We also considered the authors' perspectives in each paper, incorporating their opinions on which ethical dimension specific issues should

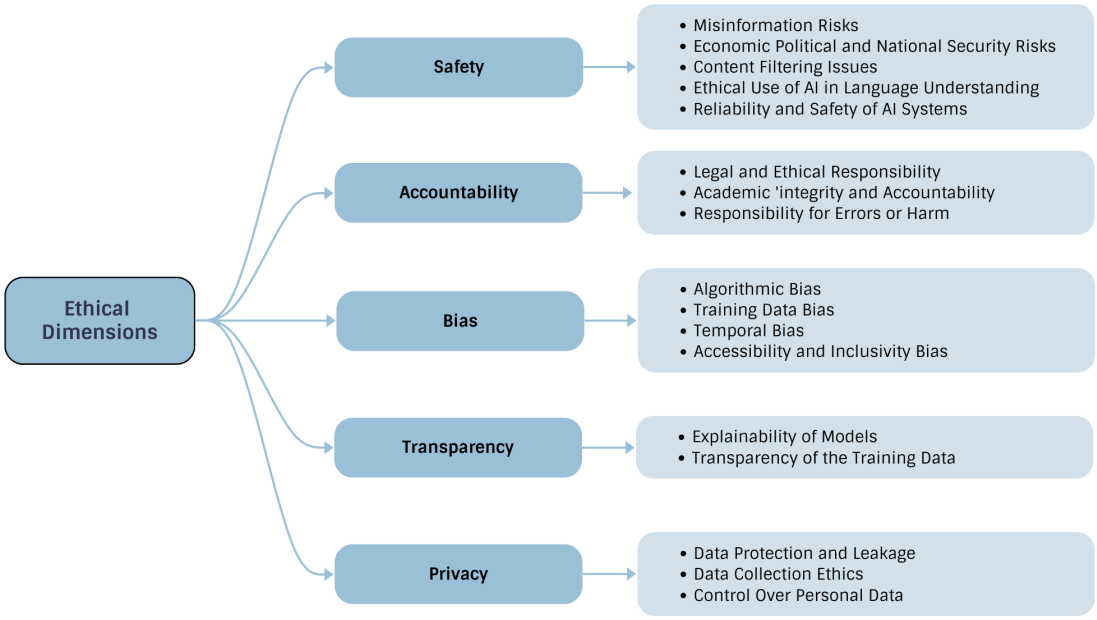

Figure 3: Ethical dimensions mapped with RQ1 themes

fall under. This ensured that our categorisation was informed not only by the frameworks and guidelines but also by the authors' interpretations. The analysis involved carefully reviewing each paper and coding the ethical issues mentioned, which were then grouped into themes within these categories. In total, we identified 130 different ethical issues across the studies, which we categorised into 17 distinct themes. Figure 3 shows the key ethical dimensions to theme mappings. Table **??** shows the primary studies addressing each dimension.

4.2.1. Safety

The theme *"Ethical Use of AI in Language Understanding"* highlights the ethical challenges related to how AI systems, particularly virtual assistants (VAs), process and respond to language, including inappropriate or harmful speech. In P10, a key concern is raised about the need to train VAs with inappropriate language for them to recognise it. This raises ethical questions about whether the VA should be able to recognise insults or offensive language and, if so, how it should respond.

The theme *"Misinformation Risk Issues"* identifies the ethical concerns related to AI systems generating and disseminating false or misleading information, which can have serious consequences. This theme was covered in P2, P6, P8, P10, P12,

P13, P14, P28, P30, P32, P38. In P12, a significant risk is highlighted: ChatGPT may produce sophisticated but hallucinatory fake information, especially in financial outputs, which could be hard to detect and lead to substantial financial losses. In P30, another concern is raised about LLMs generating misinformation, resulting in less informed users and eroding public trust in shared information.

The theme *"Economic, Political, and National Security Risks"* focuses on the dangers posed by AI systems in sensitive areas, particularly regarding the misuse of generative AI like ChatGPT. This theme is covered in P18 and includes political security risks. AI could inadvertently access and assimilate sensitive information, such as national secrets, trade secrets, or personal data, through user inputs, potentially leading to information leakage. Furthermore, military security could be threatened by the potential misuse of AI to generate attacks targeting critical infrastructure. Economic security might also be at risk, as tampering with training data could create false financial data and market predictions, undermining economic integrity.

The theme *"Content Filtering Issues"* addresses the ethical concern of LLMs, processing both accurate and harmful content due to inadequate filtering mechanisms. This theme covers P13. In P13, it is noted that LLMs can ingest and incorporate harmful content alongside accurate information, leading the AI to produce outputs that may be inappropriate or damaging.

The theme *"Reliability and Safety of AI Systems"* focuses on the ethical concerns surrounding the accuracy and dependability of AI, particularly in critical situations; this theme is covered in P7, P9, P13, P14, P19, P21, P25, P27, P28, P30, P34. In P19, a significant issue is the provision of oversimplified and erroneous safety advice, which often lacks traceability due to missing cited sources, making fact-checking difficult. This raises concerns about AI's reliability in delivering accurate and trustworthy information in safety-critical contexts. In P21, the issue of trust is discussed, especially regarding the reliability of LLMs in situations outside their training conditions. This phenomenon, known as an "out-of-distribution shift," can lead to performance failures.

4.2.2. Privacy

The theme *"Control over Personal Data"* emphasises the ethical importance of protecting individuals' privacy and ensuring they have control over how their personal information is used, particularly in AI-driven systems. As AI increasingly relies on sensitive data, especially in fields like healthcare, safeguarding this data becomes crucial to prevent misuse and ensure privacy compliance, this theme is covered in P1, P14 and P31. In P1, one key concern is the need for robust data protection and transparency in AI's decision-making, especially in healthcare, to avoid bias

and ensure fairness. In P14, an example highlights the potential for adding privacy settings to give users control over how their data is collected and shared, reinforcing the need for individual autonomy over personal information.

The theme *"Data Protection and Leakage"* addresses the ethical concerns related to the privacy and security of data in AI systems, particularly regarding the risk of unauthorized access or data breaches, this theme is covered in P1, P2, P5, P7, P8, P9, P10, P11, P12, P14, P16, P17, P22, P23, P24, P25, P26, P27, P28, P29, P30, P33, P34, P37, P38, P39. In P23, the issue is highlighted in the context of AI researchers fine-tuning pre-trained models with private data for various tasks. These models are vulnerable to privacy attacks due to their tendency to memorize training data, known as the "memorization issue," posing risks to sensitive information. In P26, concerns arise about smart education systems that collect and analyse vast amounts of student data to provide personalised learning.

The theme *"Data Collection Ethics"* focuses on the ethical issues surrounding how data is gathered for training LLMs. This theme is covered in P14 and P27. In P27, concerns are raised about scraping information from internet forums and other sources without proper consent or oversight, which raises questions about the ethicality of the data collection process. In P14, another issue is the lack of disclosing the underlying sources used in LLM training, as they are not shared with the public.

4.2.3. Bias

The theme *"Algorithmic Bias"* highlights the ethical concerns regarding how AI models can perpetuate social stereotypes and discrimination, particularly when trained on biased data. This theme is covered in P4, P5, P6, P12, P22, P24, P25, P28, P30, P34, P35, P38. In P25, an example is provided where Amazon used AI in human resources, and it was found that women were consistently rated lower than men. This bias arose because the AI analyzed resumes from the past decade, a period where most resumes came from men, indicating that the data was neither large nor diverse enough to mitigate bias. In P30, a similar concern is raised about how LLMs can perpetuate social stereotypes and introduce representational and allocational harms, leading to discrimination.

The theme *"Training Data Bias"* addresses the ethical concerns that arise when AI models are trained on datasets that either overrepresent or underrepresent certain demographic groups. This theme is covered in P1, P3, P4, P8, P10, P11, P16, P17, P28, P35. In P35, demographic bias is highlighted as a major issue, as data imbalances can cause models to exhibit biased behaviour toward specific genders, races, ethnicities, or other social groups. In P28, the concern is further emphasised, showing how biased data can perpetuate social stereotypes and lead to unfair dis-

crimination. When models are trained on skewed representations of certain groups, this can result in unjust outcomes that reinforce existing inequalities.

The theme *"Temporal Bias"* focuses on the ethical concerns arising from the time-based limitations of training data used in AI models. This theme is associated with P35. In P35, it is noted that these models are often trained on data from specific time periods or include temporal cutoffs, which can introduce bias when reporting on current events, trends, or opinions. This limitation may result in the model providing outdated or inaccurate information when addressing contemporary topics and an incomplete understanding of historical contexts due to the lack of temporally diverse data. This can affect the relevance and accuracy of AI-generated content in time-sensitive situations.

The theme *"Accessibility and Inclusivity Bias"* highlights the ethical concerns related to how AI systems may overlook the needs of individuals with disabilities or those from diverse linguistic backgrounds, this theme is covered in P14, P15, P18, P19, P20, P35. In P14, it is highlighted that accessibility features, such as screen readers, alternative text for images, or video captions, are often inadequate or absent in AI systems. Additionally, limitations in language translation capabilities may exclude non-native speakers or those with diverse linguistic needs.

4.2.4. Accountability

The theme *"Legal and Ethical Responsibility"* focuses on the challenges surrounding accountability and moral responsibility when using AI in decision-making processes, especially in critical contexts, this theme is covered in P5, P21, P34, P37, P38. In P5, it is noted that accountability is a pivotal element of AI governance, particularly when delegating tasks such as prediction or decision-making to AI systems. However, the definition of accountability in AI remains ambiguous and should be clarified based on the subject, scope, and context of its application. In P21, the issue of moral responsibility arises in high-stakes decisions, such as medical care, where traditionally, a human is held accountable. When an AI algorithm is involved, it becomes less clear who is responsible for the outcome, especially in adverse decisions. Establishing clear guidelines for when a human professional may "overrule" the AI is essential to maintaining accountability.

The theme *"Academic Integrity and Accountability"* highlights the ethical concerns surrounding using AI, particularly ChatGPT, in academic settings, this theme is covered in P11, P16, P17. In P16, the broader academic community has raised concerns about students using ChatGPT to plagiarise assignments, research papers, and other academic work, which undermines academic integrity. In P17, questions arise regarding the proper use of ChatGPT's responses for academic purposes, such as

whether AI-generated content should be credited and how accountability is managed if it is misused. The uncertainty extends to legal implications, mainly if AI-generated material is used maliciously or in ways that breach academic or ethical standards.

The theme *"Responsibility for Errors or Harm"* involves the ethical concerns related to AI-driven decisions and the risks associated with responsibility. This theme is covered in P1 and P24. In P1, it is highlighted that the availability of AI and automated decision aids can lead to a human tendency to rely too heavily on these technologies, often minimising cognitive effort. In healthcare, this over-reliance on AI-driven decisions by clinicians can lead to misleading conclusions, potentially endangering patient safety and well-being.

