Ethical Issues in Large Language Models : A Systematic Literature Review

Laakso, Atte

2024

Laakso, A, Kemell, K K & Nurminen, J K 2024, Ethical Issues in Large Language Models : A Systematic Literature Review. in T Olsson, O Sahlgren, J Parviainen, S Westerstrand, J T Harviainen, A Laitinen & J Rantala (eds), Proceedings of the Conference on Technology Ethics 2024 (Tethics 2024). CEUR Workshop Proceedings, vol. 3901, CEUR-WS.org, Aachen, pp. 42-66, Conference on Technology Ethics, Tampere, Finland, 06/11/2024. < https://ceur-ws.org/Vol-3901/ >

http://hdl.handle.net/10138/592434

Ethical Issues in Large Language Models: A Systematic Literature Review

Atte Laakso[1], Kai-Kristian Kemell[2,*] and Jukka K. Nurminen[2]

[1]Naval Academy Finland, PL 5, 00191 Helsinki, Finland
[2]University of Helsinki, Department of Computer Science, Yliopistonkatu 3, 00014 University of Helsinki, Finland

Abstract

Large Language Models (LLMs), and Generative AI (GenAI) more generally, have been the center of much attention in both media and research following recent technical advances. In the wake of the recent surge of users services like ChatGPT and GitHub Copilot have seen, there has also been much discussion on the potential negative impacts of these tools, from job loss to cheating in schools by using AI, to security issues in AI-generated code. To this end, a large number of papers has recently been published on the ethical issues and negative impacts of LLMs, across a wide variety of scientific disciplines. To help tie this recent discussion on the ethical issues of LLMs to the existing discussion on AI ethics, we conduct a Systematic Literature Review (SLR) and review 116 papers. We extract 434 individual ethical issues from these papers, based on which we identify 39 different categories of ethical issues. We then map these 39 categories of ethical issues into the seven requirements for trustworthy AI found in the Ethics Guidelines for Trustworthy AI (AI HLEG) in order to understand how these concerns related to the existing discussion on AI ethics. While various new practical issues are identified in the process, on a conceptual level the issues found in LLMs are related to the ones already identified in AI ethics and can be related to existing AI ethics principles. Overall, this SLR provides a summary of the current discussion on ethical issues in LLMs.

Keywords

Artificial Intelligence Ethics, Large Language Models, ChatGPT, Systematic Literature Review, Ethical Principles

1. Introduction

Recent advances in the field of Generative AI (GenAI) have brought GenAI systems into the spotlight in AI ethics as well. For the most part, the focus has been on Large Language Models (LLMs) in particular. These new types of systems have resulted in another surge of discussion on ethical issues and risks associated with ML, especially in the media. We have also noticed a similar surge in research papers focusing on LLMs, with or without a focus on ethical aspects.

LLMs have had a massive impact on making ML systems available for consumer use, primarily through B2C cloud services that require no technical ML knowhow. For example, ChatGPT became the fastest growing service in history thus far, surpassing TikTok, Facebook, and other online services with a history of quick growth [1]. This surge in AI use has resulted in

7th Conference on Technology Ethics (TETHICS2024), November 6–7, 2024, Tampere, Finland
*Corresponding author.
atte.laakso@mil.fi (A. Laakso); kai-kristian.kemell@helsinki.fi (K. Kemell); jukka.k.nurminen@helsinki.fi (J. K. Nurminen)
0000-0002-0225-4560 (K. Kemell); 0000-0001-5083-1927 (J. K. Nurminen)

widespread discussion on the potential impacts of AI, as, for example, educators have actually felt firsthand the impacts of services such as ChatGPT in their classes [2]. Up until recently, AI was largely used by organizational actors, while private individuals were usually passive "users" (e.g., given recommendations by AI-powered recommendation systems) or just objects of data collection for AI systems. Moreover, especially LLM systems have a wide range of potential applications and can perform various tasks, whereas conventional AI systems are typically more narrow expert systems. Due to these reasons, among others, LLMs systems present new challenges from the point of view of AI ethics as well, as various actors are currently exploring their use across a multitude of potential use contexts.

Much of the existing discussion in the field of AI ethics has focused on the *development* of ethical ML systems. E.g., principles that ML systems need to adhere to in order to be ethical (see for example [3, 4]), or tools and methods to support the development of ethical AI systems (see [5]). On the contrary, as these GenAI tools are now widely available as both online services and models to be deployed locally for consumer use, various papers have recently been published on the ethical challenges associated with the *use* of these tools and its impacts instead of just development. Indeed, recent AI ethics discussion on GenAI has placed more emphasis on the ethical *use* of AI systems as well. Overall, a great number of papers have been published recently on the ethical challenges associated with GenAI and LLMs, across a number of disciplines (e.g., as seen in [6] or in this SLR).

These developments may present some points of reflection for AI ethics researchers. For example, to what extent are these issues new on a conceptual level? Are there any *new* issues related to fairness that should be considered, or is the novelty in the changed practical context? For example, if we are concerned about the biased output of LLMs, we are ultimately still concerned about bias just as we were concerned about it in relation to decision-making support systems. Yet, do these new system use contexts nonetheless present new challenges for AI ethics research? Given that AI ethics overall is a field ripe with various competing concepts and definitions [3, 7, 4], relating this new discussion to the existing discussion is valuable for the field.

To better understand the implications of LLMs for the existing AI ethics discussion, we consider systematically reviewing this recent surge of publications to be beneficial for the field. Thus, in this paper, we conduct a Systematic Literature Review (SLR) of publications discussing ethical challenges associated with GenAI, and specifically Large Language Models (LLMs). We utilize the Ethics Guidelines for Trustworthy AI [8] as a framework for conducting the SLR and reporting its results, mapping the identified risks and ethical issues under the principles found in said guidelines, in order to relate this new LLM discussion to the existing discussion in AI ethics. Based on the SLR, we provide a comprehensive list of the ethical issues existing research has now associated with LLM systems. We then relate these findings to existing literature on AI ethics to discuss their implications for the field.

2. Related Work

We consider literature reviews related to AI ethics related work in the context of this paper. Thus, the most closely related paper we have identified is that of Weidinger et al. [6] (also found

via the SLR as [S97]). The study in question proposes a taxonomy for grouping and discussing risks related to language models. While not an SLR, the aims of the paper are ultimately similar. The authors summarize risks associated with language models into 21 risks, grouped into 6 risk categories. We further discuss their results in relation to ours in Section 5, which we consider to compliment ours, and vice versa.

While various other literature (grey or otherwise) reviews related to AI exist, these are less related to our work in this SLR. Some reviews related to AI ethics include include the following. Jobin et al. [3] review AI ethics guidelines, summarizing the principles present within the guidelines in order to provide an overview of the most common principles, and to help systematize the discussion around the various principles. A similar review of guidelines is presented by Hagendorff [4]. Morley et al. [5] review *tools* for AI ethics, mapping them across five principles and different stages of the development process. Vakkuri & Abrahamsson [7] conduct a systematic mapping study on the key concepts used in the AI ethics discussion. Khan et al. [9] also conduct an SLR on AI ethics principles, with a focus on challenges associated with each principle. Finally, Selter et al. [10] conduct an SLR on ethics *and* morality in AI.

3. Literature Review Methodology

This section discusses the SLR protocol. Section 3.1 discusses the search strategy. Section 3.2 discusses the inclusion and exclusion process. Section 3.3 discusses the data extraction.

