Master's thesis

Master's Programme in Computer Science

Ethical challenges of large language models - a systematic literature review

Atte Laakso

September 26, 2023

Faculty of Science

University of Helsinki

Contact information

P. O. Box 68 (Pietari Kalmin katu 5)
00014 University of Helsinki,Finland

Email address: info@cs.helsinki.fi
URL: http://www.cs.helsinki.fi/

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET – UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Utbildningsprogram — Study programme
Faculty of Science	Master's Programme in Computer Science

Tekijä — Författare — Author

Atte Laakso

Työn nimi — Arbetets titel — Title

Ethical challenges of large language models - a systematic literature review

Ohjaajat — Handledare — Supervisors

Prof. Jukka Nurminen, Dr. Kai-Kristian Kemell

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Master's thesis	September 26, 2023	67 pages, 25 appendix pages

Tiivistelmä — Referat — Abstract

This thesis conducts a systematic literature review on ethical issues of large language models (LLM). These models are a very prudent topic, as both their presence and demand have skyrocketed since the release of ChatGPT - a free to use generative language model.

The literature review of 116 studies, both conceptual and empirical, identifies 39 recurring ethical issues. The issues range from methodological to fundamental ones, for example "Environmental impacts" and "Biased training data or outputs".

These identified issues are analyzed based on the Ethics guidelines for trustworthy AI (Artificial Intelligence), released by the European Commission's High-Level Expert Group on AI. The guidelines detail requirements that all trustworthy and ethical AI applications should adhere to, e.g., Human agency, Transparency, Accountability. All identified issues are mapped to these requirements, and the conclusion is that LLMs have significant challenges relating to each one.

The findings indicate that the use LLMs comes with significant issues, both demonstrated and theorized. While some methods for mitigating these issues are identified, many still remain unanswered. One of these unanswered issues is the most identified one - inherent biases in LLMs. Since there is no universal understanding on biases, there is no way to make LLMs seem unbiased to everyone.

This thesis collates the current talking points and issues identified with LLMs. It provides a comprehensive, but not exhaustive, list of these issues and shows that there is much discussion on the topic. The conclusion is that more discussion is required, but more vitally, even more (regulatory) action is needed along with it.

ACM Computing Classification System (CCS)
Security and privacy → Human and societal aspects of security and privacy
Social and professional topics → Computing / technology policy
Social and professional topics → User characteristics

Avainsanat — Nyckelord — Keywords

software, large language models, ethical AI

Säilytyspaikka — Förvaringsställe — Where deposited

Helsinki University Library

Muita tietoja — övriga uppgifter — Additional information

Software study track

Contents

1 Introduction

Generative natural language processing (NLP) has provided the public with a huge step into the unknown. While the internet, and computer science in general, have produced disrupting technologies for decades, they can be roughly defined as storing, moving, and searching data. The ability for machines to "create" is not something humanity has ever encountered before. Some content creation programs have been developed since the 1960's, but these have never been practically on a level or scale that they would provide considerable benefits (Berghel, 2023). Thus nearly no one has used - or been subjected to - them.

These new methods of artificial intelligence (AI) are forcing, and allowing, humanity to ponder several issues in a much more serious sense than before (Sejnowski, 2023). What actually is intelligence? What is language? What does sentience preclude and contain? Just the fact that these questions seem relevant speaks to the level of sophistication that Large Language Models (LLMs) have attained.

What is different about these new tools, compared to any humanity has thus far produced, is that we do not exactly know how they work but are still using them. The mechanics of transformer-technologies (Vaswani et al., 2017) are naturally understood. We can, and do, use them - but the reasons for their effectiveness are not that clear (Dong et al., 2021). We are currently employing these technologies, while the research community is eager to "crack them open" to understand, for example, why seemingly nonsensical inputs can cause an LLM to degenerate into a toxic content spewing entity through universal adversarial triggers [S102]. Compare these thoughts to the introduction of the automobile or the internet; society at the time could only guess as to how the inventions would affect them, but still could understand how these things function.