4.2.5. Transparency

The theme of *"Explainability of Models"* addresses the transparency challenges posed by the complexity of LLMs. In P38, the "black box" nature of LLMs is emphasised, highlighting the difficulty in understanding their decision-making processes due to the vast number of parameters they contain—often in the millions or billions. This theme is covered in P2, P5, P8, P10, P11, P12, P13, P15, P16, P21, P22, P24, P34, P38, and P39. This complexity makes it nearly impossible to explain how these models generate their outputs fully. In P39, the mystery surrounding the emergent capabilities of LLMs is further discussed as researchers remain uncertain about what drives these behaviors.

The theme *"Transparency of the Training Data"* focuses on the ethical issues arising from the lack of clarity regarding the data used to train AI models. This theme is covered in P6 and P19. In P6, it is noted that the training data is often treated as a trade secret, and while efforts are underway to reverse-engineer which data was used, companies neither confirm nor deny the accuracy of these guesses. This opacity raises concerns about the integrity and representativeness of the training data. In P19, the lack of transparency becomes particularly problematic when ChatGPT provides answers or recommendations, leaving users without a clear understanding of the sources behind its advice. Unlike human experts who can cite specific references, ChatGPT operates on a probabilistic model, making it difficult to trace or explain the origins of its responses.

RQ1 Key Takeaways

Ethical concerns centre on privacy and bias, while accountability remains critically underexplored or entirely absent in the education and public safety domains. Safety issues (misinformation, reliability) and transparency challenges (model explainability, opaque training data) persist across high-stakes applications. This distribution suggests current frameworks prioritise technical fixes over governance mechanisms, leaving the question of "who is responsible when AI fails" systematically unaddressed.

4.3. RQ2 Results

Table 6: Mapping of RQ2 Themes to Ethical Dimensions

Theme	**Accountability**	**Transparency**	**Bias**	**Privacy**	**Safety**
Ethical Frameworks & Interdisciplinary Collaboration	✓	✓			
Accountability & Continuous Oversight in AI systems	✓		✓		
Bias Mitigation & Fairness in AI systems			✓		
User Empowerment & Transparency in AI Interactions		✓		✓	
Enhancing Trust & Interpretability in AI systems	✓				✓

To address RQ2, "What strategies are used to address the ethical dimensions?", we conducted a thorough review of the studies. We extracted the mitigation strategies and recommendations presented in each paper. A mitigation strategy refers to a specific approach or action proposed or implemented to directly address or reduce a particular risk, issue, or challenge. In contrast, a recommendation is a suggestion or guidance proposed by the authors for future actions, often highlighting potential solutions or best practices that do not require immediate implementation. These strategies were categorized accordingly, and we further explored their evaluation status. To assess evidence strength, we classified each mitigation strategy as *fully evaluated* (quantitative/qualitative assessment with clear outcome measures, e.g., user study, trial, benchmark) or *not evaluated*. Studies reporting only limited empirical probing without clear outcome measures were conservatively coded as *not evaluated*. Two authors independently applied this rubric. Only 5/39 studies (12.8%) fully evaluated mitigation strategies. After identifying the evaluation status, we conducted a thematic analysis and categorised various mitigation strategies or recommendations, reflecting the proposed or practical solutions presented. Table 6 maps RQ2 themes to ethical dimensions.

- Ethical Frameworks and Interdisciplinary Collaboration: Establishing ethical guidelines for LLMs requires input from diverse disciplines to address the complex ethical challenges they present, this theme is covered in P1, P2, P4, P6, P16, P18, P24, P25, P30, P35, P37, P38. In P1, the recommendation emphasizes the need to establish ethical guidelines and regulatory frameworks for AI in healthcare, advocating for interdisciplinary collaboration to ensure that ethical standards are comprehensive and adaptable to advancements in technology. This collaboration is essential to create frameworks that are both effective and applicable across various domains. In P2, a Machine Ethics approach is proposed, suggesting that ethical standards and reasoning should be directly embedded within AI systems. This approach would enable AI to make ethically informed decisions autonomously, integrating ethical principles into the core of AI functionality. In P6, there is a focus on the legal dimensions of AI, suggesting that the training phase of AI may be covered under fair use; however, clearer guidelines are needed to inform system users. These examples illustrate the necessity of establishing ethical frameworks that are continually refined through interdisciplinary cooperation, ensuring that LLMs operate ethically and align with societal expectations.

- *Bias Mitigation and Fairness in AI systems*: Addressing biases and promoting fairness in LLMs requires a multifaceted approach, with strategies implemented at various stages of the AI development process, this theme is covered in P1, P2, P3, P4, P6, P8, P16, P19, P25, P28, P35. In P3, a mitigation strategy is employed using a fill-in-the-blank method in a GPT-2 model to ensure that context is carefully considered during predictions, focusing on binary racial decisions between “White" and “Black". This method involves examining racial bias by masking racial references within the context, prompting the model to assign probabilities and observe potential biases. In P4, a combination of techniques is used to handle biases systematically: pre-processing methods transform input data to reduce biases before training, while in-processing techniques modify learning algorithms to eliminate discrimination during training. Additionally, post-processing methods are applied to adjust the model's output after training, particularly when retraining is not an option, treating the model as a black box. These strategies showcase a comprehensive effort to mitigate biases at different levels of the AI development pipeline, aiming to create fairer and more equitable AI systems.

- *Enhancing Trust and Interpretability in AI Systems*: Building trust in LLMs relies heavily on making AI systems transparent and understandable to users,

especially in critical fields like healthcare, this theme includes P1, P7, P9, P15, P20, P24, P30. In P1, a recommendation is made to focus on improving the interpretability of AI algorithms so that healthcare professionals can clearly understand and trust the decisions generated by these systems. This involves making the decision-making processes of AI transparent, allowing professionals to verify and rely on AI outputs confidently. In P7, a mitigation strategy involves using leading questions to guide ChatGPT when it fails to generate valid responses, enabling users to refine the system's outputs until they are accurate iteratively. This interaction fosters trust by giving users greater control over the AI's response quality. In P9, the HCMPI method is recommended to reduce data dimensions, focusing on extracting only relevant K-dimensional information for healthcare chatbot systems. This reduction makes the AI's reasoning clearer and more interpretable, allowing users to follow the underlying logic without getting overwhelmed by excessive data. These strategies aim to enhance AI systems by making them more interactive and interpretable.

- *User Empowerment and Transparency in AI Interactions*: Empowering users and ensuring transparency in AI systems are critical for fostering ethical interactions and trust, this theme is covered in P1, P2 P2, P5, P8, P10, P11, P12, P13, P16, P17, P24, P28, P29, P30, P31, P33. In P2, a recommendation is made to emphasize critical reflection throughout conversational AI's design and development phases. This includes giving users more control over their interactions with AI agents and being transparent about the AI's non-human nature and limitations. Such transparency allows users to understand the AI's capabilities and limitations, enabling more informed decision-making. In P8, a mitigation strategy involves techniques like logit output verification and proactive detection of hallucinations. These strategies are paired with participatory design, where users actively shape AI systems, ensuring that the AI's responses remain accurate and meaningful. In P30, further mitigation strategies include functionality audits to assess whether LLM applications meet their intended goals and impact audits to evaluate how AI affects users, specific groups, and the broader environment. These strategies prioritize user involvement and clarity, fostering a transparent and user-centered AI ecosystem.

- *Accountability and Continuous Oversight in AI Systems*: Ensuring that AI systems are accountable and continuously monitored is essential to maintain ethical standards and protect users, this theme is covered in P5, P7, P9, P10, P11, P12, P13, P20, P24, P25, P30, P31, P32, P35. In P24, a recommendation is made for AI tools like ChatGPT to be evaluated by regulators, specifically

in healthcare settings, to ensure safety, efficacy, and reliable performance. This highlights the importance of oversight from regulatory bodies to ensure that AI applications adhere to established standards. In P9, a mitigation strategy involves the Healthcare Chatbot-based Zero Knowledge Proof (HCZKP) method, which enables the use of data without making it visible. This approach reduces the need for extensive data collection, safeguarding privacy, and ensuring ethical data handling. Additionally, the strategy recommends adopting data minimization principles by collecting only necessary essential data and decentralizing patient data during feedback training. These strategies emphasize the need for ongoing oversight and accountability in how AI systems manage and utilize data, particularly in sensitive environments.

Although many recommendations remain conceptual, several mitigation strategies have been empirically evaluated, with mixed outcomes. In NLP, pre-processing, in-processing, and post-processing techniques (P4) have reduced measurable bias but often fail to generalize across different domains, limiting the impact of achieving fairness in practice. In healthcare, the HCMPI method and HCZKP protocol (P9) have successfully reduced the sensitive patient data exposure without affecting the performance of the chatbot, these results are still confined to controlled evaluations in the experiment, rather than long term practice. In education, the responsible chatbot framework supported self-regulation and cognitive engagement, showing feasibility in a small scale experiment; however, robust empirical trials are needed to achieve more promising results (P14). The results have shown that while mitigation strategies show potential, more empirical evidence are needed to establish their reliability and effectiveness in practical scenarios.

RQ2 key Takeaways

Only 5 of 39 studies (12.8%) empirically evaluated their proposed mitigation strategies, revealing a critical gap between conceptual proposals and validated solutions. Bias mitigation techniques (pre-processing, in-processing, post-processing) show technical promise but limited cross-domain generalization. Privacy safeguards like HCZKP protocol succeeded in controlled healthcare experiments yet lack long-term real-world validation. Transparency and accountability strategies remain predominantly theoretical, with few tested implementations. The heavy reliance on untested conceptual frameworks means most proposed solutions lack evidence of practical effectiveness, particularly for governance and oversight challenges.

4.4. RQ3 Results

Table 7: Mapping of RQ3 Themes to Ethical Dimensions

Theme	**Accountability**	**Transparency**	**Bias**	**Privacy**	**Safety**
Data-Related Challenges				✓	✓
Technical Challenges			✓	✓	✓
Legality Challenges			✓		✓
Ethical Standard Challenges	✓				✓
Individual Use Challenges	✓		✓		
Ethical Dilemma Challenges		✓	✓		

While numerous mitigation strategies have been proposed to address the ethical concerns surrounding the use of LLMs, many studies indicate that significant challenges persist even after these strategies are implemented. Authors in several papers have explicitly acknowledged that the mitigation efforts, while promising, often fall short due to various technical, legal, and governance barriers. These challenges highlight the complexity of achieving effective ethical governance in LLM applications. To better understand these issues, we have categorised the identified challenges into key themes, each reflecting the unique difficulties encountered during the practical implementation of these mitigation strategies (also summarised in Table 8 for readability).