3.1. Search Strategy

For this SLR, we searched for literature from three databases: *ACM Digital Library*, *IEEE Xplore*, and *Scopus*. We employed the following search string for all three, with potential minor adjustments to adhere to the constraints of the search engine of each database:

"large language model" OR "chatgpt" AND "ethic*" OR "fair*" OR "transpar*" OR "explaina*" OR "trustworth*" OR "human agenc*" OR "oversight" OR "privacy" OR "diversity*" OR "discrimin*" AND "generat*"*

The first part of the search string was used to limit the search to these specific types of ML systems. ChatGPT as a specific search term was included due to being the most popular service at this time, to the point where it regularly appeared in titles and abstracts of various papers. The second part of the search string comprises the requirements for ethical AI found in the Ethics Guidelines for Trustworthy AI (henceforth AI HLEG) [8], with some of the principles omitted due to being too general and resulting in a vast number of unrelated results. The last part of the search string was added to further omit irrelevant results, such as various papers using LLMs for data analysis without discussing them. We searched for literature published after 2020 due to recent advances in GenAI. While the idea of GenAI predates this limitation, the AI ethics discussion on them generally does not, with most of these papers motivated by recent practical developments. We further discuss the limitations of the search protocol in Section 5.

Inclusion criteria	Exclusion criteria
In English	Not in English
Accessible to us	Behind a paywall and not self-archived
Peer-reviewed	Not peer-reviewed (or not confirmable)
Is a conference paper or a journal article	Is a book, book chapter, editorial, pre-print, letter, note, or keynote
Is focused on GenAI/ LLMs	Is not focused on GenAI/ LLMs
Is focused on discussing ethical issues or contains substantial discussion on them otherwise	Is not focused on ethical issues in GenAI/ LLMs, or does not contain substantial discussion on them

Table 1
Inclusion and exclusion criteria of the SLR.

3.2. SLR Process

The searches were conducted in June 2023. The initial searches resulted in a total of 1647 studies. After this, the results were iteratively evaluated and included or excluded in stages. This proceeded as follows: (1) inclusion/exclusion based on document type and removal of duplicate results *(1337 left, 310 excluded)* , (2) inclusion/exclusion based on title and abstract *(157 left, 1180 excluded)*, and (3) inclusion/exclusion based on full text *(116 included for review, 41 excluded)*. The inclusion and exclusion criteria are detailed in Table 1.

For clarity, we opted to only include papers explicitly discussing LLMs, even if general NLP literature could be considered relevant. As for "substantial discussion" on ethical issues, we decided to consider any paper that discusses ethical issues in the abstract to fulfill this criterion. Studies discussing relevant topics, such as cybersecurity, were excluded if they only focused on technical factors (e.g., attack vectors) without any ethical aspects being discussed in the abstract (or the full paper, in the third round).

3.3. Information Extraction

The following information was extracted from each of the included 116 final papers: (a) paper type (conference/journal/book and conceptual/empirical/lit. review), (b) types of ethical issues identified, (c) any mitigation measures proposed in relation to the ethical issues discussed, (d) database, (e) title, (f) authors, (g) DOI, (h) publication venue, (i) publication year.

Where possible, we utilized direct citations from the paper to extract the ethical issues and the mitigation proposals. In cases where the argumentation was more spread out, the key points were manually summarized and included in the spreadsheet brackets ([]). After this data had been extracted, the ethical issues and mitigation proposals were further analyzed by allocating them into categories. These categories were iteratively synthesized from the data so that (a) recurring issues could be best highlighted but so that (b) there would be as little overlap as

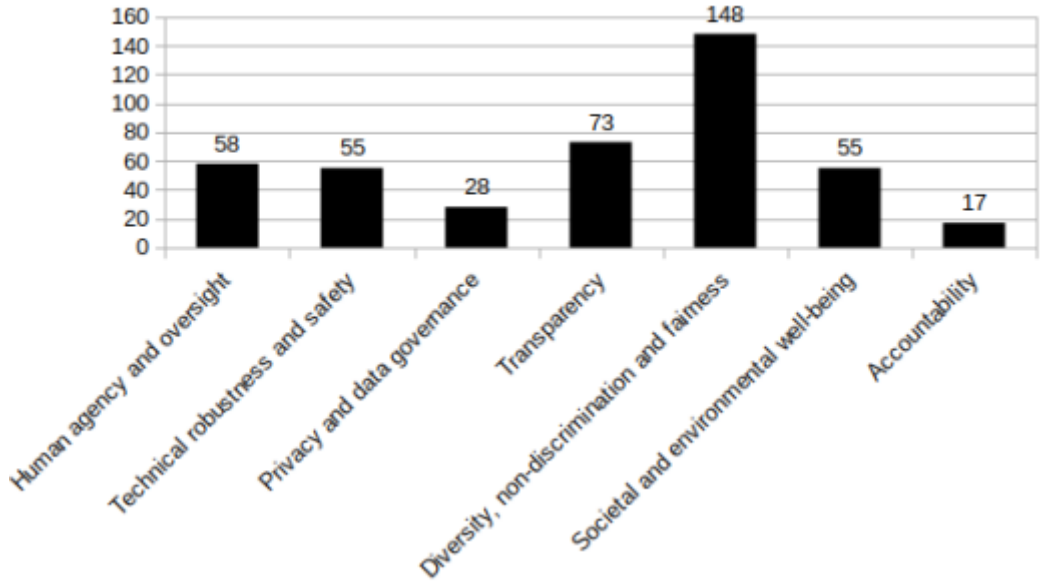

Figure 1: 434 individual ethical issues from literature mapped into the requirements of AI HLEG [8]

possible between the categories.

4. Results

Based on the SLR process, we identified 39 categories of ethical issues currently associated with LLMs in existing literature. These categories were synthesized based on 434 individual instances of ethical issues present in the 116 final papers. Figure 1 presents how these 434 issues are distributed between the seven requirements of the AI HLEG. As we report our results, we cite the 116 final papers using the following citation scheme: [Sn] where n is the number of the [S]tudy the way these 116 papers are presented in Appendix 1.

4.1. Human Agency and Oversight

The requirement of human agency and oversight posits that AI should support human autonomy and decision-making, while also allowing for human oversight of the AI. Human actors should be able to make informed decisions regarding the AI, and the AI should support humans in making more informed decisions. Human oversight, then, is intended to ensure that the AI does not undermine human autonomy or cause other undesirable effects. [8] The following categories of issues associated with human agency and oversight were identified:

- **Loss of learning or ease of cheating** [S1, S5, S7, S26, S28, S29, S30, S31, S33, S34, S42, S49, S59, S63, S70, S90, S98, S100, S103, S108, S110]
- **Fake news or misinformation** [S2, S3, S25, S30, S38, S42, S48, S67, S81, S93, S96, S97, S101, S103, S107, S110]

- **Echo chambers** [S2, S25, S101]
- **Self-acting AI** [S31, S64, S114]
- **Influence through suggestions** [S8, S38, S66, S114]
- **Manipulation** [S19, S23, S26, S38, S81, S93, S94, S97, S102, S104, S110]

Loss of learning or ease of cheating was one of the most commonly discussed ethical issues for LLMs. The discussion in the literature was primarily forward-facing, focusing on future issues. Only three papers [S59, S63, S110] were empirical in nature. To mitigate such issues, literature proposes (a) using teaching approaches that leverage LLMs and other AI tools while making it difficult to use them to cheat, such as in-person assignments, group work, and oral presentations, (b) providing better instructions to make students less inclined to use them, and (c) to better combat cheating with LLMs.

Fake news and other misinformation and disinformation. LLMs can be utilized to generate vast amounts content, and LLM-generated content is already not easily identifiable by humans and only becoming more difficult to discern. As a mitigation measure, literature recommends AI solutions for detecting AI-generated content, but this is acknowledged as a moving goalpost.

Echo chambers in LLMs manifest in that LLMs are not prone to challenging the views of their users and typically passively agree with the user, providing the content requested, barring any guardrails.

Self-acting AI refers to LLMs making decisions without sufficient human oversight, ranging from issues such as anthropomorphic LLMs expressing human-like emotions (e.g., "I'm sorry to hear that") to unobserved harmful actions.