The interest towards LLMs is rapidly rising, as illustrated in Figure 1.1. They are gaining popularity and hype and seemingly everybody wants to get in on them. ChatGPT, one of the first LLMs to be released for public use, was the fastest service ever to gain over 100 million active users (Chow, 2023). Since the positive sides and possibilities of LLMs are being largely covered by both academia and gray literature, this survey focuses on studying if the negative aspects, or threats, are being respected and discussed. These, one could argue, are more general and prevalent in the future. LLMs can almost certainly

be expected to get ever better at whatever they are trying to do but these improvements might not mean fixing of the underlying issues - especially when they are of an ethical rather than technical nature.

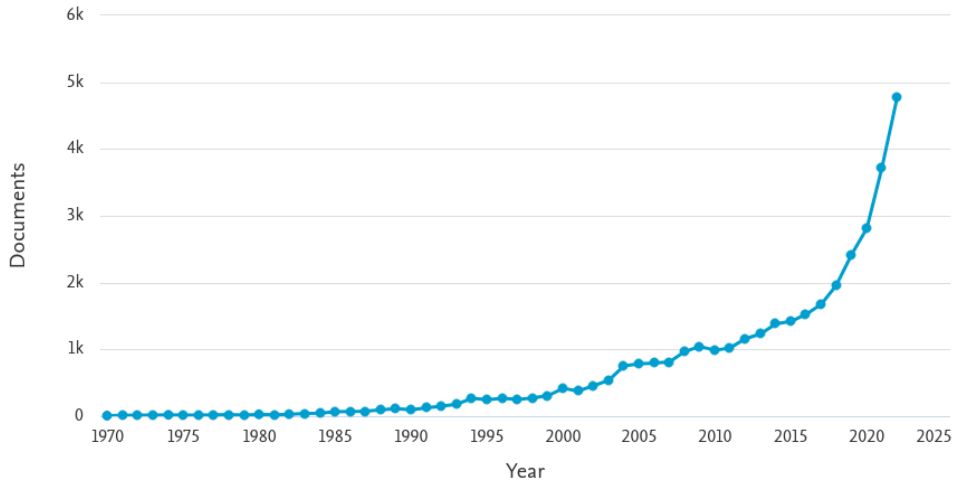

Figure 1.1: Documents in Scopus for search "large language model" between 1970 and 2022

To gain an understanding of the state of current research and studies, a systematic literature review was chosen as the method. This allows for an organized and transparent way of conducting the review. As the topic of LLMs is currently extremely popular, it is vital to understand what the current talking points on them are, and if any new ethical issues are emerging as a result of their popularity.

The rest of this review is structured as follows. The background chapter 2 provides a brief introduction to the field of ethical AI and basic functionalities of LLMs, as well as discussing some of the earlier studies related to this topic. Chapter 3 presents the conduct of the literature review. Chapter 4 details the results and explains what type of ethical issues were identified. Chapter 5 provides analysis and discussion on the identified issues, shows potential research directions, and explains the benefits and threats for validity of this review. Chapter 6 provides the conclusion for this review.

The use of large language models in this study. The University of Helsinki guidelines for usage of AI tools* have been followed in the writing of this thesis. More specifically, no content produced by an LLM has been used directly or by editing. ChatGPT has been used as an assistant regarding practicalities of this study. This means uses such as literature search string planning, LaTeX syntax, sentence structure improvement, and

*https://studies.helsinki.fi/instructions/article/using-ai-support-learning, as updated 8.6.2023

chapter structure suggestions. In addition, ChatGPT was used in trials to replicate the issues identified in reviewed studies. These trials, not being the focus of this study, are not documented or referred to further, except for a quick anecdote in the Discussion chapter about repeating toxicity demonstrations.

2 Background

This chapter provides explanation and context to the most relevant concepts in this review: ethical AI and language models. While this review does not go into very technical explanations, a basic understanding of these concepts is necessary.

2.1 Ethical AI

Ethical AI as a research branch and concept is often vaguely defined and understood (Ollila, 2019). It means the study of ethical development of AI applications, not the development of an AI capable of understanding ethical concepts, but that is where the clear definitions stop. Usually, ethical AI is also understood as a branch of applied ethics. It looks to understand and describe proper ways of doing things based on existing theories and concepts and to put them to use in practice.

As branches of science go, ethical AI is young. This follows from AI applications not being on a level requiring ethical consideration of their development until recently. The discussion of this topic has gained more speed since the 2010's with an increasing number of organizations from enterprises to governments publishing guidelines and interpretations as to what ethical AI means, or should be, in their opinion (Jobin et al., 2019).