Figure 4 summarises how the five ethical dimensions (RQ1) relate to mitigation strategies (RQ2) and the main challenge categories reported when implementing these strategies (RQ3).

- *Technical Challenges*: Technical challenges in implementing mitigation strategies for LLMs often revolve around the complexity of integrating advanced techniques within diverse and sensitive contexts, this theme is covered in P1, P2, P4, P7, P10, P11, P12, P14, P20, P21, P22, P25, P28, P29, P30, P32, P33. In P1, the use of Natural Language Processing (NLP) on Electronic Health Records (EHR) data highlights multiple challenges, such as accurately identifying clinical entities, ensuring privacy protection, handling spelling errors, and managing the de-identification of sensitive information. These challenges are compounded by the lack of labeled data, the difficulty of detecting negations, and the complexities of deciphering numerous medical abbreviations. In P2, Gonen and Goldberg (2019) critique current debiasing methods for word vectors. While these techniques produce high scores on self-defined metrics, they often only superficially hide biases rather than genuinely eliminate them. P25

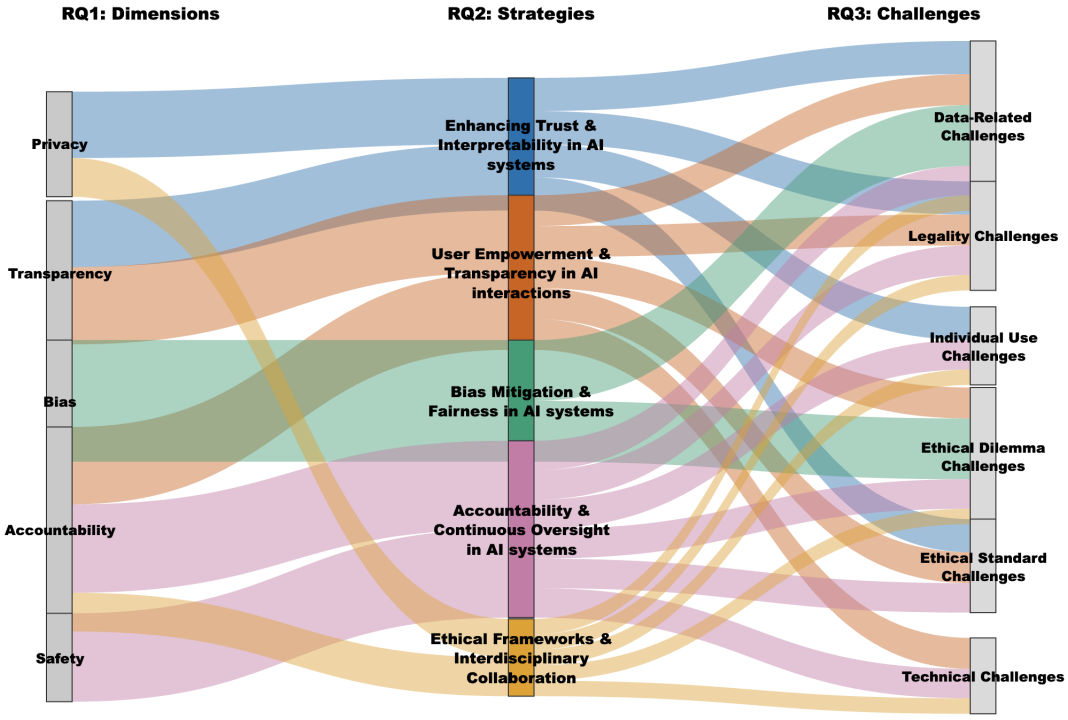

Figure 4: Flows Between Ethical Dimensions (RQ1), Strategies (RQ2), and Challenges (RQ3)

Table 8: RQ3 at-a-Glance: Challenge Categories and Typical Manifestations

Challenge category	**Typical issues observed**
Technical	limited robustness/generalization; debiasing limits; ongoing security testing (red teaming)
Ethical standards	evolving regulations; lack of global standardization; context-sensitive requirements
Ethical dilemmas	transparency vs. secrecy; fairness vs. learned associations; privacy vs. oversight
Individual use	over-trust/misuse; uncritical adoption; stakeholder gaming or agenda-driven behavior
Legality	IP/confidentiality risks; censorship and moderation ambiguity; unclear legal standards
Data-related	data quality/coverage; limited access; future scarcity of high-quality training data

emphasizes that Red Teaming, a method used to identify security vulnerabilities, must be a continuous process and not a one-time solution. It requires a

persistent commitment to security throughout the development and operation of AI systems, with careful consideration of both legal and ethical implications. These examples demonstrate the multifaceted technical hurdles that complicate the practical application of ethical mitigation strategies in LLMs.

- *Ethical Standard Challenges*: Implementing ethical standards in the context of LLMs is another challenging task, particularly due to the dynamic and diverse regulatory and cultural environments in which these technologies operate, this theme inclues P2, P8, P10, P30, P32, P34. In P10, the proposed recommendations are based on the analysis of the current regulatory landscape and anticipated regulations by the European Commission. However, as regulations continue to evolve, these recommendations may need to be revised to remain applicable and relevant. This highlights the importance of staying updated with the latest legal requirements in each jurisdiction to ensure ongoing compliance. P34 underscores the significance of incorporating a user-centered approach in cybersecurity, recognizing that addressing human factors can enhance the effectiveness of these measures. Yet, as cybersecurity threats evolve, there is a need for continuous development of strategies to stay ahead of emerging challenges. In P8, it is noted that a one-size-fits-all approach to ethical standards is unlikely to be effective, as LLMs span diverse cultural and global contexts. The absence of standardized or universally regulated frameworks for LLMs adds to the complexity, requiring adaptable and context-sensitive solutions. These examples illustrate the challenges of creating consistent and effective ethical standards in a rapidly changing and diverse landscape.

- *Ethical Dilemma Challenges*: Ethical dilemmas in the implementation of LLMs often arise from conflicting values and considerations, especially when navigating fairness, transparency, and privacy. In P3, the issue of learned associations within AI models highlights a significant ethical dilemma, this theme is covered in P3, P5, P8, P9, P10. For example, in P3, language models trained on data that frequently describe suspects and criminals as "black males" may reinforce stereotypes, raising concerns about legal equality. This brings up the ethical question of whether such associations, which exist in the training data, should be allowed in AI systems deployed for legal purposes, where fairness and equality are paramount. In P5, another ethical dilemma is identified in balancing transparency and protecting non-public information. While transparency is crucial for understanding an AI system's behavior and decision-making process, companies often prioritize commercial secrecy to protect proprietary technology and business models. These conflicts illustrate the complexities of ethical

decision-making in AI, where satisfying one ethical principle may compromise another, necessitating careful consideration and context-specific solutions.

- *Individual Use Challenges*: Implementing mitigation strategies for LLMs encounters specific challenges related to how individuals use and interpret AI-generated content, this theme is covered in P13, P25, P27, P30. In P13, even with established guidelines and recommendations, reports have surfaced in the healthcare and wellness sectors where individuals consult LLMs like ChatGPT for health-related matters and take its advice without proper scrutiny. Despite mitigation efforts, the widespread sharing of such AI-generated content on social media demonstrates the difficulty in ensuring that users critically assess AI advice. In P25, engaging multiple stakeholders introduces further complications, even when measures are in place to involve diverse perspectives. Some stakeholders may use their involvement to advance personal agendas, reinforce existing biases, or misuse sensitive data, undermining transparency and ethical intentions. These challenges illustrate the complexities in managing individual behavior and ensuring responsible AI use, despite mitigation strategies aimed at fostering ethical and informed engagement with LLMs.

- *Legality Challenges*: Legal challenges in implementing mitigation strategies for LLMs often involve navigating the complexities of intellectual property, censorship, and transparency, this theme is covered in P7, P8, P30. In P7, a significant concern is the potential leakage of business secrets and proprietary information when users interact with LLMs such as ChatGPT. If proprietary code is inadvertently shared during AI interactions, it may become part of the chatbot's knowledge base, raising issues of copyright infringement and the preservation of business confidentiality. This poses a risk for organizations that depend on protecting sensitive information. In P8, censorship within LLMs, while intended to prevent harmful outputs, introduces legal dilemmas. There is often no clear or objective standard for determining what content is harmful, which can lead to the suppression of free speech or creative expression. Overly restrictive censorship may also hinder important debates, while a lack of transparency around censorship policies can create distrust in the AI system. These examples underscore the legal complexities of implementing effective mitigation strategies, where balancing ethical considerations with regulatory compliance is a persistent challenge.

- *Data-related Challenges*: Data-related challenges are a critical factor in the implementation of mitigation strategies for LLMs, as they affect the quality,

availability, and reliability of the datasets used in AI development, this theme is covered in P11, P12, P28, P32. In P28, concerns are raised about the future limitations of data collection and usage in machine learning. Research indicates that high-quality language data could be exhausted by 2026, with lower-quality data potentially running out by 2060. This forecast suggests that the limited availability of suitable datasets may constrain the future development and improvement of LLMs, affecting their ability to perform effectively and ethically. In P32, the quality of data is further questioned, as a substantial portion of source material comes from preprint servers that lack rigorous peer review. This reliance on unverified data can limit the generalizability and reliability of LLMs, particularly when data is drawn from diverse and variable contexts. These examples highlight the significant data challenges encountered when implementing mitigation strategies, where data quality, scope, and future availability play a crucial role in shaping the effectiveness of LLM interventions.

RQ3 Key Takeaways

Implementation challenges affect 87.2% of studies, with technical barriers (NLP accuracy, bias detection limits) being the most prevalent (17 studies). Data constraints raise concerns that identifying high-quality training data may be exhausted by 2026. Legal challenges appear in three concrete forms: IP disputes over proprietary training data, censorship and transparency trade-offs in content moderation, and a lack of clear legal standards for AI-generated content. Ethical dilemmas reveal fundamental tensions: transparency and commercial secrecy, bias removal and associations in legal contexts. Significantly, user behavior challenges demonstrate that guidelines alone fail without enforcement mechanisms; individuals still uncritically accept AI-generated medical advice despite explicit warnings.

5. Discussion

Our systematic mapping reveals that ethical concerns around LLMs are not merely "multifaceted" in the abstract: they are weighted very differently across domains, lag behind rapid technical change, and are governed by fragmented, sometimes conflicting frameworks. Across the studies we analyzed, three recurring tensions stand out: (1) Ethical priorities shift more across domains than across high-level principles; (2) Mitigation strategies are often treated as one-off fixes rather than living processes; and (3) The most sophisticated technical mitigations are also the hardest to scale. In addition, end users and marginalized communities are still rarely

positioned as co-designers of LLM systems. Below, we unpack these tensions and translate them into concrete recommendations.