Influence through suggestions. AI outputs are observed to affect user attitudes towards concepts [S38], and influence the entire creative process even if not accepted into the final text directly [S66]. Concerns have also been raised over AI affecting users' self-image [S8].

In terms of **manipulation**, literature discusses both purposeful manipulation and "nudging", as well as unintended manipulation resulting from biased training data. The black box nature of LLMs and the lack of alternative answers unless actively prompted are also considered to negatively impact human agency [S102]. As mitigation measures, literature suggests human-in-the-loop, clearly disclosing AI-generated content, and increased regulation.

4.2. Technical Robustness and Safety

This requirement, on a general level, posits that AI systems should be designed in a way that minimizes any harm they may cause, while preventing unacceptable harm. This includes cybersecurity issues, accuracy of the models being used, having fallback plans in case of issues, as well as reliability and reproducibility [8]. The following issues related to technical robustness and safety were identified in literature:

- **Inaccurate results** [S1, S6, S14, S25, S29, S34, S42, S52, S64, S82, S84, S93, S103, S107, S110, S112]
- **Dangerous content** [S76, S91, S97, S107]
- **Alignment** [S3, S13, S25]

- **Bias scoring solutions** [S10, S21, S65, S67, S79, S80, S81, S82, S85, S91, S95, S97, S105, S113, S116]
- **Data leakage or unintended memorization** [S21, S39, S40, S41, S44, S45, S55, S56, S68, S81, S86, S87, S88, S89, S97, S99, S115]

Inaccurate results of LLMs are a widely discussed issue in literature. Examples include attributing nationalities solely based on names [S112] and repeating common misconceptions found online [S107]. This is particular a problem in LLMs as humans are inclined to look at given content as a whole rather than zooming in on details, and LLMs are very capable of producing content that overall looks professional [S29]. Inaccuracies also tend to increase with model size [S107]. Better training data is generally suggested to mitigate these issues, whereas current models tend to use erratic data scoured from all over the Internet.

Dangerous content refers to potential for physical harm, such as dangerous advice. This includes both overtly unsafe (e.g., ”drink poison”) and indirectly unsafe content (”eat a Carolina Reaper”) [S76]. Children are considered most vulnerable in this regard, but adults can also fall victim to, e.g., bad legal or home appliance maintenance advice [S97].

Alignment refers to AI working in alignment with ethical values humans wish to impose on it, and is discussed in technical papers as well. In this SLR, we associated papers discussing unintentionally produced unethical content to relate to alignment. Such unintentional nonalignment is generally attributed issues with training data.

Bias scoring and solutions were discussed as a problem rather than a solution in many papers, as creating bias-free bias measures is a challenge in and of itself [S21, S67, S116]. Models may be trained to score well on a specific metric as a form of ”fairwashing” [S21]. In the context of LLMs, new issues may arise, as, e.g., filtering out certain words without context can eliminate reclaimed slur-words and removal of discourse of minorities. As mitigation measures, literature calls for standardized and validated bias measurement methods, as well as new types of metrics.

Data leakage or unintended memorization refers to LLMs memorizing their training data verbatim and occasionally outputting direct excerpts of it, which is considered a safety and robustness concern here as much as it is a data privacy concern. This is an issue especially for private data, which may also be purposefully extracted by malicious actors as opposed to simply being accidentally spilled. Several studies empirically demonstrate data leakage in existing LLMs [S40, S87, S88, S89]. Existing defensive techniques are criticized [S55, S88, S89], with the scale of LLMs working against them due to the equally large attack space it provides. Memorization is observed particularly for unique or rare data items, and it seems to increase with model capacity [S41]. As mitigation measures, literature proposes techniques for reducing memorization (e.g., [S41, S68, S87, S115]) and overall calls for improved techniques and designs to combat it.

4.3. Privacy and Data Governance

This requirement focuses on issues related to data and privacy in the development of AI systems. It encompasses both input data as well as system outputs. Current ML systems are particularly data intensive, and LLMs even moreso, emphasizing the importance of ethical issues related to data. The following issues related to this principle were identified from the literature:

- **Data gathered without consent** [S1, S2, S44, S81]
- **Privacy and data security** [S9, S11, S12, S13, S23, S26, S30, S47, S49, S54, S56, S57, S67, S68, S72, S89, S108, S109, S110]
- **Data management** [S39, S82, S87, S108, S111]

Data gathered without consent is a much-discussed issue in relation to LLMs due to current LLMs being typically trained on massive sets of data collected from all over the Internet. E.g., private data that is publicly available online may have ended up there as a result of a data breach rather than being uploaded with consent [S44], and utilizing it may cause harm [S81]. As the conceptually but practically challenging simple solution, literature recommends the use of data sets where data is gathered with consent [S44].

Privacy and data security is a widely discussed issue in LLMs, as it is with AI overall. In LLMs, aside from the training data posing various issues, the data collected from the users is also a potential issue [S9, S110]. Children in particular may be prone to sharing personal information with LLMs while interacting with them, although this is an issue for adults as well, especially with anthropomorphized AI [S97]. Various mitigation measures are discussed in literature, though each with their practical problems: automated de-identification [S47], pseudonymization and other de-identification methods [S11, S45, S47, S56, S97], and synthetic data [S47], as well as humans-in-the-loop [S50] and better sourcing of training data [S44, S97].

Data management poses challenges for LLM developers due to the massive amounts of training data used by current LLMs, making traditional data management methods non-effective [S111], posing both ethical and legal data management issues [S39]. As a mitigation measure, literature calls for increased governance [S39, S97] (although existing approaches are seen as insufficient), more auditable models [S56], and better documentation of training data sets [S71].

4.4. Transparency

Transparency is focused on making AI applications more understandable to various stakeholders, including their users, through more specific aspects such as traceability, explainability, and communication [8]. Traceability and explainability focus on system outputs and their understandability, although explainability also includes declarations of trade-offs made during the development of the system. Communication focuses on how the AI represents itself. For example, whether the system makes it clear the user is interacting with an AI. The following issue categories were identified in relation to transparency:

- **LLM as an author** [S1, S2, S24, S27, S52, S70, S94]
- **Academic integrity of source tracking** [S5, S6, S9, S20, S23, S24, S25, S31, S33, S34, S37, S42, S43, S44, S49, S62, S63, S70, S81, S90, S94, S96, S98, S100, S103, S108, S110]
- **Unfair decision making** [S12, S58, S69, S81, S109]
- **Lack of transparency** [S14, S26, S53, S65, S66, S78, S81, S95, S100, S101]
- **Copyright infringement** [S2, S19, S24, S26, S37, S39, S44, S50, S52, S72, S90, S98, S103, S104]
- **Unfactual training data** [S13, S23, S42, S49, S60, S67, S91, S97, S100, S110]

LLM as an author was discussed specifically in relation to academic writing, with the discussion stemming from LLMs actively being credited as authors in some recent publications. Literature argues that LLMs should not be credited as authors, as they are not accountable for their mistakes, and thus cannot be credited for their actions either. If authors insist such crediting, it might be more prudent to credit the developers of the tool, the party that owns it, or the individuals whose data it was trained on [S94].

A large number of papers raise concerns related to **academic integrity and source tracking** in the context of LLMs, as an increasing number of research papers (or manuscripts) is generated with the aid of LLMs or entirely by LLMs. Publishers have expressed concerns over their open access publications being used to train LLMs [S24]. As mitigation measures, literature again calls for authenticity checks for identifying AI-generated content [S9, S20, S24, S43, S44, S60, S62, S100], although false positives need to be seriously considered given the ramifications for researchers [S62]. Additionally, literature stresses the importance of guidelines and rules for using LLMs [S6, S60, S63, S98, S100, S110].

Unfair decision-making, here, refers to improper results that cannot be easily traced. Many of these issues are related to issues with training data, or the goals of the model training process. For example, when prioritizing who receives healthcare first, a value judgement is made [S12], intentionally or unintentionally. Allocational harms are more often considered unintentional harms [S58, S69, S81] and stem from training data that contains biases.