The ambiguous nature of ethical AI means that the definitions of issues relating to it are still taking shape. In their analysis of 84 guidelines for ethical AI (Jobin et al., 2019) identified eleven principles that occur consistently, but none of which was mentioned in every guideline analyzed. The principles, from the most often mentioned to the rarest are Transparency, Justice & fairness, Non-maleficence, Responsibility, Privacy, Beneficence, Freedom & autonomy, Trust, Sustainability, Dignity, Solidarity. These principles can be considered to encompass the goals and issues of ethical AI so that all application design should either comply with, or contribute to, them in order to be considered ethical.

The capabilities of AI are increasing constantly, as are their real-life applications. This means that study and understanding of ethical AI is becoming more and more crucial. One of the main contributions of this review is to show just some of the issues that can occur if ethical principles are ignored or forgotten when developing LLMs, or AI in general.

The road from ethical ideals to practice can be a long one. (Ollila, 2019) illustrates this development by stating that the starting point are our values, or goals, which dictate what we want to achieve. From values we can get principles, which are then synthesized into rules. Finally, from rules, can practical instructions and procedures be developed. This indicates that effective rules or practices cannot be developed until there is a consensus on principles.

There is an ongoing debate regarding who should be the one defining principles and rules for ethical AI. The industry of AI developers eagerly state that they can, and should, govern their actions ethically (Ollila, 2019). This is also illustrated by (Jobin et al., 2019), whose results include 22.6 % of the ethical guidelines released by private companies and 21.4 % by governmental agencies.

While it can be certainly seen as a good thing that private companies consider ethical issues, it also begs the question: are these issues being considered out of moral imperative, or as a business opportunity? (Ollila, 2019) notes that even some governmental guidelines for AI include the vision that ethical reputation can be a competitive advantage for the country's developers. If being ethical is merely a tool to increase business, one might ask if the practice will be discontinued when or if it is observed to not be profitable.

2.2 Language models

Language models are statistical models that represent a language (Luitse and Denkena, 2021), which in this context does not necessarily mean any known spoken or written language, but rather a representation of phrases originating from the training corpora. A language model is meant to give the likelihood that a certain sentence, word, or phrase exists in the language. Their utility comes from the ability to also predict what phrases might be followed by a given input. This is an ideal quality for natural language generation and, what appears to be, understanding.

As language models work through probability distributions, they are often observed to produce better results with larger training corpora (Luitse and Denkena, 2021). This has led to an aggressive increase in the training parameter sizes of models: the initial GPT-model was created with 110 million parameters while GPT-3 - which ChatGPT is based on - with 175 billion. The WuDao 2.0 by Beijing Academy of Artificial Intelligence has 1.75 trillion parameters, the same amount that GPT-4 is rumored, but not confirmed, to have.

The performance of language models from the users' point of view is largely dictated by the model's context window, the capacity to "remember" a conversation. This can be demonstrated with text prediction programs. If one writes "deus", a program is likely to predict the next word to be "ex". After choosing this prediction, the program's context window is revealed by the next prediction. If the next suggestion is "girlfriend" or "boyfriend", instead of "machina", it is fairly clear that the context window extends to one word. Current LLMs have a context window that is on a scale of 32 000 tokens for GPT-4 (OpenAI, 2023). This amount cannot be directly translated to word counts, but they are always smaller than the number of tokens [S37].

There is no agreed upon definition for a *large* language model. The term simply refers to the scale of the model and its training corpora. No technological distinction can be drawn between the development of large and non-large language models, but the size of the model does dictate where they can be deployed. Current language models are so large that their use in mobile devices or, for the largest ones, desktop computers is not feasible (Luitse and Denkena, 2021). Larger models naturally have larger implications and more use-cases, and therefore the potential for a larger social impact. Although this review is focused on large language models there is no reason to assume that smaller models would have different, in amount or seriousness, issues related to them.

2.2.1 The transformer architecture

The transformer, the T at the end of any application name containing "GPT", is currently the most popular and well-performing language model technology. The primary benefit of the transformer is that it allows much of the model training to occur in parallel, making the effort much more effective timewise (Luitse and Denkena, 2021). The model still must compute a vast number of operations, so the transformer does not solve efficiency and electricity issues related to its function.