5.1. Domain-specific ethical priorities and implementation gaps

Our RQ1 analysis shows that the ethical dimensions, i.e., safety, privacy, accountability, bias, and transparency, are weighted differently across application domains. Safety is especially salient in healthcare and education and other public-facing contexts, where systems must prioritise safeguarding students and other vulnerable groups [56]. In more general AI applications, safety appears less frequently, while concerns shift towards bias and transparency in model development and deployment [57]. For example, in LLM-based recruitment tools that screen resumes and generate shortlist recommendations, the primary ethical concerns are bias (e.g., systematically favouring candidates from particular genders, ethnicities, or universities) and transparency (e.g., being able to explain why a candidate was shortlisted or rejected). Physical safety is not the central issue in this context; instead, the focus is on fairness.

Privacy emerges as a non-negotiable dimension in healthcare, where data sensitivity and regulatory mandates (e.g., HIPAA) make strong protections indispensable [21]. In economic and auditing contexts, privacy remains relevant but is often traded off against transparency and explainability, given the need for traceability of financial decisions and regulatory oversight [58]. Bias is most visible in legal and economic domains, where predictive models risk entrenching racial or socioeconomic disparities [59], whereas in some general AI applications transparency and interpretability are foregrounded even when fairness is less explicitly addressed. Accountability is particularly emphasised where legal liability or public trust are at stake, such as healthcare and government settings [60], but tends to recede in more experimental or early-stage deployments [61]. Transparency is a concern while using AI generated content in education [62].

Real-world deployments illustrate both the operationalisation of these priorities and persistent implementation gaps. In documentation systems, for example, safety and accountability are enforced through mandatory human oversight, with clinicians required to review all AI-generated notes before they are entered into the medical record [63]. At the governance level, frameworks such as the National Academy of Medicine's AI Code of Conduct emphasise continuous monitoring and equity metrics as core accountability mechanisms [64]. Yet a scoping review of 692 FDA-approved AI/ML medical devices found that only 3.6% reported race or ethnicity data, 81.6% omitted age information, and 99.1% provided no socioeconomic data [65]. This gap between governance frameworks that foreground equity and transparency, and

devices that rarely report basic demographic information, indicates that domain-specific ethical priorities often remain aspirational rather than fully implemented.

5.1.1. Cross-domain contrasts and under-addressed dimensions

While all five ethical dimensions appear across the 39 studies, Table 4 shows that they are not distributed evenly across domains, and this unevenness helps explain where implementation gaps persist. Cybersecurity work is heavily privacy-oriented (6 instances), with limited attention to safety (2) and bias (1), and no coverage of transparency or accountability (0, 0). Public safety studies, by contrast, emphasise safety (4) but show no explicit accountability (0) and minimal transparency (1). Education covers multiple dimensions, yet accountability remains low (1) relative to privacy (3) and transparency (2). Healthcare is the most balanced domain, with consistently high attention across all five dimensions (7/7/7/7/5), reflecting higher-stakes deployment and stronger compliance pressure. Across domains, this pattern indicates that governance-oriented dimensions, especially accountability (and, in some domains, transparency) remain systematically under-addressed relative to privacy and bias. This distribution aligns with our broader evidence gap: only a small proportion of studies empirically evaluate mitigation strategies, while many proposals remain conceptual (see RQ2 synthesis and Section 5.6). As a result, domains that most need accountability mechanisms in practice (e.g., public-facing and public-safety contexts) are also those where accountability is least operationalised in the mapped literature.

Recommendation: Make domain priorities concrete and auditable

1. **Domain-specific profiles** of high-level principles (e.g., safety, privacy, bias) be developed for sectors such as healthcare, education and finance, explicitly stating which dimensions are non-negotiable.
2. **Minimum reporting requirements** (e.g., for demographic coverage, performance by subgroup) be linked to these profiles, so that priorities like equity and transparency become auditable rather than symbolic.
3. **Oversight mechanisms** (such as mandatory human review in clinical documentation) be explicitly tied to the ethical risks that dominate each domain.

Such profiles help move from generic principles to concrete obligations that reflect domain-specific stakes.

5.2. Ethics as a living, continuous process

LLMs and their deployment environments change quickly: training data, model architectures, integration points and user populations all evolve over time. Our

review indicates that ethical strategies for bias, privacy and transparency are rarely designed with this dynamism in mind. Frameworks such as the EU AI Act and the NIST AI Risk Management Framework emphasise ongoing risk management and adaptation, but the majority of empirical work we identified still focuses on pre-deployment assessment or initial roll-out [54, 66]. There is comparatively little evidence on systematic post-deployment re-evaluation.

This static treatment of ethics is problematic because many failures emerge only after deployment, when models encounter new data distributions, novel uses or previously under-represented user groups [67]. Strategies that were adequate for earlier-generation systems may underperform for models trained on more diverse or sensitive data, or when LLMs are repurposed for higher-stakes decisions. The critique of "checkbox" ethics by Kijewski et al. [68] is reflected in our mapping: many guidelines define what should be considered at design time, but few specify how those considerations should be revisited as systems evolve.

Several organisations are beginning to frame ethics guidance explicitly as a living process. NASA, for instance, describes its ethical AI policies as "an evolving, living set of AI policies, principles, and guidelines" designed to remain responsive to emerging challenges and advances [69]. Similarly, frameworks such as Microsoft's Responsible AI Standard and IEEE guidelines foreground continuous accountability and monitoring rather than one-off compliance checks [70]. However, implementing such adaptive approaches requires operational structures for monitoring, auditing and updating systems over their lifecycle [71].

Recommendation: Operationalise ethics as a living process

1. **Tie ethics to lifecycle milestones**, linking ethical reviews to multi-year regulatory cycles (e.g., EU AI Act, NIST updates) and major model changes (new training data, architecture updates, or deployment contexts).
2. **Mandate post-deployment monitoring** for key metrics (e.g., performance by subgroup, privacy incidents, safety events), rather than treating evaluation as a one-time activity.
3. **Maintain an ethics change log** documenting detected issues, mitigation actions and residual risks, so that governance bodies can trace how ethical commitments evolve in practice.

Treating guidelines as living documents helps keep them specific, current and aligned with the behaviour of deployed LLMs.

5.3. Fragmented frameworks and cross-jurisdiction consistency

Implementing ethical standards for LLMs is further complicated by fragmented governance. Our mapping confirms prior observations that organisations must navigate a "bewildering variety" of AI ethics guidelines, many of which partially overlap but diverge in emphasis and legal force [59, 72]. The EU AI Act introduces binding requirements with substantial penalties (up to €35 million or 7% of global revenue) for non-compliance [73], whereas frameworks from NIST and IEEE are often voluntary and context-dependent. International organisations operating across jurisdictions thus face both compliance and harmonisation challenges [74, 75].

Fragmentation increases compliance costs and creates uneven protection for users. In healthcare, the FDA's January 2025 guidance for AI-enabled medical devices requires documentation across seven domains, adding another layer of domain-specific expectations [76]. In finance, institutions must juggle model risk guidance from bodies such as the Federal Reserve, FDIC and OCC [77, 78]. The US Government Accountability Office (GAO) noted in May 2025 that credit unions have no AI-specific oversight at all, despite rising adoption [79]. Large technology firms have responded by building internal governance structures; Microsoft, for example, operationalises six Responsible AI principles through dedicated governance offices and sector-specific guidance [80]. However, survey evidence from Thomson Reuters shows that only a minority of legal organisations have formal generative AI policies despite growing use [81], suggesting that fragmented frameworks disproportionately burden smaller firms and departments.

Recommendation: Establish a shared core of ethical principles

1. **Defining a core global baseline** of ethical principles for AI design, development, deployment and use (e.g., fairness, privacy, safety, transparency, accountability) that regulators and standard-setters can adopt and adapt.
2. **Layering domain- and jurisdiction-specific guidance** on top of this baseline, rather than creating entirely separate frameworks, to support organisations operating across regions.
3. **Providing implementation templates** (e.g., documentation structures, risk registers) to reduce compliance overhead for smaller organisations.

A shared core with adaptable layers can improve consistency while still allowing for local and sector-specific requirements.

5.4. Designing mitigations that scale in practice

Even when ethical principles and frameworks are clear, implementing mitigations at scale remains challenging. Our review identifies recurrent concerns about

the resource intensity and context-specificity of bias mitigation, privacy protection and transparency practices [82–84]. Many proposed interventions require specialised expertise and bespoke configuration, which can limit uptake beyond well-resourced organisations [85–87].

Empirical examples illustrate both the potential and limits of scalable approaches. Privacy-preserving techniques such as federated learning have demonstrated technical success at scale; the MELLODDY consortium, for instance, processed more than 2.6 billion proprietary data points across ten competing pharmaceutical companies while protecting each partner's data and improving aggregate model performance [88]. Yet a 2024 review found that only 5.2% of federated learning studies had reached real-world clinical implementation, despite promising technical results [89, 90]. This indicates that technical viability alone does not guarantee practical scalability.

Bias mitigation shows similar tensions. Recent evidence demonstrates that LLM outputs can change when only gender is altered: a study using gender-swapped social care case notes found that Google's Gemma downplayed women's health needs relative to men's (while Llama 3 showed no difference), and a hiring-style experiment observed that ChatGPT generated and evaluated CVs in ways that favoured older men over equivalently qualified women [91, 92]. Some of the most impactful corrective actions in adjacent domains have deliberately avoided algorithmic complexity. US nephrology adopted race-free eGFR equations in 2021, and the Organ Procurement and Transplantation Network's 2023 policy retroactively credited waiting time to Black kidney candidates, with early evidence of substantial time credits and higher transplant rates [93, 94]. Similarly, the American Heart Association's PREVENT equations omit race and allow optional inclusion of a ZIP code-based Social Deprivation Index to represent social risk [95]. These reforms scaled precisely because they simplified models and decision criteria rather than introducing more intricate technical fixes.

Recommendation: Prioritise scalable and simple mitigations

1. **Scalability be treated as a first-order design constraint** for mitigation strategies, not an afterthought, with explicit consideration of resource and expertise requirements.
2. **Simple, policy-level interventions** (e.g., removal of problematic variables, standardised reporting and review procedures) be prioritised where they offer comparable protection to complex technical fixes.
3. **Federated and privacy-preserving techniques** be paired with operational plans for deployment, governance and evaluation, to avoid remaining at

the proof-of-concept stage.
Designing mitigations with scalability in mind increases the likelihood that ethical commitments will be implemented beyond early adopters and well-resourced institutions.

5.5. Who defines "ethical"? End-user engagement and collaboration

Finally, our findings underline that ethical LLM deployment depends not only on technical and regulatory mechanisms but also on who is involved in defining and assessing harms. Historically, AI development has prioritised performance metrics over societal impact, with limited structured input from affected communities [96, 97]. This contributes to misalignment between LLM behaviour and the needs and values of those most affected, particularly marginalised groups.