Lack of transparency is an issue in current LLMs that are predominantly black boxes, resulting in various issues. LLM users may create so-called ”algorithmic folk theories” where they attribute models with too much authority and fail to understand their limitations [S66]. This lack of transparency also increases reliance on the largest LLM developers [S101] and hinders research on LLMs. Moreover, current LLMs generally suffer from documentation debt [S81]. Better documentation and communication are proposed as mitigation measures.

Copyright infringement, here, refers to creative ownership rather than data ownership (discussed elsewhere). Current LLMs do not indicate whose creations were used to produce which outputs. LLMs do not credit creators and are themselves incapable of considering legal or ethical issues [S2]. No mitigation measures are proposed due to the technical limitations of current LLMs, and relevant legislation is still pending.

Unfactual training data refers to factually incorrect training data, which is particularly an issue for LLMs due to the vast amounts of (poorly curated) training data used to train them. This is also a challenge as new discoveries are constantly made in science, occasionally rendering obsolete information that was used to be considered a fact. Current LLMs are often trained on static data and not actively updated.

4.5. Diversity, Non-Discrimination, and Fairness

A trustworthy AI application needs to be fair and respect all people in an equal manner [8]. This includes attributes such as cultural background, beliefs, orientations, socioeconomic background, etc. This also includes *access* to AI applications, as well as taking different stakeholders into account and including them in the development process to what extent possible. The following ethical concerns connected to this requirement were identified in existing literature:

- **Biased training data or outputs** [S1, S2, S3, S5, S9, S10, S15, S16, S17, S18, S19, S20, S21,

S22, S23, S24, S26, S30, S35, S36, S37, S38, S42, S46, S48, S49, S51, S54, S57, S58, S60, S61, S64, S65, S66, S68, S69, S71, S72, S73, S74, S77, S81, S82, S83, S84, S85, S90, S91, S93, S94, S95, S96, S97, S98, S99, S100, S101, S103, S104, S105, S108, S109, S112, S113]

- **Discriminatory results** [S2, S5, S12, S17, S30, S46, S49, S58, S72, S73, S77, S80, S81, S84, S89, S93, S97, S98, S104, S109, S116]
- **Lack of global definition for bias or fairness** [S9, S21, S58, S60, S65, S84, S91, S93, S116]
- **Non-binary gender neglected** [S72, S73, S74, S75, S81, S84, S93, S109, S113, S116]
- **Toxic content** [S2, S3, S22, S30, S48, S54, S67, S74, S75, S81, S82, S89, S91, S101, S109, S111]
- **Promoting inequality** [S5, S7, S9, S12, S20, S23, S30, S33, S34, S49, S60, S61, S63, S67, S71, S77, S81, S82, S91, S93, S97, S101, S105, S108, S109, S110, S113]

Biased training data or outputs were the most common issue discussed in the literature reviewed for this SLR. The 10 most commonly discussed types of bias are biases related to gender, age, sexual orientation, physical appearance, disability, ethnicity, socioeconomic, religion, and culture, as well as cross-sectional bias [S16, S105]. More conceptual types of bias are discussed in [S51]: confidence, recency, majority label, and common token bias. As mitigation measures, literature proposes better training data management (both manual annotation and technological innovation) [S16, S35, S42, S46, S72, S105], pre-curated training corpora [S93, S94, S113], and making the NLP community more diverse [S16, S42].

Discriminatory results were another widely discussed issue in the reviewed literature, which focused on gender discrimination (e.g., [S46, S72]) and discrimination against marginalized demographic groups (e.g., [S58, S84]). This included a number of papers presenting empirical results [S46, S72, S73, S77, S80, S84, S109]. For mitigation measures, literature proposes multicultural development teams [S58], continuous evaluation [S77], and adversarial triggers and prompt engineering for testing purposes [S77, S83].

The **lack of a global definition for bias and fairness** presents issues for LLMs and AI systems at large. Currently, ethics reflected in the development of LLMs are those of western, white populations [S9, S116]. This also relates to the larger discussion on the limitations of algorithmic fairness. Few mitigation measures are proposed.

Non-binary gender is neglected often in LLMs due to the constraints of language, it being easier to study a model with a binary designation of gender. As bias specifically in LLMs is a new practical issue overall, few solutions are currently discussed in literature to this more specific bias issue

Toxic content illustrates the problems of extracting training data from the Internet with little curation. Data used to train LLMs often includes, e.g., content from Reddit that is from banned or quarantined subreddits [S54]. Many of the reviewed studies discussing this issue also demonstrate it in practice. As for mitigation measures, though toxicity scoring tools similar to bias scoring ones exist, they have similar problems [S22], and using smaller, curated sets of data hinders performance.

It is argued that LLMs **promote inequality** in different ways. E.g., by offering better performance against subscription fees [S7], as a result of those with power producing most content used to train them [S113], and by exhibiting poorer performance in languages other

than English [S67, S91, S113]. Mitigation measures are shared between other bias and fairness issues.

4.6. Societal and Environmental Well-being

According to AI HLEG [8], AI systems should be designed in a sustainable manner, referring to societal and environmental well-being. This includes the environmental cost of the entire system life-cycle and supply chain, as well as socially conscious design of AI. The guidelines acknowledge, in relation to this requirement, that AI systems can be incredibly beneficial to society, but also similarly detrimental if developed and implemented poorly. The following issues were identified in relation to this requirement:

- **Loss of social skills** [S26, S30]
- **Reduced value of education** [S29, S30, S43, S60, S108, S110]
- **Paper or credential generation** [S20, S24, S25, S35, S42, S60, S62, S81]
- **Purposeful toxic or immoral content** [S50, S74, S92, S97, S99, S103, S104, S107, S109]
- **Job loss or class divide** [S2, S3, S9, S12, S13, S30, S37, S49, S97, S104]
- **Training data pruning** [S2, S52]
- **Environmental impacts** [S4, S17, S32, S42, S52, S71, S72, S81, S97, S101, S106]
- **Fairwashing** [S84, S106]
- **Replacement of traditional learning** [S1, S2, S49, S53, S57]

Loss of social skills may occur in the future if increased use of AI results in decreased development of social skills. As this is a hypothetical future issue, discussed to raise awareness of its potential future impacts, no mitigation methods are presented to tackle it currently.

LLMs may **reduce the value of education** if students become too reliant on LLM use and feel that learning and memorizing some things themselves is not worth the effort. Using LLMs to pass tests may also result in students being ill-prepared for future careers [S29]. Literature suggests teaching AI literacy as one way of addressing such issues [S108].

Paper or credential generation. LLMs are capable of producing superficially viable research content, which can lead to an inflated number of papers being produced. E.g., some individuals may use LLMs to produce massive amounts of self-referential studies with little scientific novelty [S20, S24]. As mitigation measures, literature proposes guidelines and demanding disclosure of LLM use [S35] and technologies for identifying LLM-generated content.

Aside from unintentional unethical content, LLMs can be used to produce **purposeful toxic or immoral content**. E.g., to generate malware [S92], or for social engineering attacks [S92, S99]. Adversarial triggers can be used to generate unexpected outputs seemingly unprompted [S99]. Mitigation measures proposed are similar to those proposed for toxic and biased content overall, in addition to stock responses for inappropriate prompts [S91].

Like many new technologies before, LLMs may result in **job loss or class divide**. In particular, LLMs are seen to concern groups of professionals that have historically not been under such threats before, namely white-collar workers [S3] and creative professionals [S37, S97].

Training data pruning for LLMs is currently often conducted by human actors from developing countries [S2]. These individuals are subjected to large amounts of unwanted and inappropriate content in the process, which can be extreme and emotionally damaging.