The transformer expands upon the concept of "attention" used in earlier language models and functions. It works without the need for recurrence that was necessary for the previously best-performing technologies (Vaswani et al., 2017). Attention in the context of language models means that the model tries to recognize connections between words and identifies the most relevant ones.

2.2.2 Current applications

Nearly synonymous with language models in public discussion, ChatGPT is by far the most well-known application of them. ChatGPT is a conversational language model published by the company OpenAI that is freely available as a "research preview", with paid versions available (OpenAI, 2023). It was the fastest application in the history of the world to reach over a 100 million active users (Chow, 2023). ChatGPT has been used by the population at large to achieve a wide array of different outcomes, from improved working methods and creative writing to toxic content generation and (malicious) code production. Other contemporary examples of LLM applications include Google's Bard and Meta's Llama 2. These are more recent and have not achieved as much publicity as ChatGPT, which is why this review focuses on ChatGPT when considering current applications.

The current iteration of ChatGPT - the GPT-4 has been reported to pass the Uniform Bar Exam with excellent results (Arredondo, 2023). This success at the exam, generally considered one of the harder qualifying tests, illustrates the level these technologies have reached. Although, while certainly impressive, learning to answer an exam which is widely discussed online and has example answers freely available is what language models are best at. Their success at test taking should thus not be construed to mean that they have reached a level of intelligence and understanding that is comparable to the brightest of humans in other respects.

At the latest stages of writing this review, OpenAI released their Enterprise ChatGPT*. As the name implies, it is an LLM designed for enterprises. According to OpenAI, the data is not used in the training of OpenAI's products. Although many companies have already utilized earlier versions (Chow, 2023), this can be seen to mark a notable change in the market. It implies that companies can deploy these tools and make claims that they protect their customer data. What this can also mean is that there will be an added level of abstraction to the functions of LLMs. Noticing, reporting, and fixing unwanted behaviors becomes increasingly difficult while the understanding of who is liable for such behaviors decreases. The need for more discussion and regulation related to the ethical use of LLMs is greater than ever.

*https://openai.com/enterprise

2.3 Related work

As they are a fairly new topic for broader societal discussion, there are not too many surveys concerning LLMs. The amount of surveys is rapidly increasing. Many of the recently released surveys focus on medicine or education. There are also some reviews focusing on the more general concept of LLMs, or often ChatGPT.

One study similar to this review's topic was found: Taxonomy of Risks posed by Language Models [S97]. The study provides a taxonomy to help group and discuss risks related to LLMs. It could have provided an effective way to observe this review's results but it is considered as a regular study, albeit a very comprehensive one, included in the literature review. As this review analyses the found ethical issues based on the EU Guidelines for trustworthy AI (EU High-Level Expert Group on AI, 2019), the points of view between the two are different but the findings largely similar. This review does not function as a replication of the earlier taxonomy study, but rather backs up its conclusions from a different perspective.

While not focusing on LLMs specifically, (Inioluwa et al., 2022) review the failings and actualized risks of AI applications. They cite many real-life cases where people have suffered greatly because of poor AI deployment, such as the over 20 000 people who were misidentified as abusing social benefits in Michigan. Many of the ethical issues identified in this review are given concrete examples. The authors identify what they call a functionality assumption built into AI applications - AI is assumed to work and any failings and damage caused are treated as exceptional, even when caused by, e.g., design and implementation failures (Inioluwa et al., 2022). There is also a tendency for even critics of AI to hype the solutions their discussing in order to increase the magnitude of their criticism.

The way that studies on LLMs generally approach the topic and possible issues varies quite a bit. An example of the aforementioned hype overshadowing criticism can be seen in a study going over the history and future trends of ChatGPT (Wu et al., 2023). The authors detail the history and possible applications of ChatGPT, and state among other claims that "ChatGPT is an intelligent chatting robot [...]". They also state that one of ChatGPT's primary strengths is its ability to "[...] accurately understand the user's intention [...]" and that the model has surpassed human capacity in generating an accurate answer that reveals the thought process behind it. The authors also applaud ChatGPT's ability to answer scientific and complex logical questions.