Interdisciplinary collaboration has been promoted as one route to address these gaps, bringing together engineers, legal experts, ethicists and domain specialists [98, 99]. Examples such as the American Medical Informatics Association's 2023–24 multi-stakeholder initiative on AI-enabled clinical decision support — which convened over 200 clinicians, patients, regulators, ethicists and industry partners to co-develop recommendations including standardised AI labelling and a national safety-reporting mechanism — demonstrate the potential of structured collaboration [100]. At the same time, deployments such as Los Angeles Unified School District's "Ed" generative AI chatbot, which was suspended a few months after launch amid vendor collapse and concerns about transparency and data protection, illustrate the risks of insufficient stakeholder engagement and oversight [101].

High-profile controversies such as the COMPAS risk assessment tool, which misclassified Black defendants as high-risk nearly twice as often as White defendants who did not reoffend, while underestimating risk for White defendants who did reoffend [102], show the consequences of deploying systems without robust community input or contestability mechanisms. More constructive models remain concentrated in well-resourced institutions, such as IBM's AI Fairness 360 toolkit, which translates fairness research into practical bias metrics and mitigation algorithms across domains like finance, healthcare and education [103]. Reviews of explainable AI in medicine consistently highlight that trustworthy deployment requires substantial investment in transparency, governance and interdisciplinary teams [104]. Yet our mapping suggests that structured engagement with marginalised communities is still more often recommended than implemented, and power imbalances, differing priorities and communication barriers continue to hinder meaningful participation [105, 106].

Recommendation: Centre affected communities in LLM design and oversight

We encourage LLM developers and deployers to:
1. **Establish structured engagement mechanisms** such as community advisory boards, participatory design workshops and iterative user testing with representatives of marginalised groups, including racial and ethnic minorities, people with disabilities, LGBTQ+ communities and Indigenous peoples.
2. **Integrate community feedback into governance**, for example by incorporating stakeholder input into risk registers, model cards and deployment decisions.
3. **Support contestability**, by providing clear channels for users to challenge and appeal LLM-assisted decisions, and by tracking how such challenges lead to system changes.
Centring affected communities and interdisciplinary expertise helps ensure that LLM ethics move beyond abstract principles to address concrete, context-specific harms.

5.6. Methodological pathways for empirically validating ethical mitigation strategies

Our mapping indicates that 26 of 39 studies remain conceptual, highlighting the need for more systematic and replicable evaluation protocols. We propose here some future research directions outlining methodological pathways for empirically validating ethical mitigation strategies.

1. Standardized multi-metric benchmarking tests can be used to evaluate trade-offs across safety, bias and transparency outcomes under controlled settings, for example, scenario-driven evaluation with multiple metrics instead of accuracy reporting only [107].

2. Behavioral and adversarial testing can be incorporated to uncover the missing failure models by static benchmarks, examples include red teaming protocols, which search for harmful model behaviors and test-suite methods for NLP models [108, 109].

3. For privacy-focused mitigations, empirical validation can specifically measure leakage reduction using attack-based evaluations like membership inference and training-data extraction attacks, comparing models before and after mitigation under consistent threat models [36, 110].

4. For bias-focused mitigations, evaluation should include targeted bias benchmarks that operationalize stereotypical preferences, complemented by domain-specific datasets if applicable [111, 112].

5. For safety-related mitigations, empirical validation can include risk-focused benchmarks and stress tests that identify toxic degeneration and truthfulness failure models, which can be useful in domains like health [113, 114].

5.7. Implications for Applied Soft Computing: Opportunities for Ethical Mitigation

Our mapping highlights a persistent gap between conceptual governance proposals and empirically validated mitigation strategies (RQ2), as well as recurring implementation barriers, including limited robustness, domain shift, and data constraints (RQ3). These gaps suggest clear opportunities for the Applied Soft Computing (ASOC) community, since soft computing methods are designed to handle uncertainty, trade-offs, and complex constraints, all of which are central to ethical deployment of LLM-based systems.

Safety and reliability under uncertainty. Safety issues identified in RQ1 (e.g., hallucination, misinformation, reliability failures) and technical challenges in RQ3 indicate that mitigation must operate under uncertain model behaviour. Fuzzy inference or fuzzy risk scoring can serve as a lightweight control layer that converts signals (task criticality, uncertainty proxies, user context) into conservative actions such as abstention, escalation to humans, or stricter filtering in high-risk settings.

Bias–utility trade-offs and limited generalisation. RQ2 shows bias mitigation is among the most common technical strategies, yet studies report limited cross-domain generalisation. Evolutionary or other metaheuristic multi-objective optimisation offers a direct way to tune mitigation settings by explicitly balancing competing objectives (accuracy/utility vs fairness vs safety/privacy), rather than relying on a single fixed configuration.

Operationalising transparency and accountability. Our results indicate that transparency and accountability strategies remain predominantly conceptual and are less frequently validated empirically (RQ2). Hybrid AI approaches (e.g., LLMs combined with symbolic constraints or rule-based components, such as fuzzy rules) can make governance requirements executable and provide auditable decision pathways aligned with accountability needs.

6. Threats to Validity

In this section, we discuss the threats to the validity of our study.

Internal Validity: One of the key threats to internal validity in a systematic mapping study (SMS) is selection bias, which can arise from subjective interpretation during the study selection process. To mitigate this, we employed multiple strategies. We identified a set of keywords considered relevant to the ethical use of LLMs, tested them, and consulted a university librarian with expertise in systematic review search design, who recommended adding controlled vocabulary terms to refine the keywords. We conducted searches across six databases to ensure a wider variety of studies could be included. Pilot tests were performed and validated by all authors to ensure the reliability of the results. Additionally, forward snowballing was performed to capture studies that may not have been initially identified. Despite these measures, we acknowledge that selection bias may still have influenced the set of studies we analysed. Studies published outside our chosen computing databases were not included. These may focus on different ethical issues or propose alternative mitigation strategies. Therefore, our results should be viewed as reflecting only a focused picture of the broader literature.

Construct Validity: Construct validity in our study refers to the extent to which the selected studies are relevant and appropriate to our research goals. To address this, we selected papers that directly addressed our research questions (RQs) and excluded those that focused solely on LLMs without discussing the ethical issues associated with their use or deployment. We also held meetings and discussions to establish the inclusion and exclusion criteria for the selected papers. The timeframe for our SMS was set to 2023-July 2024, so any study published after July 2024 would not be reflected in our results. A potential limitation in this timeframe is that new studies published after our cutoff will not be included; therefore, our synthesis may not capture the most up-to-date developments. Additionally, as our study focused primarily on existing academic literature, it did not incorporate practitioner insights to a large extent, which might offer different perspectives on ethical concerns and mitigation strategies in practice.

External Validity: Our mapping primarily draws on studies originating in Western jurisdictions, each with its own cultural norms and regulatory frameworks. As a result, the universal ethical dimensions and mitigation strategies we identify may not directly translate into non-Western settings with different legal requirements or value systems. We intend to incorporate empirical work and guidelines from diverse regions, such as Asia, Africa, and Latin America, into our future work to validate findings across different cultural and legal landscapes. Furthermore, by excluding papers under four pages, we might have omitted brief but potentially relevant contributions. However, this criterion helped us focus on studies with sufficient

methodological and conceptual content. We would revisit these studies in our future work to gather more information. We also note that by focusing our study on English language publications, our findings may lean toward Western perspectives. This potential influence can be seen both in ethical priorities (for example, emphasised attention to privacy and accountability) and in feasible mitigation strategies. As a result, the conclusions, such as the significance of privacy concerns, should be interpreted as reflective of the literature base from this study, rather than universally generalizable across all cultural and regulatory contexts.

Conclusion Validity: One of the major threats to conclusion validity in systematic mapping studies is bias in data extraction. In our study, the data extraction process was guided by our RQs, ensuring that the selected data were directly relevant to our study objectives. To mitigate potential bias, we used Google Forms to facilitate coding, enabling us to systematically categorise data through thematic analysis. This approach enabled us to group the findings according to predefined codes and themes derived from the existing literature. As new themes emerged during the coding process, they were incorporated as needed. We held regular meetings to discuss and refine the data extraction and analysis processes, ensuring agreement on selecting relevant data and how it should be presented. Another concern regarding validity is the quality of evidence and the potential for publication bias. We acknowledge that a majority of our included studies (26 of 39) are conceptual, which may introduce author perspective bias into the reported strategies and skew the prominence of specific ethical dimensions. Moreover, our synthesis may be affected by publication bias, since positive or novel conceptual contributions are more likely to appear in the literature. To mitigate this, we searched across multiple databases and manually checked reference lists; we will also incorporate empirical investigations to validate the results from the conceptual studies. Consequently, our conclusions about the ethical dimensions and the reported effectiveness of mitigation strategies may present an optimistic view, as conceptual papers often propose solutions without rigorous testing. Our results should be interpreted as highlighting broad trends and gaps in the literature, rather than providing definitive measures in practice.

7. Conclusion

This systematic mapping study of 39 papers on generative AI ethics reveals three novel insights through cross-domain analysis. First, we demonstrate a critical evaluation gap: only 5 of 39 studies (12.8%) conducted a full empirical evaluation of their proposed mitigation strategies, while 26 papers (66.7%) proposed strategies that

remain conceptual and untested. Second, we identify a dimension and strategy misalignment: bias was the most discussed concern but received substantial conceptual attention and lacked empirical validation. Third, we find that although several technical challenges were addressed through algorithmic measures and ethical challenges were addressed through governance, there was limited cross-disciplinary integration. Our analysis identified 130 ethical issues across five dimensions: safety, transparency, accountability, privacy and bias, revealing that while numerous mitigation strategies have been proposed, implementation challenges remain across technical, legal, and ethical standards, individual use, and data-related domains.

7.1. Future Research Directions and Policy Implications

Our findings point to three critical directions for future work:

From conceptual proposals to validated solutions The high ratio of strategies lacking full empirical validation indicates a systemic issue in research. Future work should not only propose new strategies but also systematically test existing ones. Research could prioritise replication, validation and comparative effectiveness studies that close the gap between conceptual proposals and deployable solutions.

Developing domain-specific evaluation standards Our cross-domain analysis revealed that domains with established evaluation methodologies (healthcare with protocols) show higher validation rates than emerging domains (education, legal services). Future research should develop standardised evaluation benchmarks, for example, appropriate metrics for ethical dimensions in practice, diversity requirements and protocols for post-deployment.

Operationalising dimensions with strategy We identified that certain ethical dimensions, like bias, were discussed in 62% of papers, but had low validation rates. This suggests a lack of measurable outcomes. Future work should focus on developing concrete metrics for abstract ethical dimensions like accountability and bias. How do we empirically measure improvements in "accountability" when implementing measures? How to construct a valid bias test across different cultural contexts? These measurement challenges require work beyond algorithmic metrics.