Environmental impacts of LLMs include (1) direct impacts from energy use, (2) secondary impacts from emissions, (3) impacts resulting from the system influencing human behavior, and (4) resources needed for hardware. Few operators publish figures on environmental impacts [S106], and as the scale of LLMs continues to grow, they require more and more computational power [S106]. Notably, increasing model fairness tends to decrease sustainability as well [S17]. The following mitigation measures are proposed: increased (and standardized) reporting [S17, S71, S81], more efficient solutions [S17, S32], running systems in more carbon friendly regions [S81].

Fairwashing (cf. greenwashing) is discussed in literature in relation to hypothetical certificates that may be granted to systems and their developers without reliable means of verifying compliance to the methods the certificate aims to measure, as well as overall marketing of LLMs and AI.

Replacement of traditional learning, while closely related to reduced value of education, which in this SLR refers to similar issues from the perspective of students, is more focused on the entire education system. Thus, it includes other actors than learners as well. E.g., for educators, the use of LLMs in and of itself warrants ethical consideration when it comes to tasks such as (automated) grading [S53, S57]. Mitigation measures include AI literacy education from an early age [S2].

4.7. Accountability

The seventh and final requirement of the AI HLEG [8] is accountability. This requirement posits that AI systems must be responsible, that it must be possible to properly audit their use and content, and that developers need to minimize and report negative impacts, address trade-offs, and have processes for redress if and when issues do arise. This requirement spans the entire system lifecycle as well. The following issues related to this requirement were identified in the SLR:

- **Corporate influence** [S2, S10, S13, S25, S96, S101, S116]
- **Cost of privacy** [S45, S56, S68, S89]
- **Cost of AI monitoring** [S70, S83, S93]
- **Ambiguity of accountability** [S13, S26, S28]

Corporate influence is seen as an issue in relation to LLMs as large, multinational corporations dominate the LLM development and consumer market, and do so largely unregulated at present. These corporations safeguard their position by obscuring the training data and program code [S13] (and own the large cloud computing facilities for training), and control research by publications and recruitment [S101, S116]. Smaller practitioners struggle to enter the market. As mitigation measures, literature calls for more regulation [S2, S10, S23, S25, S96] and better funding for independent researchers [S101, S116].

Cost of privacy is an issue in LLMs in that it can often be dialed down to improve efficiency or profitability. The massive amounts of training data make comprehensive human overview of the training material for LLMs practically impossible [S56, S89]. Due to it being so expensive to improve model privacy, forgoing it may be tempting, especially if the model is for internal

organization use [S45], but organizations should remember that such models should not be shared or used in situations where this may be more likely to cause issues.

As for the **cost of AI monitoring**, continuously monitoring, testing, and amending LLM performance after deployment can be highly resource-intensive. In particular, re-training models is especially expensive in the context of LLMs, which may be required to address ethical issues when there is a will to do so.

Ambiguity of accountability presents various issues for LLMs. The fact that LLMs do not clearly emulate any parts of their training data also has implications for copyright and academic referencing, and may also be relevant for toxic and otherwise inappropriate content [S13]. E.g., if an LLM makes a social media post that breaks the terms of service or is illegal, pinning the responsibility on a human actor may be difficult, if it is needed. As a solution, literature suggests enforcing accountability through regulation.

5. Threats to Validity

The SLR presented in this paper has its limitations. The SLR protocol itself has the following limitations. First, we have limited the search to literature published after 2020. This was done due to much of the LLM-specific discussion being related to very recent advances in the field (and, e.g., the introduction of ChatGPT), so as to more reliably and easily exclude irrelevant results. However, this will have also excluded any less recent papers with the foresight to discuss these issues prior to 2021. Second, we have excluded some of the requirements (robustness, safety, governance, societal, environmental, and accountability) of the AI HLEG [8] from our search keywords due to the high number of irrelevant results they produced. Doing so may have resulted in excluding relevant literature, but we have nonetheless identified various papers and issues related to these concepts as well in this SLR. Third, the principle-based approach in and of itself poses some potential limitations. There is no universally agreed-upon set of principles for AI ethics [3]. In using the AI HLEG requirements for trustworthy AI [8] as keywords, any literature focusing on different principles, or no principles at all, may have been excluded as a result. Finally, we have included ChatGPT as a part of the search string due to its particular relevance at the time, but other services such as GitHub Copilot could have also been included, in addition to alternatives such as Llama (or Llama 2).

Additionally, it is also prudent to acknowledge potential limitations related to the data extraction and analysis. The analysis process was carried out by the first author, although the second and third author discussed and planned the process with the first author. This leaves room for subjective interpretation on part of the first author in terms of categorizing the identified ethical issues. However, we argue that the use of an existing ethical framework (AI HLEG [8]) has served to leave less room for subjective interpretation, improving the quality of the analysis.

6. Discussion and Conclusions

In this paper, we conducted an SLR of literature on ethical issues and risks associated with LLMs and generative AI. Based on 116 papers, we identified 434 individual ethical issues, based

on which we proposed 39 categories of ethical issues related to LLMs. We then mapped these 39 issue categories to the seven requirements for ethical AI found in the Ethics Guidelines for Trustworthy AI [8]. Based this analysis, we propose several practical and theoretical implications. Our findings have implications primarily for researchers interested in AI ethics, although they can also inform practitioners about potential issues they might want to be aware of, depending on their project or system context.

First, we conducted our SLR using AI HLEG [8] as a framework. This framework consists of seven requirements (human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental well-being, and accountability). The issues and concerns we identified in this SLR could all be reasonably mapped to these requirements during the analysis. These requirements correspond with the most common AI ethics principles [3]. Thus, we argue that, on a conceptual level, the ethical issues identified in LLMs are similar to the ones previously acknowledged in literature.

However, when looking at each issue in more detail, some still present new practical challenges, while others are indeed familiar ones. As an example, the trade-off between accuracy and explainability (or interpretability) is well-documented in existing literature (e.g., [11]), while literature on LLMs discusses trade-offs between accuracy and privacy [S11, S45, S47, S56, S91, S97] and trade-offs between fairness and sustainability [S17]. Similarly, the capability of LLMs to generate convincing content that looks human-made presents various practical issues, as we have extensively discussed in this SLR. On the other hand, job loss resulting from automation, for example, is an issue that has been discussed in relation to various technologies over the decades and centuries, including computers overall.

Secondly, based on the literature, it seems that there is more emphasis on *user responsibility* as well, in addition to developer or organization responsibility that has typically been the focus in AI ethics (e.g., as discussed in AI HLEG [8] and in the various guidelines aimed at organizations developing AI). Indeed, it may also be worthwhile it to consider the role of the user in producing inappropriate or biased outputs. Literature discusses the use of LLMs purposely for unethical purposes [S50, S74, S92, S97, S99, S103, S104, S107, S109], as well as the importance of teaching AI literacy in the future [S2, S108]. For example, if developers take measures to block policy violating prompts [S48] and users then actively look for ways to circumvent them on purpose, is there a point where the responsibility for such misuse should fall on the user instead? Literature agrees that such systems will inevitably become more popular, and that educating their users on good practices for their use is vital.

Thirdly, our principle-based SLR approach compliments study of Weidinger et al. [6]. Based on workshops with professionals, they identify six risk categories for language models (”I. Discrimination, Hate speech and Exclusion, II. Information Hazards, III. Misinformation Harms, IV. Malicious Uses, V. Human-Computer Interaction Harms, and VI. Environmental and Socioeconomic harms”) and then differentiate between already observed risks and hypothetical, future risks for each category. Together, these papers provide an overview of the ethical issues and risks associated with LLMs currently.

Finally, as for practical implications, the issues identified in this SLR may inform any interested practitioner in potential issues and risks they may want to be aware of depending on their project or system context. In addition to identifying potential ethical issues in this SLR, we

have also listed the solution suggestions for each ethical issue found in literature, which may help practitioners tackle any issues they may consider relevant for their system context.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT for Sections 3.1 and 3.2 (SLR protocol): Drafting content. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] Reuters, ChatGPT sets record for fastest-growing user base - analyst note, 2023]. URL: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

[2] C. K. Lo, What is the impact of ChatGPT on education? A rapid review of the literature, Education Sciences 13 (2023) 410.