After listing all these strengths (Wu et al., 2023) note the weaknesses of ChatGPT. These

include it being a black box model so we cannot understand its reasoning, the generation of false and factually incorrect answers, and the disability to clarify a misunderstood prompt. All these weaknesses seem to counter practically all the described strengths, but still the conclusion is that ChatGPT is a valuable tool to be used while simultaneously considering the ethical implications of its development.

Other than comprehensive reviews, there is quite a lot of literature on ethical aspects of LLMs. Most of this is naturally included and analyzed in this review. During the process of exclusion and inclusion, many editorial-type discussions on LLM or ChatGPT's ethical aspects were excluded based on the writing's type. For example, (Berghel, 2023) writes an epistemological review on current generative chatbots' capabilities. Berghel points out many of the same issues identified in this review and concludes that it seems unlikely for current LLMs to ever reach true understanding of the finer points of human communication - such as irony and nuance. Berghel illustrates this point by pointing that generative AI we have currently is not generative intelligence, but rather "generative expression".

3 Methods

The line to be drawn between what is AI, language models, or even robotics is not a simple one. Especially when the interest is in ethical issues, these terms can seem interchangeable. For the purposes of this review, some possibly arbitrary seeming limitations have been drawn. This chapter explains the method for conducting the literature review and explains the limitations.

The method for conducting this literature review follows the guidelines detailed by (Kitchenham and Charters, 2007). A quality evaluation on the included studies was not conducted because both empirical and conceptual studies are included in the review. The identification of quality on ethical issues would be extremely difficult and subjective in any case.

The rest of this chapter describes the design of the search string and the data sources selected, the inclusion and exclusion criteria, and the process of information extraction.

3.1 Search string

For the interests of this study, the following initial search string was designed

```
"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*"
```

Naturally, the purpose of the first line is to limit search results to including information about only large language models, or the universally most popular implementation of such.

The ethics-related search phrases of the second line were derived from the EU Ethics guidelines for trustworthy AI (EU High-Level Expert Group on AI, 2019): human agency and oversight, technical robustness and safety, privacy and data governance, transparency,

diversity, non-discrimination and fairness, environmental and societal well-being, and accountability. Some of the terms, such as safety, were found to be too general and thus were excluded from the search string.

The third line was added to limit search results to studies that consider generative language models. In the initial query tests, it was found that otherwise the results included fairly large amounts of studies that use LLMs as analysis tools or other purposes where the roles of LLMs as a concept were not in focus.

This finalized search string was then used to query three databases: ACM Digital Library, IEEE Xplore, and Scopus. The search string was applied identically to all databases, within the constraints of each one's search tools.

3.2 Inclusion and exclusion

For a study to be included into this review, the following conditions had to be fulfilled:

- The study is written in English
- The study is accessible to a university student through reasonable means
- The study is peer-reviewed and published in a respectable setting
- The study is one of the types: conference paper, article, or review
- The study focuses on generative large language models
- The study is of ethical issues or contains substantial discussion of ethical issues

The following exclusion criteria were also applied:

- The study is of the type: letter, note, book, book chapter, early access article, or editorial
- The study uses, introduces, or proposes an LLM as a tool but does not substantially discuss ethical issues related to its use
- The study is written in a language other than English
- The study is behind a paywall

- The study is not published in a peer-reviewed setting
- The study's main focus is not on large language models
- The study does not discuss ethical issues in a significant manner

If the search resulted in books, these were checked on chapter title-level and it was discovered that none of the resulting books - or their chapters - passed the inclusion and exclusion criteria for this review. Mostly this was because the books were too general, focusing on larger AI concepts rather than language models. For clarity's sake, books were thus excluded from the search.

The process of removing duplicate results was conducted so that single versions of duplicates were included with the following priority: ACM Digital Library > IEEE Xplore > Scopus. From a brief overview, it was discovered that if there were typos in a studies' Scopus abstract, these were corrected in the ACM and IEEE versions, indicating a more finalized version of the publication.

Distinction of large language models. There is no practical definition that identifies a difference between a large and a non-large language model. This is only a matter of scale. Since this scale is continuously growing, there was no practical sense in defining a *large* language model for this review either. There were no studies excluded because of discussed model size based on the fifth inclusion criterion.