The ethical deployment of LLMs focuses on closing the gaps we identified: between conceptual proposals and validated solutions, between ethical dimensions in attention and action, and between disciplines in problem-solving approaches. Future work must move beyond identifying concerns to systematically testing and deploying solutions that protect the millions of users already interacting with these systems.

Acknowledgements

Prof John Grundy is supported by the Australian Research Council (ARC) under the Future Fellowship grant number FL190100035. The authors thank our librarian (wishing to remain anonymous) for her expert assistance in designing and validating our search strategy.

References

[1] M. Alawida, S. Mejri, A. Mehmood, B. Chikhaoui, O. Isaac Abiodun, A comprehensive study of chatgpt: advancements, limitations, and ethical considerations in natural language processing and cybersecurity, Information 14 (8) (2023) 462.

[2] C. Arora, J. Grundy, M. Abdelrazek, Advancing requirements engineering through generative ai: Assessing the role of llms, in: Generative AI for Effective Software Development, Springer, 2024, pp. 129–148.

[3] A. A. Linkon, M. Shaima, M. S. U. Sarker, N. Nabi, M. N. U. Rana, S. K. Ghosh, M. A. Rahman, H. Esa, F. R. Chowdhury, et al., Advancements and applications of generative artificial intelligence and large language models on business management: A comprehensive review, Journal of Computer Science and Technology Studies 6 (1) (2024) 225–232.

[4] K. Huang, F. Wang, Y. Huang, C. Arora, Prompt engineering for requirements engineering: A literature review and roadmap, arXiv preprint arXiv:2507.07682 (2025).

[5] D. H. Hagos, R. Battle, D. B. Rawat, Recent advances in generative ai and large language models: Current status, challenges, and perspectives, IEEE Transactions on Artificial Intelligence (2024).

[6] S. Bengesi, H. El-Sayed, M. K. Sarker, Y. Houkpati, J. Irungu, T. Oladunni, Advancements in generative ai: A comprehensive review of gans, gpt, autoencoders, diffusion model, and transformers., IEEE Access (2024).

[7] S. Reddy, Generative ai in healthcare: an implementation science informed translational path on application, integration and governance, Implementation Science 19 (1) (2024) 27.

[8] Y. J. P. Bautista, C. Theran, R. Aló, Ethical considerations of generative ai: A survey exploring the role of decision makers in the loop, in: Proceedings of the AAAI Symposium Series, Vol. 3, 2024, pp. 391–398.

[9] N. Bontridder, Y. Poullet, The role of artificial intelligence in disinformation, Data & Policy 3 (2021) e32.

[10] F. Germani, G. Spitale, N. Biller-Andorno, The dual nature of ai in information dissemination: Ethical considerations, JMIR AI 3 (2024) e53505–.

[11] T. W. Sanchez, M. Brenman, X. Ye, The ethical concerns of artificial intelligence in urban planning, Journal of the American Planning Association (2024) 1–14.

[12] A. Rezaei Nasab, M. Dashti, M. Shahin, M. Zahedi, H. Khalajzadeh, C. Arora, P. Liang, Fairness concerns in app reviews: A study on ai-based mobile apps, ACM Transactions on Software Engineering and Methodology (2024).

[13] B. Kitchenham, L. Madeyski, D. Budgen, Segress: Software engineering guidelines for reporting secondary studies, IEEE Transactions on Software Engineering 49 (3) (2022) 1273–1298.

[14] K. Petersen, S. Vakkalanka, L. Kuzniarz, Guidelines for conducting systematic mapping studies in software engineering: An update, Information and software technology 64 (2015) 1–18.

[15] F. Li, N. Ruijs, Y. Lu, Ethics & ai: A systematic review on ethical concerns and related strategies for designing with ai in healthcare, Ai 4 (1) (2022) 28–53.

[16] J. Morley, L. Floridi, L. Kinsey, A. Elhalal, From what to how: an initial review of publicly available ai ethics tools, methods and research to translate principles into practices, Science and engineering ethics 26 (4) (2020) 2141–2168.

[17] E. Atlam, M. Almaliki, A. Alfahaid, I. Gad, G. Elmarhomy, M. Alwateer, A. Ahmed, Slm-aie: A systematic literature map of artificial intelligence ethics (2024).

[18] J. Dewey, J. H. Tufts, Ethics, DigiCat, 2022.

[19] S. Sivasubramaniam, D. H. Dlabolová, V. Kralikova, Z. R. Khan, Assisting you to advance with ethics in research: an introduction to ethical governance and application procedures, International Journal for Educational Integrity 17 (2021) 1–18.

[20] D. C. Poff, Academic ethics and academic integrity, in: Encyclopedia of business and professional ethics, Springer, 2023, pp. 11–16.

[21] A. Jobin, M. Ienca, E. Vayena, The global landscape of ai ethics guidelines, Nature machine intelligence 1 (9) (2019) 389–399.

[22] S. L. Anderson, M. Anderson, Ai and ethics, AI and Ethics 1 (1) (2021) 27–31.

[23] J.-C. Põder, Ai ethics–a review of three recent publications (2021).

[24] E. Kazim, A. S. Koshiyama, A high-level overview of ai ethics, Patterns 2 (9) (2021).

[25] M. Coeckelbergh, AI ethics, Mit Press, 2020.

[26] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, L. Floridi, The ethics of algorithms: Mapping the debate, Big Data & Society 3 (2) (2016) 2053951716679679.

[27] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, et al., Ai4people—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations, Minds and machines 28 (2018) 689–707.

[28] R. Eitel-Porter, Beyond the promise: implementing ethical ai, AI and Ethics 1 (1) (2021) 73–80.

[29] E. Hickman, M. Petrin, Trustworthy ai and corporate governance: the eu's ethics guidelines for trustworthy artificial intelligence from a company law perspective, European Business Organization Law Review 22 (2021) 593–625.

[30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Advances in neural information processing systems 27 (2014).

[31] A. Vaswani, Attention is all you need, Advances in Neural Information Processing Systems (2017).

[32] T. B. Brown, Language models are few-shot learners, arXiv preprint arXiv:2005.14165 (2020).

[33] V. Alto, Modern Generative AI with ChatGPT and OpenAI Models: Leverage the capabilities of OpenAI's LLM for productivity and innovation with GPT3 and GPT4, Packt Publishing Ltd, 2023.

[34] E. M. Bender, T. Gebru, A. McMillan-Major, 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.

[35] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature machine intelligence 1 (5) (2019) 206–215.

[36] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, in: 2017 IEEE symposium on security and privacy (SP), IEEE, 2017, pp. 3–18.

[37] J. Li, W. Xiao, C. Zhang, Data security crisis in universities: identification of key factors affecting data breach incidents, Humanities and Social Sciences Communications 10 (1) (2023) 1–18.

[38] H. Kibriya, W. Z. Khan, A. Siddiqa, M. K. Khan, Privacy issues in large language models: a survey, Computers and Electrical Engineering 120 (2024) 109698.

[39] T. Madiega, Artificial intelligence act, European Parliament: European Parliamentary Research Service (2021).

[40] E. Jillson, Aiming for truth, fairness, and equity in your company's use of ai, Federal Trade Commission 19 (2021).

[41] N. AI, Artificial intelligence risk management framework: Generative artificial intelligence profile (2024).

[42] S. Migliorini, China's interim measures on generative ai: Origin, content and significance, Computer Law & Security Review 53 (2024) 105985.

[43] I. S. Association, et al., The ieee global initiative on ethics of autonomous and intelligent systems. ieee. org. retrieved march 12, 2021 (2017).

[44] S. Pichai, Ai at google: our principles, The Keyword 7 (2018) (2018) 1–3.

[45] OpenAI, Openai safety standards, https://openai.com/safety-standards/ (2024).

[46] M. C. Buiten, Towards intelligent regulation of artificial intelligence, European Journal of Risk Regulation 10 (1) (2019) 41–59.

[47] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems, version 2, http://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf (2017).

[48] N. AI, Artificial intelligence risk management framework (ai rmf 1.0) (2023).

[49] Microsoft, Microsoft responsible ai standard, version 2: General requirements, https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf (2022).

[50] A. Oketunji, M. Anas, D. Saina, Large language model (llm) bias index—llmbi, Data & Policy (2023).

[51] M. T. Ribeiro, S. Singh, 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, 2016, pp. 1135–1144.

[52] B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Artificial intelligence and statistics, PMLR, 2017, pp. 1273–1282.

[53] P. on AI, Annual report 2021 (2021).
URL https://partnershiponai.org/resource/annual-report-2021/

[54] G. Palumbo, D. Carneiro, V. Alves, Objective metrics for ethical ai: a systematic literature review, International Journal of Data Science and Analytics (2024) 1–21.

[55] C. Wohlin, Guidelines for snowballing in systematic literature studies and a replication in software engineering, in: Proceedings of the 18th international conference on evaluation and assessment in software engineering, 2014, pp. 1–10.

[56] D. Leslie, Understanding artificial intelligence ethics and safety, arXiv preprint arXiv:1906.05684 (2019).

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

[58] V. Dignum, Responsible artificial intelligence: how to develop and use AI in a responsible way, Vol. 2156, Springer, 2019.

[59] T. Hagendorff, The ethics of ai ethics: An evaluation of guidelines, Minds and machines 30 (1) (2020) 99–120.

[60] A. Anawati, H. Fleming, M. Mertz, J. Bertrand, J. Dumond, S. Myles, J. Leblanc, B. Ross, D. Lamoureux, D. Patel, et al., Artificial intelligence and social accountability in the canadian health care landscape: A rapid literature review, PLOS Digital Health 3 (9) (2024) e0000597.

[61] H. Smith, Clinical ai: opacity, accountability, responsibility and liability, Ai & Society 36 (2) (2021) 535–545.

[62] M. Zahariev, et al., Legal and ethical challenges from copyright perspective of implementing artificial intelligence in education, in: Conference Proceedings. The Future of Education 2024, 2024.

[63] K. Lawrence, V. S. Kuram, D. L. Levine, S. Sharif, C. Polet, K. Malhotra, K. Owens, Informed consent for ambient documentation using generative ai in ambulatory care, JAMA Network Open 8 (7) (2025) e2522400–e2522400.

[64] N. A. of Medicine, An artificial intelligence code of conduct for health and medicine: Essential guidance for aligned action (2025).

[65] V. Muralidharan, B. A. Adewale, C. J. Huang, M. T. Nta, P. O. Ademiju, P. Pathmarajah, M. K. Hang, O. Adesanya, R. O. Abdullateef, A. O. Babatunde, et al., A scoping review of reporting gaps in fda-approved ai medical devices, NPJ Digital Medicine 7 (1) (2024) 273.

[66] N. H. Shah, M. A. Pfeffer, M. Ghassemi, The need for continuous evaluation of artificial intelligence prediction algorithms, JAMA Network Open 7 (9) (2024) e2433009–e2433009.