[3] A. Jobin, M. Ienca, E. Vayena, The global landscape of ai ethics guidelines, Nature Machine Intelligence 1 (2019) 389–399.

[4] T. Hagendorff, The ethics of AI ethics: An evaluation of guidelines, Minds and Machines 30 (2020) 99–120.

[5] 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 (2020) 2141–2168.

[6] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh, et al., Taxonomy of risks posed by language models, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 214–229.

[7] V. Vakkuri, P. Abrahamsson, The key concepts of ethics of artificial intelligence, in: 2018 IEEE international conference on engineering, technology and innovation (ICE/ITMC), IEEE, 2018, pp. 1–6.

[8] Ethics guidelines for trustworthy ai, 2019. URL: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[9] 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.

[10] J.-L. Selter, K. Wagner, H. Schramm-Klein, Ethics and morality in AI-a systematic literature review and future research, ECIS 2022 Research Papers. 60. (2022).

[11] M. Ananny, K. Crawford, Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability, new media & society 20 (2018) 973–989.

Appendix 1

This appendix contains the full list of studies included in this SLR (Table 2), clearly separated from the references used outside Section 4 that were not a part of the SLR itself. As mentioned in Section 3, these papers were each given an ID (S1-S116) that was used to reference them in the paper (e.g., "[S52]").

ID	Full reference (APA)
S1	Anderson, S. S. (2023). "Places to stand": Multiple metaphors for framing ChatGPT's corpus. *Computers and Composition*, 68, 102778.
S2	Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... & Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, 102642.
S3	Taecharungroj, V. (2023). "What can ChatGPT do?" Analyzing early reactions to the innovative AI chatbot on Twitter. Big Data and Cognitive Computing, 7(1), 35.
S4	Lakim, I., Almazrouei, E., Abualhaol, I., Debbah, M., & Launay, J. (2022, May). A holistic assessment of the carbon footprint of noor, a very large Arabic language model. In*Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models* (pp. 84-94).
S5	Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*, 1-15.
S6	Perkins, M. (2023). Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. *Journal of university teaching & learning practice, 20(2), 07.*
S7	Yeo, M. A. (2023). Academic integrity in the age of artificial intelligence (AI) authoring apps. *Tesol Journal*, 14(3), e716.
S8	Poddar, R., Sinha, R., Naaman, M., & Jakesch, M. (2023, April). AI Writing Assistants Influence Topic Choice in Self-Presentation. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-6).
S9	Luo, W., He, H., Liu, J., Berson, I. R., Berson, M. J., Zhou, Y., & Li, H. (2024). Aladdin's Genie or Pandora's Box for early childhood education? Experts chat on the roles, challenges, and developments of ChatGPT. *Early Education and Development*, 35(1), 96-113.
S10	Steinborn, V., Dufter, P., Jabbar, H., & Schütze, H. (2022, July). An information-theoretic approach and dataset for probing gender stereotypes in multilingual masked language models. In *Findings of the Association for Computational Linguistics: NAACL 2022* (pp. 921-932).
S11	Vakili, T., & Dalianis, H. (2021). Are clinical BERT models privacy preserving? The difficulty of extracting patient-condition associations. In AAAI 2021 Fall Symposium on Human Partnership with Medical AI: Design, Operationalization, and Ethics (AAAI-HUMAN 2021), Virtual Event, November 4-6, 2021.

S12	Haluza, D., & Jungwirth, D. (2023). Artificial intelligence and ten societal megatrends: An exploratory study using GPT-3. *Systems*, 11(3), 120.
S13	Kolides, A., Nawaz, A., Rathor, A., Beeman, D., Hashmi, M., Fatima, S., ... & Jararweh, Y. (2023). Artificial intelligence foundation and pre-trained models: Fundamentals, applications, opportunities, and social impacts. *Simulation Modelling Practice and Theory*, 126, 102754.
S14	Pearce, H., Ahmad, B., Tan, B., Dolan-Gavitt, B., & Karri, R. (2022, May). Asleep at the keyboard? assessing the security of github copilot's code contributions. In *2022 IEEE Symposium on Security and Privacy (SP)* (pp. 754-768). IEEE.
S15	Kirk, H. R., Jun, Y., Volpin, F., Iqbal, H., Benussi, E., Dreyer, F., ... & Asano, Y. (2021). Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in neural information processing systems*, 34, 2611-2624.
S16	Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2), 1-21.
S17	Hessenthaler, M., Strubell, E., Hovy, D., & Lauscher, A. (2022). Bridging fairness and environmental sustainability in natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7817–7836, Abu Dhabi, United Arab Emirates.* Association for Computational Linguistics.
S18	Shihadeh, J., Ackerman, M., Troske, A., Lawson, N., & Gonzalez, E. (2022, September). Brilliance bias in GPT-3. In *2022 IEEE Global Humanitarian Technology Conference (GHTC)* (pp. 62-69). IEEE.
S19	Kreps, S., & Jakesch, M. (2023). Can AI communication tools increase legislative responsiveness and trust in democratic institutions?. *Government Information Quarterly*, 40(3), 101829.
S20	Salvagno, M., Taccone, F. S., & Gerli, A. G. (2023). Can artificial intelligence help for scientific writing?. *Critical care*, 27(1), 75.
S21	Balkir, E., Kiritchenko, S., Nejadgholi, I., & Fraser, K. C. (2022). Challenges in applying explainability methods to improve the fairness of NLP models. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*
S22	Welbl, J., Glaese, A., Uesato, J., Dathathri, S., Mellor, J., Hendricks, L. A., ... & Huang, P. S. (2021). Challenges in detoxifying language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2447–2469, Punta Cana, Dominican Republic. Association for Computational Linguistics.
S23	Kooli, C. (2023). Chatbots in education and research: A critical examination of ethical implications and solutions. *Sustainability*, 15(7), 5614.
S24	Lund, B. D., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., & Wang, Z. (2023). ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, 74(5), 570-581.

S25	De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., & Rizzo, C. (2023). ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Frontiers in Public Health*, 11, 1166120.
S26	Sallam, M., Salim, N. A., Barakat, M., & Ala'a, B. (2023). ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations. *Narra J*, 3(1).
S27	Dowling, M., & Lucey, B. (2023). ChatGPT for (finance) research: The Bananarama conjecture. *Finance Research Letters*, 53, 103662.
S28	Rahman, M. M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences*, 13(9), 5783.
S29	Mrabet, J., & Studholme, R. (2023, March). ChatGPT: A friend or a foe?. In *2023 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)* (pp. 269-274). IEEE.
S30	Bahrini, A., Khamoshifar, M., Abbasimehr, H., Riggs, R. J., Esmaeili, M., Majdabadkohne, R. M., & Pasehvar, M. (2023, April). ChatGPT: Applications, opportunities, and threats. In *2023 Systems and Information Engineering Design Symposium (SIEDS)* (pp. 274-279). IEEE.
S31	Abdullah, M., Madain, A., & Jararweh, Y. (2022, November). ChatGPT: Fundamentals, applications and social impacts. In *2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 1-8). Ieee.
S32	Gill, S. S., & Kaur, R. (2023). ChatGPT: Vision and challenges. *Internet of Things and Cyber-Physical Systems*, 3, 262-271.
S33	Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in education and teaching international*, 61(2), 228-239.
S34	Choi, E. P. H., Lee, J. J., Ho, M. H., Kwok, J. Y. Y., & Lok, K. Y. W. (2023). Chatting or cheating? The impacts of ChatGPT and other artificial intelligence language models on nurse education. *Nurse Education Today*.
S35	Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ digital medicine*, 6(1), 75.
S36	da Silva, D. A., Louro, H. D. B., Goncalves, G. S., Marques, J. C., Dias, L. A. V., da Cunha, A. M., & Tasinaffo, P. M. (2021). Could a conversational ai identify offensive language?. *Information*, 12(10), 418.
S37	Mirowski, P., Mathewson, K. W., Pittman, J., & Evans, R. (2023, April). Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-34).
S38	Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., & Naaman, M. (2023, April). Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI conference on human factors in computing systems* (pp. 1-15).