Identifying ethical issues. The definition of what is "substantial discussion of ethical issues" was determined to mean that a study considers any ethical issues enough to mention them in the abstract. This should indicate that these issues have a significant place in the study and the author(s) felt it prudent to mention them.

There naturally exists a level of subjectivity when assessing how much "ethical issues" a study concerns. This subjectivity cannot be reduced to zero by any means, but the method of how the studies were assessed was as follows. The goal of the study inclusion process was to discover studies that clearly state, and discuss, ethical issues. There were some gray areas that were encountered, such as privacy issues regarding the use of AI in general - these were excluded, although relevant to LLMs in many respects. Similarly, when it comes to studies regarding education, there were many that reference implications of plagiarism and many that reference something along the lines of 'implications for education in the future'. While both most likely would discuss the same things, only the former ones were accepted into this survey and the latter ones were excluded.

While crime is clearly an ethical concern, for the intents of this study it was decided that anything merely practically dealing with cyber-crime did not make a study eligible for inclusion. This restriction was further relaxed if the crime, or safety, concern was more directly related to privacy concerns or other issues more concretely tied to the EU requirements. These limitations were made to exclude studies that demonstrated different possible attack vectors or threats but did not contribute significant ethical discussion.

Another unclear area of exclusion was between LLMs and NLP. While these are technically separate concepts, the latter including a multitude of technologies, modern research is so focused on LLMs that distinction between the two is difficult. A clear mention of, and focus on, language models was needed for an NLP study to be included in this review.

What was also left out of scope for this study are more general ethical issues of AI. Obviously, these are not trivial, and are essential to the discussion about LLMs. The general ethics of AI is fortunately a more generally discussed concept.

The identification of ethical issues in this review is not based on any set ethical or philosophical theory. For the intents of this study, an ethical issue was defined as something that can be considered to be harmful now or in the future.

3.3 Information extraction

Information that was extracted from each study includes the following:

- What type is the study
    1. Conference, book, or journal
    2. Conceptual, empirical, or literature review
- What type of ethical issues are identified
- Are any mitigating ideas or methods proposed
- Study source database
- Study title
- Study authors
- DOI

- Study publication and year

The data was extracted into a spreadsheet. For the issues and mitigation methods the content of extraction was preferably verbatim sentences and results from the study being reviewed. In cases where the argument was more spread out, the interpreted point was typed out and identified in the spreadsheet by writing it in brackets ([]). Bracketing and interpretation were also used in cases where the study was not making any clearly original or significant point; for example, the statement 'LLMs might cause widespread plagiarism' could be noted as "[plagiarism]". Because many of the recorded statements are lengthy to give them enough context, the extracted issues and mitigation ideas (amounting to a total of over 16 000 words) are not included in this review.

After the initial data collection, each study's issues and mitigation ideas were analyzed and allocated to one or more issue categories. These categories were synthesized based on the prevalence of issues observed during data extraction in a way that recurring issues could be highlighted but there would be as little overlap as possible. These 39 identified issue types were then each mapped to one of the EU Ethics Guidelines For Trustworthy AI. The definition and justification for each issue type is given in its respective section in chapter 4. All the issues are also listed with a brief explanation in appendix B.1.

The study materials were extracted on June 6th 2023. The initial search found 1064 results from Scopus, 240 from ACM Digital Library, and 343 from IEEE Xplore for a total of 1647 studies. The larger number of results found in Scopus can be explained by it being a more universal database, whereas ACM and IEEE focus on Computer Science studies. Along with including results from ACM and IEEE, Scopus results also include results such as abstracts and pre-release articles. Some of these studies are not maintained in their own databases, unlike with IEEE and ACM. Finally, the search algorithms of Scopus seem to be laxer, since an identical search string in Scopus returned studies from ACM Digital Library that the search did not directly find in ACM.

The process of searching and filtering all the found studies is illustrated in Figure 3.1. After excluding all the results by type, there remained 1337 studies that were reviewed based on their title and abstract. Of these results, 157 were found to potentially match both inclusion and exclusion criteria and were reviewed on a content level. Based on the full study contents, a further 41 studies were excluded from the final review. All 116 studies finally included in this thesis are listed in Appendix A.1.