[67] R. Ortega-Bolaños, J. Bernal-Salcedo, M. Germán Ortiz, J. Galeano Sarmiento, G. A. Ruz, R. Tabares-Soto, Applying the ethics of ai: a systematic review of tools for developing and assessing ai-based systems, Artificial Intelligence Review 57 (5) (2024) 110.

[68] S. Kijewski, E. Ronchi, E. Vayena, The rise of checkbox ai ethics: a review, AI and Ethics (2024) 1–10.

[69] E. McLarney, Y. Gawdiak, N. Oza, C. Mattmann, M. Garcia, M. Maskey, S. Tashakkor, D. Meza, J. Sprague, P. Hestnes, et al., Nasa framework for the ethical use of artificial intelligence (ai) (2021).

[70] M. Srikumar, R. Finlay, G. Abuhamad, C. Ashurst, R. Campbell, E. Campbell-Ratcliffe, H. Hongo, S. R. Jordan, J. Lindley, A. Ovadya, et al., Advancing ethics review practices in ai research, Nature Machine Intelligence 4 (12) (2022) 1061–1064.

[71] Y. Li, S. Goel, Making it possible for the auditing of ai: A systematic review of ai audits and ai auditability, Information Systems Frontiers (2024) 1–31.

[72] E. Prem, From ethical ai frameworks to tools: a review of approaches, AI and Ethics 3 (3) (2023) 699–716.

[73] European Commission, Article 99: Penalties (AI act service desk) (2024). URL https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-99

[74] A. A. Khan, M. A. Akbar, M. Fahmideh, P. Liang, M. Waseem, A. Ahmad, M. Niazi, P. Abrahamsson, Ai ethics: an empirical study on the views of practitioners and lawmakers, IEEE Transactions on Computational Social Systems 10 (6) (2023) 2971–2984.

[75] V. Vakkuri, K.-K. Kemell, Implementing ai ethics in practice: An empirical evaluation of the resolvedd strategy, in: Software Business: 10th International Conference, ICSOB 2019, Jyväskylä, Finland, November 18–20, 2019, Proceedings 10, Springer, 2019, pp. 260–275.

[76] U.S. Food and Drug Administration, Artificial intelligence-enabled device software functions: Lifecycle management and marketing submission recommendations: Draft guidance for industry and food and drug administration staff, draft guidance document issued January 7, 2025 (Jan. 2025). URL https://www.fda.gov/media/184856/download

[77] Office of the Comptroller of the Currency, Sound practices for model risk management: Supervisory guidance on model risk management, oCC Bulletin 2011-12, issued April 4, 2011 (Apr. 2011).
URL https://www.occ.gov/news-issuances/bulletins/2011/bulletin-2011-12.html

[78] Federal Deposit Insurance Corporation, Adoption of supervisory guidance on model risk management, financial Institution Letter FIL-22-2017, issued June 7, 2017 (Jun. 2017).
URL https://www.fdic.gov/news/financial-institution-letters/2017/fil17022.html

[79] U.S. Government Accountability Office, Artificial intelligence: Use and oversight in financial services, report to Congressional Committees, GAO-25-107197, issued May 19, 2025 (May 2025).
URL https://www.gao.gov/assets/gao-25-107197.pdf

[80] Microsoft Corporation, Microsoft responsible ai standard, v2: General requirements, for external release (Jun. 2022).
URL https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf

[81] Thomson Reuters Institute, 2024 generative ai in professional services: Perceptions, usage & impact on the future of work, special report on generative AI in professional services (2024).
URL https://www.thomsonreuters.com/content/dam/ewp-m/documents/thomsonreuters/en/pdf/reports/tr4322226_rgb.pdf

[82] C. Deng, Y. Duan, X. Jin, H. Chang, Y. Tian, H. Liu, H. P. Zou, Y. Jin, Y. Xiao, Y. Wang, et al., Deconstructing the ethics of large language models from long-standing issues to new-emerging dilemmas, arXiv preprint arXiv:2406.05392 (2024).

[83] A. A. Khan, S. Badshah, P. Liang, M. Waseem, B. Khan, A. Ahmad, M. Fahmideh, M. Niazi, M. A. Akbar, Ethics of ai: A systematic literature review of principles and challenges, in: Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering, 2022, pp. 383–392.

[84] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, M. Srikumar, Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai, Berkman Klein Center Research Publication (2020-1) (2020).

[85] R. Binns, Fairness in machine learning: Lessons from political philosophy, in: Conference on fairness, accountability and transparency, PMLR, 2018, pp. 149–159.

[86] W. Hoffmann-Riem, Artificial intelligence as a challenge for law and regulation, Regulating artificial intelligence (2020) 1–29.

[87] C. Cath, Governing artificial intelligence: ethical, legal and technical opportunities and challenges, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 376 (2133) (2018) 20180080.

[88] W. Heyndrickx, L. Mervin, T. Morawietz, N. Sturm, L. Friedrich, A. Zalewski, A. Pentina, L. Humbeck, M. Oldenhof, R. Niwayama, et al., Melloddy: Cross-pharma federated learning at unprecedented scale unlocks benefits in qsar without compromising proprietary information, Journal of chemical information and modeling 64 (7) (2023) 2331–2344.

[89] Z. L. Teo, L. Jin, N. Liu, S. Li, D. Miao, X. Zhang, W. Y. Ng, T. F. Tan, D. M. Lee, K. J. Chua, et al., Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture, Cell Reports Medicine 5 (2) (2024).

[90] I. Siniosoglou, S. Bibi, K.-F. Kollias, G. Fragulis, P. Radoglou-Grammatikis, T. Lagkas, V. Argyriou, V. Vitsas, P. Sarigiannidis, Federated learning models in decentralized critical infrastructure, Shaping the Future of IoT with Edge Intelligence (2024) 95–115.

[91] S. Rickman, Evaluating gender bias in large language models in long-term care, BMC Medical Informatics and Decision Making 25 (1) (2025) 274.

[92] D. Guilbeault, S. Delecourt, B. S. Desikan, Age and gender distortion in online media and large language models, Nature (2025) 1–9.

[93] L. A. Inker, N. D. Eneanya, J. Coresh, H. Tighiouart, D. Wang, Y. Sang, D. C. Crews, A. Doria, M. M. Estrella, M. Froissart, et al., New creatinine-and cystatin c–based equations to estimate gfr without race, New England Journal of Medicine 385 (19) (2021) 1737–1749.

[94] A. L. Hoffman, S. G. Westphal, D. Wekesa, C. D. Miles, Impact of optn policy 3.7 d providing waiting time modification for candidates affected by race-inclusive egfr calculations: early results from a single center, Clinical Transplantation 38 (3) (2024) e15273.

[95] S. S. Khan, K. Matsushita, Y. Sang, S. H. Ballew, M. E. Grams, A. Surapaneni, M. J. Blaha, A. P. Carson, A. R. Chang, E. Ciemins, et al., Development and validation of the american heart association's prevent equations, Circulation 149 (6) (2024) 430–449.

[96] K. Murphy, E. Di Ruggiero, R. Upshur, D. J. Willison, N. Malhotra, J. C. Cai, N. Malhotra, V. Lui, J. Gibson, Artificial intelligence for good health: a scoping review of the ethics literature, BMC medical ethics 22 (2021) 1–17.

[97] A. Hagerty, I. Rubinov, Global ai ethics: a review of the social impacts and ethical implications of artificial intelligence, arXiv preprint arXiv:1907.07892 (2019).

[98] S. Pink, E. Quilty, J. Grundy, R. Hoda, Trust, artificial intelligence and software practitioners: an interdisciplinary agenda, AI & SOCIETY (2024) 1–14.

[99] M. Al-kfairy, D. Mustafa, N. Kshetri, M. Insiew, O. Alfandi, Ethical challenges and solutions of generative ai: An interdisciplinary perspective, in: Informatics, Vol. 11, MDPI, 2024, p. 58.

[100] S. Labkoff, B. Oladimeji, J. Kannry, A. Solomonides, R. Leftwich, E. Koski, A. L. Joseph, M. Lopez-Gonzalez, L. A. Fleisher, K. Nolen, et al., Toward a responsible future: recommendations for ai-enabled clinical decision support, Journal of the American Medical Informatics Association 31 (11) (2024) 2730–2739.

[101] J. Young, An education chatbot company collapsed. where did the student data go, EdSurge News. EdSurge (2024).

[102] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias, in: Ethics of data and analytics, Auerbach Publications, 2022, pp. 254–264.

[103] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, et al., Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias, IBM Journal of Research and Development 63 (4/5) (2019) 4–1.

[104] M. Frasca, D. La Torre, G. Pravettoni, I. Cutica, Explainable and interpretable artificial intelligence in medicine: a systematic bibliometric review, Discover Artificial Intelligence 4 (1) (2024) 15.

[105] S. Keles, Navigating in the moral landscape: analysing bias and discrimination in ai through philosophical inquiry, AI and Ethics (2023) 1–11.

[106] R. Gianni, S. Lehtinen, M. Nieminen, Governance of responsible ai: From ethical guidelines to cooperative policies, Frontiers in Computer Science 4 (2022) 873437.

[107] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, et al., Holistic evaluation of language models, arXiv preprint arXiv:2211.09110 (2022).

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

[109] M. T. Ribeiro, T. Wu, C. Guestrin, S. Singh, Beyond accuracy: Behavioral testing of nlp models with checklist, arXiv preprint arXiv:2005.04118 (2020).

[110] 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.

[111] M. Nadeem, A. Bethke, S. Reddy, Stereoset: Measuring stereotypical bias in pretrained language models, in: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers), 2021, pp. 5356–5371.

[112] N. Nangia, C. Vania, R. Bhalerao, S. Bowman, Crows-pairs: A challenge dataset for measuring social biases in masked language models, in: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), 2020, pp. 1953–1967.

[113] S. Gehman, S. Gururangan, M. Sap, Y. Choi, N. A. Smith, Realtoxicityprompts: Evaluating neural toxic degeneration in language models, arXiv preprint arXiv:2009.11462 (2020).

[114] S. Lin, J. Hilton, O. Evans, Truthfulqa: Measuring how models mimic human falsehoods, in: Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers), 2022, pp. 3214–3252.

8. Appendix A: Selected Primary Studies

P1 : Sathe, N., Deodhe, V., Sharma, Y., & Shinde, A. (2023, December). A Comprehensive Review of AI in Healthcare: Exploring Neural Networks in Medical Imaging, LLM-Based Interactive Response Systems, NLP-Based EHR Systems, Ethics, and Beyond. In 2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech) (pp. 633-640). IEEE.

P2 : Ruane, E., Birhane, A., & Ventresque, A. (2019). Conversational AI: Social and Ethical Considerations. AICS, 2563, 104-115.