S39	Jernite, Y., Nguyen, H., Biderman, S., Rogers, A., Masoud, M., Danchev, V., ... & Mitchell, M. (2022, June). Data governance in the age of large-scale data-driven language technology. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2206-2222).
S40	Panchendrarajan, R., & Bhoi, S. (2021). Dataset reconstruction attack against language models. In *Proceedings of AIofAI'21: 1st Workshop on Adverse Impacts and Collateral Effects of Artificial Intelligence Technologies, Montreal, CA*. CEUR Workshop.
S41	Huang, W. R., Chien, S., Thakkar, O., & Mathews, R. (2022). Detecting unintended memorization in language-model-fused ASR. In *Proceedings of Interspeech 2022*, 2808-2812.
S42	Kansteiner, W. (2022). Digital doping for historians: can history, memory, and historical theory be rendered artificially intelligent?. *History and Theory*, 61(4), 119-133.
S43	Alamleh, H., AlQahtani, A. A. S., & ElSaid, A. (2023, April). Distinguishing human-written and ChatGPT-generated text using machine learning. In *2023 Systems and Information Engineering Design Symposium (SIEDS)* (pp. 154-158). IEEE.
S44	Lee, J., Le, T., Chen, J., & Lee, D. (2023, April). Do language models plagiarize?. In *Proceedings of the ACM Web Conference 2023* (pp. 3637-3647).
S45	Lehman, E., Jain, S., Pichotta, K., Goldberg, Y., & Wallace, B. C. (2021). Does BERT pretrained on clinical notes reveal sensitive data?. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
S46	Limisiewicz, T., & Mareček, D. (2022). Don't Forget About Pronouns: Removing Gender Bias in Language Models Without Losing Factual Gender Information. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, Seattle, Washington. Association for Computational Linguistics.
S47	Vakili, T., Lamproudis, A., Henriksson, A., & Dalianis, H. (2022, June). Downstream task performance of bert models pre-trained using automatically de-identified clinical data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 4245-4252).
S48	Gradonm, K. T. (2023). Electric sheep on the pastures of disinformation and targeted phishing campaigns: The security implications of chatgpt. *IEEE Security & Privacy*, 21(3), 58-61.
S49	Qadir, J. (2023, May). Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. In *2023 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1-9). IEEE.
S50	Mori, Y., & Miyake, Y. (2022, December). Ethical Issues in Automatic Dialogue Generation for Non-Player Characters in Digital Games. In *2022 IEEE International Conference on Big Data (Big Data)* (pp. 5132-5139). IEEE.
S51	Hämäläinen, P., Tavast, M., & Kunnari, A. (2023, April). Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-19).

S52	Cooper, G. (2023). Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology*, 32(3), 444-452.
S53	Poulton, A., & Eliens, S. (2021, September). Explaining transformer-based models for automatic short answer grading. In Proceedings of the 5th International Conference on Digital Technology in Education (pp. 110-116).
S54	Nguyen, H., Malik, A., & Zink, M. (2022). Exploring Realtime Conversational Virtual Characters. *SMPTE Motion Imaging Journal*, 131(3), 25-34.
S55	He, X., Chen, C., Lyu, L., & Xu, Q. (2022). Extracted BERT Model Leaks More Information than You Think!. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 1530–1537), Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
S56	Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Raffel, C. (2021). Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)* (pp. 2633-2650).
S57	Kumar, R. (2023). Faculty members' use of artificial intelligence to grade student papers: a case of implications. *International Journal for Educational Integrity*, 19(1), 9.
S58	Ramesh, K., Sitaram, S., & Choudhury, M. (2023). Fairness in language models beyond English: Gaps and challenge. In *Findings of the Association for Computational Linguistics: EACL 2023* (pp. 2106–2119), Dubrovnik, Croatia. Association for Computational Linguistics.
S59	Biderman, S., & Raff, E. (2022, October). Fooling MOSS detection with pretrained language models. In *Proceedings of the 31st ACM international conference on information & knowledge management* (pp. 2933-2943).
S60	Dergaa, I., Chamari, K., Zmijewski, P., & Saad, H. B. (2023). From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. *Biology of sport*, 40(2), 615-622.
S61	Li, Y., Zhang, G., Yang, B., Lin, C., Wang, S., Ragni, A., & Fu, J. (2022). Herb: Measuring hierarchical regional bias in pre-trained language models. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022* (pp. 334–346). Association for Computational Linguistics.
S62	Wahle, J. P., Ruas, T., Kirstein, F., & Gipp, B. (2022). How large language models are transforming machine-paraphrased plagiarism. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 952–963). Association for Computational Linguistics.
S63	Yan, D. (2023). Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Education and Information Technologies*, 28(11), 13943-13967.
S64	Kasirzadeh, A., & Gabriel, I. (2023). In conversation with artificial intelligence: aligning language models with human values. *Philosophy & Technology*, 36(2), 27.

S65	Pikuliak, M., Beňová, I., & Bachratý, V. (2023). In-depth look at word filling societal bias measures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 3648–3665), Dubrovnik, Croatia. Association for Computational Linguistics.
S66	Bhat, A., Agashe, S., Oberoi, P., Mohile, N., Jangir, R., & Joshi, A. (2023, March). Interacting with next-phrase suggestions: How suggestion systems aid and influence the cognitive processes of writing. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (pp. 436-452).
S67	Kumar, S., Balachandran, V., Njoo, L., Anastasopoulos, A., & Tsvetkov, Y. (2022). Language generation models can cause harm: So what can we do about it? an actionable survey. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 3299–3321), Dubrovnik, Croatia. Association for Computational Linguistics.
S68	Li, X., Tramer, F., Liang, P., & Hashimoto, T. (2022). Large language models can be strong differentially private learners. In *Proceedings of the 10th International Conference on Learning Representations (ICLR 2022).*
S69	Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., & Kersting, K. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3), 258-268.
S70	Crawford, J., Cowling, M., & Allen, K. A. (2023). Leadership is needed for ethical ChatGPT: Character, assessment, and learning using artificial intelligence (AI). *Journal of University Teaching & Learning Practice*, 20(3), 02.
S71	Alshahrani, S., Wali, E., & Matthews, J. (2022, December). Learning From Arabic Corpora But Not Always From Arabic Speakers: A Case Study of the Arabic Wikipedia Editions. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)* (pp. 361-371).
S72	Borchers, C., Gala, D. S., Gilburt, B., Oravkin, E., Bounsi, W., Asano, Y. M., & Kirk, H. R. (2022). Looking for a handsome carpenter! debiasing GPT-3 job advertisements. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)* (pp. 212–224), Seattle, Washington. Association for Computational Linguistics.
S73	Kraft, A., Zorn, H. P., Fecht, P., Simon, J., Biemann, C., & Usbeck, R. (2022). Measuring gender bias in german language generation. In *Proceedings of The Informatik 2022 Workshop "Trustworthy AI in Science and Society"* (pp. 1257-1274), Hamburg. 26.-30. September 2022.
S74	Touileb, S., & Nozza, D. (2022). Measuring harmful representations in Scandinavian language models. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)* (pp. 118–125), Abu Dhabi, UAE. Association for Computational Linguistics.
S75	Nozza, D., Bianchi, F., Lauscher, A., & Hovy, D. (2022). Measuring harmful sentence completion in language models for LGBTQIA+ individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion.* Association for Computational Linguistics.

S76	Mei, A., Kabir, A., Levy, S., Subbiah, M., Allaway, E., Judge, J., ... & Wang, W. Y. (2022). Mitigating covertly unsafe text within natural language systems. In *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 2914–2926), Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
S77	Venkit, P. N., Gautam, S., Panchanadikar, R., Huang, T. H. K., & Wilson, S. (2023). Nationality bias in text generation. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 116–122, Dubrovnik, Croatia. Association for Computational Linguistics.
S78	Longoni, C., Fradkin, A., Cian, L., & Pennycook, G. (2022, June). News from generative artificial intelligence is believed less. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 97-106).
S79	Dev, S., Sheng, E., Zhao, J., Amstutz, A., Sun, J., Hou, Y., ... & Chang, K. W. (2021). On measures of biases and harms in NLP. *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022* (pp. 246–267). Association for Computational Linguistics.
S80	Akyürek, A. F., Paik, S., Kocyigit, M. Y., Akbiyik, S., Runyun, Ş. L., & Wijaya, D. (2022). On measuring social biases in prompt-based multi-task learning. In *Findings of the Association for Computational Linguistics: NAACL 2022* (pp. 551–564), Seattle, United States. Association for Computational Linguistics.
S81	Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623).
S82	Vashishtha, A., Prasad, S. S., Bajaj, P., Chaudhary, V., Cook, K., Dandapat, S., ... & Choudhury, M. (2023, May). Performance and Risk Trade-offs for Multi-word Text Prediction at Scale. In *Findings of the Association for Computational Linguistics: EACL 2023* (pp. 2226-2242).
S83	Abid, A., Farooqi, M., & Zou, J. (2021, July). Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 298-306).
S84	Qian, R., Ross, C., Fernandes, J., Smith, E., Kiela, D., & Williams, A. (2022). Perturbation augmentation for fairer NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 9496–9521), Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
S85	Nozza, D., Bianchi, F., & Hovy, D. (2022). Pipelines for social bias testing of large language models. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*. Association for Computational Linguistics.
S86	Elmahdy, A., Inan, H. A., & Sim, R. (2022). Privacy leakage in text classification: A data extraction approach. In *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing* (pp. 13–20), Seattle, United States. Association for Computational Linguistics.

S87	Zhao, X., Li, L., & Wang, Y. X. (2022). Provably confidential language modelling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 943–955), Seattle, United States. Association for Computational Linguistics.
S88	Mireshghallah, F., Goyal, K., Uniyal, A., Berg-Kirkpatrick, T., & Shokri, R. (2022). Quantifying privacy risks of masked language models using membership inference attacks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 8332–8347), Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
S89	Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., ... & Irving, G. (2022). Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 3419–3448), Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
S90	Ibrahim, H., Asim, R., Zaffar, F., Rahwan, T., & Zaki, Y. (2023). Rethinking homework in the age of artificial intelligence. *IEEE Intelligent Systems*, 38(2), 24-27.
S91	Dinan, E., Abercrombie, G., Bergman, S. A., Spruit, S., Hovy, D., Boureau, Y. L., & Rieser, V. (2022). SafetyKit: First aid for measuring safety in open-domain conversational systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics.
S92	Grbic, D. V., & Dujlovic, I. (2023, March). Social engineering with ChatGPT. In *2023 22nd International Symposium INFOTEH-JAHORINA (INFOTEH)* (pp. 1-5). IEEE.
S93	Sheng, E., Chang, K. W., Natarajan, P., & Peng, N. (2021). Societal biases in language generation: Progress and challenges. 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)* (pp. 4275–4293), Online. Association for Computational Linguistics.
S94	Gero, K. I., Liu, V., & Chilton, L. (2022, June). Sparks: Inspiration for science writing using language models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (pp. 1002-1019).
S95	Sayenju, S., Aygun, R., Franks, B., Johnston, S., Lee, G., & Modgil, G. (2022, December). Stereotype and Categorical Bias Evaluation via Differential Cosine Bias Measure. In *2022 IEEE International Conference on Big Data (Big Data)* (pp. 5082-5089). IEEE.
S96	Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & De Choudhury, M. (2023, April). Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1-20).
S97	Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P. S., Mellor, J., ... & Gabriel, I. (2022, June). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 214-229).
S98	Sun, G. H., & Hoelscher, S. H. (2023). The ChatGPT storm and what faculty can do. *Nurse Educator*, 48(3), 119-124.

S99	Heidenreich, H. S., & Williams, J. R. (2021, July). The earth is flat and the sun is not a star: The susceptibility of gpt-2 to universal adversarial triggers. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 566-573).
S100	Dalalah, D., & Dalalah, O. M. (2023). The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT. *The International Journal of Management Education*, 21(2), 100822.
S101	Luitse, D., & Denkena, W. (2021). The great transformer: Examining the role of large language models in the political economy of AI. *Big Data & Society*, 8(2), 20539517211047734.
S102	Rosenberg, L. (2023, March). The metaverse and conversational AI as a threat vector for targeted influence. In *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 0504-0510). IEEE.
S103	Murugesan, S., & Cherukuri, A. K. (2023). The rise of generative artificial intelligence and its impact on education: the promises and perils. Computer, 56(5), 116-121.
S104	Weisz, J. D., Muller, M., He, J., & Houde, S. (2023). Toward general design principles for generative AI applications. In *Joint Proceedings of the ACM IUI Workshops 2023*, March 2023, Sydney, Australia. CEUR-WS.
S105	Anoop, K., Gangan, M. P., Deepak, P., & Lajish, V. L. (2022). Towards an enhanced understanding of bias in pre-trained neural language models: A survey with special emphasis on affective bias. In *Responsible Data Science: Select Proceedings of ICDSE 2021* (pp. 13-45). Singapore: Springer Nature Singapore.
S106	Hershcovich, D., Webersinke, N., Kraus, M., Bingler, J. A., & Leippold, M. (2022). Towards climate awareness in NLP research. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 2480–2494), Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
S107	Lin, S., Hilton, J., & Evans, O. (2021). Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3214–3252), Dublin, Ireland. Association for Computational Linguistics.
S108	Su, J., & Yang, W. (2023). Unlocking the power of ChatGPT: A framework for applying generative AI in education. *ECNU Review of Education*, 6(3), 355-366.
S109	Steed, R., Panda, S., Kobren, A., & Wick, M. (2022, May). Upstream mitigation is not all you need: Testing the bias transfer hypothesis in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3524-3542).
S110	Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. Smart Learning Environments, 10(1), 15.
S111	Luccioni, A., & Viviano, J. (2021, August). What's in the box? an analysis of undesirable content in the Common Crawl corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 182-189).

S112	Ladhak, F., Durmus, E., Suzgun, M., Zhang, T., Jurafsky, D., McKeown, K., & Hashimoto, T. B. (2023, May). When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 3206-3219).
S113	Gururangan, S., Card, D., Dreier, S. K., Gade, E. K., Wang, L. Z., Wang, Z., ... & Smith, N. A. (2022). Whose language counts as high quality? measuring language ideologies in text data selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 2562–2580), Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
S114	Liu, Y., Mittal, A., Yang, D., & Bruckman, A. (2022, April). Will AI console me when I lose my pet? Understanding perceptions of AI-mediated email writing. In *Proceedings of the 2022 CHI conference on human factors in computing systems* (pp. 1-13).
S115	Li, H., Song, Y., & Fan, L. (2022). You don't know my favorite color: Preventing dialogue representations from revealing speakers' private personas. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 5858–5870), Seattle, United States. Association for Computational Linguistics.
S116	Talat, Z., Névéol, A., Biderman, S., Clinciu, M., Dey, M., Longpre, S., ... & Van Der Wal, O. (2022, May). You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode#5–Workshop on Challenges & Perspectives in Creating Large Language Models* (pp. 26-41).

Table 2: Full list of papers included in the SLR.