P3 : Malic, V. Q., Kumari, A., & Liu, X. (2023, December). Racial skew in fine-tuned legal AI language models. In 2023 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 245-252). IEEE.

P4 : Bansal, R. (2022). A survey on bias and fairness in natural language processing. arXiv preprint arXiv:2204.09591.

P5 : Bang, J., Lee, B. T., & Park, P. (2023, August). Examination of ethical principles for llm-based recommendations in conversational ai. In 2023 International Conference on Platform Technology and Service (PlatCon) (pp. 109-113). IEEE.

P6 : Lofstead, J. (2023, September). Economic, Societal, Legal, and Ethical Considerations for Large Language Models. In 2023 Fifth International Conference on Transdisciplinary AI (TransAI) (pp. 155-162). IEEE.

P7 : Khoury, R., Avila, A. R., Brunelle, J., & Camara, B. M. (2023, October). How secure is code generated by chatgpt?. In 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 2445-2451). IEEE.

P8 : Jiao, J., Afroogh, S., Xu, Y., & Phillips, C. (2024). Navigating llm ethics: Advancements, challenges, and future directions. arXiv preprint arXiv:2406.18841.

P9 : Cai, Z., Chang, X., & Li, P. (2023, November). HCPP: A Data-Oriented Framework to Preserve Privacy during Interactions with Healthcare Chatbot. In 2023 IEEE International Performance, Computing, and Communications Conference (IPCCC) (pp. 283-290). IEEE.

P10 : Piñeiro-Martín, A., García-Mateo, C., Docío-Fernández, L., & Lopez-Perez, M. D. C. (2023). Ethical challenges in the development of virtual assistants powered by large language models. Electronics, 12(14), 3170.

P11 : Parray, A. A., Inam, Z. M., Ramonfaur, D., Haider, S. S., Mistry, S. K., & Pandya, A. K. (2023). ChatGPT and global public health: applications, challenges, ethical considerations and mitigation strategies.

P12 : Khan, M. S., & Umer, H. (2024). ChatGPT in finance: Applications, challenges, and solutions. Heliyon, 10(2).

P13 : Harrer, S. (2023). Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. EBioMedicine, 90.

P14 : Chauncey, S. A., & McKenna, H. P. (2023). A framework and exemplars for ethical and responsible use of AI Chatbot technology to support teaching and learning. Computers and Education: Artificial Intelligence, 5, 100182.

P15 : Ansarullah, S. I., Kirmani, M. M., Alshmrany, S., & Firdous, A. (2024). Ethical issues around artificial intelligence. In A Biologist s Guide to Artificial Intelligence (pp. 301-314). Academic Press.

P16 : Patton, D. U., Landau, A. Y., & Mathiyazhagan, S. (2023). ChatGPT for social work science: Ethical challenges and opportunities. Journal of the Society for Social Work and Research, 14(3), 553-562.

P17 : Wu, X., Duan, R., & Ni, J. (2024). Unveiling security, privacy, and ethical concerns of ChatGPT. Journal of Information and Intelligence, 2(2), 102-115.

P18 : Guo, D., Chen, H., Wu, R., & Wang, Y. (2023). AIGC challenges and opportunities related to public safety: a case study of ChatGPT. Journal of Safety Science and Resilience, 4(4), 329-339.

P19 : Oviedo-Trespalacios, O., Peden, A. E., Cole-Hunter, T., Costantini, A., Haghani, M., Rod, J. E., ... & Reniers, G. (2023). The risks of using ChatGPT

to obtain common safety-related information and advice. Safety science, 167, 106244.

P20 : Head, C. B., Jasper, P., McConnachie, M., Raftree, L., & Higdon, G. (2023). Large language model applications for evaluation: Opportunities and ethical implications. New directions for evaluation, 2023(178-179), 33-46.

P21 : MacIntyre, M. R., Cockerill, R. G., Mirza, O. F., & Appel, J. M. (2023). Ethical considerations for the use of artificial intelligence in medical decision-making capacity assessments. Psychiatry research, 328, 115466.

P22 : Grote, T., & Berens, P. (2024). A paradigm shift?—On the ethics of medical large language models. Bioethics, 38(5), 383-390.

P23 : Behnia, R., Ebrahimi, M. R., Pacheco, J., & Padmanabhan, B. (2022, November). Ew-tune: A framework for privately fine-tuning large language models with differential privacy. In 2022 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 560-566). IEEE.

P24 : Wang, C., Liu, S., Yang, H., Guo, J., Wu, Y., & Liu, J. (2023). Ethical considerations of using ChatGPT in health care. Journal of Medical Internet Research, 25, e48009.

P25 : Mitsunaga, T. (2023, October). Heuristic Analysis for Security, Privacy and Bias of Text Generative AI: GhatGPT-3.5 case as of June 2023. In 2023 IEEE International Conference on Computing (ICOCO) (pp. 301-305). IEEE.

P26 : Gan, W., Qi, Z., Wu, J., & Lin, J. C. W. (2023, December). Large language models in education: Vision and opportunities. In 2023 IEEE international conference on big data (BigData) (pp. 4776-4785). IEEE.

P27 : Kshetri, N. (2023). Cybercrime and privacy threats of large language models. IT Professional, 25(3), 9-13.

P28 : Zhuo, T. Y., Huang, Y., Chen, C., & Xing, Z. (2023). Red teaming chatgpt via jailbreaking: Bias, robustness, reliability and toxicity. arXiv preprint arXiv:2301.12867.

P29 : Zhang, X., Xu, H., Ba, Z., Wang, Z., Hong, Y., Liu, J., ... & Ren, K. (2024). Privacyasst: Safeguarding user privacy in tool-using large language model agents. IEEE Transactions on Dependable and Secure Computing.

P30 : Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023). Auditing large language models: a three-layered approach. AI and Ethics, 1-31.

P31 : Curzon, J., Kosa, T. A., Akalu, R., & El-Khatib, K. (2021). Privacy and artificial intelligence. IEEE Transactions on Artificial Intelligence, 2(2), 96-108.

P32 : Bano, M., Hoda, R., Zowghi, D., & Treude, C. (2024). Large language models for qualitative research in software engineering: exploring opportunities and challenges. Automated Software Engineering, 31(1), 8.

P33 : Khoje, M. (2024, February). Navigating Data Privacy and Analytics: The Role of Large Language Models in Masking conversational data in data platforms. In 2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC) (pp. 1-5). IEEE.

P34 : Jeyaraman, M., Balaji, S., Jeyaraman, N., & Yadav, S. (2023). Unraveling the ethical enigma: artificial intelligence in healthcare. Cureus, 15(8).

P35 : Ferrara, E. (2023). Should chatgpt be biased? challenges and risks of bias in large language models. arXiv preprint arXiv:2304.03738.

P36 : Akinci D'Antonoli, T., Stanzione, A., Bluethgen, C., Vernuccio, F., Ugga, L., Klontzas, M. E., ... & Koçak, B. (2023). Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. Diagnostic and Interventional Radiology, Epub-ahead.

P37 : Belzner, L., Gabor, T., & Wirsing, M. (2023, October). Large language model assisted software engineering: prospects, challenges, and a case study. In International Conference on Bridging the Gap between AI and Reality (pp. 355-374). Cham: Springer Nature Switzerland.

P38 : Liyanage, U. P., & Ranaweera, N. D. (2023). Ethical considerations and potential risks in the deployment of large language models in diverse societal contexts. Journal of Computational Social Dynamics, 8(11), 15-25.

P39 : Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J. (2024). Transforming assessment: The impacts and implications of large language models and generative ai. Educational Measurement: Issues and Practice, 43(2), 16-29.

9. Appendix B: Database Search Strings

Database	Search String
IEEE	(("Large Language Model*" OR LLMs) AND (guideline* OR standard* OR framework* OR compliance OR principles OR practices OR governance OR impact OR oversight OR algorithmic OR policy OR policies) AND (development OR deployment OR use OR design OR implementation) AND (ethics OR ethical OR moral OR bias OR fairness OR transparency OR accountability OR privacy OR security OR sustainability OR responsible OR trustworthiness OR equit* OR inclus* OR diversity OR legal OR rights OR cultural))
ACM DL	(([All: "large language model*"] OR [All: llms]) AND ([All: guideline*] OR [All: standard*] OR [All: framework*] OR [All: compliance] OR [All: principles] OR [All: practices] OR [All: governance] OR [All: impact] OR [All: oversight] OR [All: algorithmic] OR [All: policy] OR [All: policies]) AND ([All: development] OR [All: deployment] OR [All: use] OR [All: design] OR [All: implementation]) AND ([All: ethics] OR [All: ethical] OR [All: moral] OR [All: bias] OR [All: fairness] OR [All: transparency] OR [All: accountability] OR [All: privacy] OR [All: security] OR [All: sustainability] OR [All: responsible] OR [All: trustworthiness] OR [All: equit*] OR [All: inclus*] OR [All: diversity] OR [All: legal] OR [All: rights] OR [All: cultural]))
ProQuest	((noft("Large Language Model*") OR noft(LLMs)) AND (noft(guideline*) OR noft(standard*) OR noft(framework*) OR noft(compliance) OR noft(principles) OR noft(practices) OR noft(governance) OR noft(impact) OR noft(oversight) OR noft(algorithmic) OR noft(policy) OR noft(policies)) AND (noft(development) OR noft(deployment) OR noft(use) OR noft(design) OR noft(implementation)) AND (noft(ethics) OR noft(ethical) OR noft(moral) OR noft(bias) OR noft(fairness) OR noft(transparency) OR noft(accountability) OR noft(privacy) OR noft(security) OR noft(sustainability) OR noft(responsible) OR noft(trustworthiness) OR noft(equit*) OR noft(inclus*) OR noft(diversity) OR noft(legal) OR noft(rights) OR noft(cultural)))
Web of Science	(("Large Language Model*" OR llms) AND (guideline* OR standard* OR framework* OR compliance OR principles OR practices OR governance OR impact OR oversight OR algorithmic OR policy OR policies) AND (development OR deployment OR use OR design OR implementation) AND (ethics OR ethical OR moral OR bias OR fairness OR transparency OR accountability OR privacy OR security OR sustainability OR responsible OR trustworthiness OR equit* OR inclus* OR diversity OR legal OR rights OR cultural))
Wiley Online Library	(("Large Language Model*" OR llms) AND (guideline* OR standard* OR framework* OR compliance OR principles OR practices OR governance OR impact OR oversight OR algorithmic OR policy OR policies) AND (development OR deployment OR use OR design OR implementation) AND (ethics OR ethical OR moral OR bias OR fairness OR transparency OR accountability OR privacy OR security OR sustainability OR responsible OR trustworthiness OR equit* OR inclus* OR diversity OR legal OR rights OR cultural))
Science Direct	("Large Language Model" OR "LLMs") AND ("guideline" OR "standard") AND ("development" OR "design") AND ("ethics" OR "moral")