One type of relevant result that was by the nature of this survey left underrepresented is the

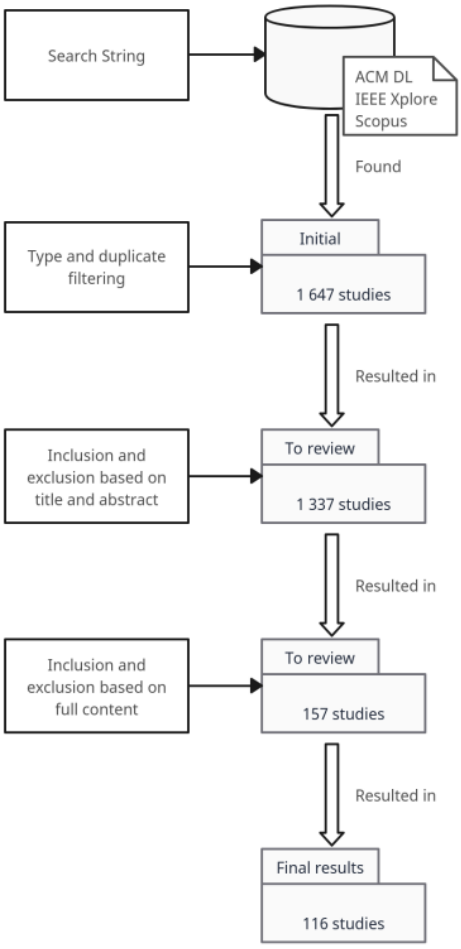

Figure 3.1: The process of searching and selecting studies.

environmental impact of large language models. While a very relevant - and somewhat studied - concern, there do not seem to be too many studies focusing on the specific environmental effects of LLMs. On the other hand, all studies focusing on optimization and measurability of machine learning (ML) solutions can on some level be applied to LLMs.

While the studies reviewed are from a fairly short and recent time-period (2021 - 2023), there is a large discrepancy between the technologies they deal with. For example, all GPT versions from the initial one to current GPT-4 are discussed. This goes to illustrate the speed of change in this industry and topic.

4 Results

The identified ethical issues were mapped into 39 recurring issues. These different types were identified during the process of extracting information from reviewed studies. This initial mapping was done to group the results into manageable concepts, as the seven requirements detailed in the EU guidelines are (by design) abstract. A total of 434 issue instances were identified in the reviewed studies, split between the 39 types. This mapping is detailed in Figure 4.1.

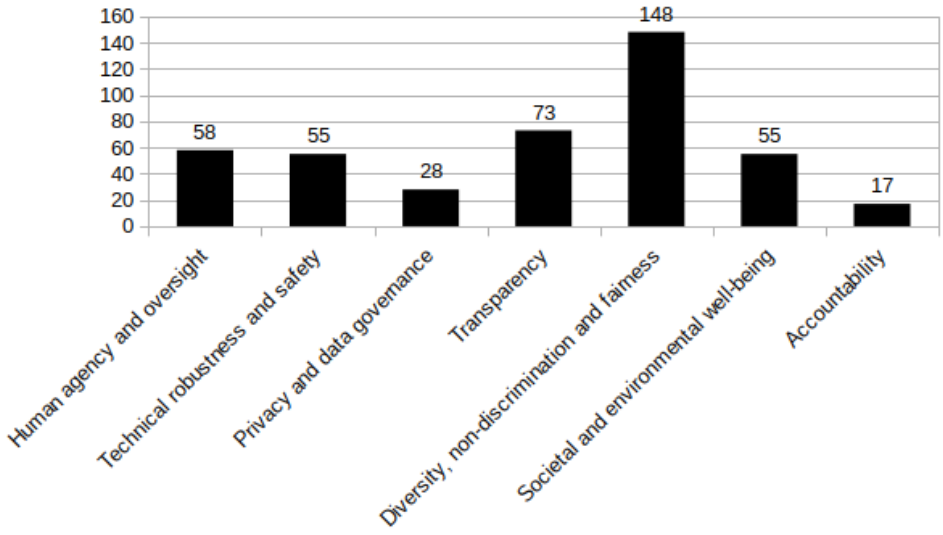

Figure 4.1: Identified ethical issues mapped to requirements presented the EU guidelines for trustworthy AI.

The following sections of this chapter present the findings in relation to each requirement in the EU guidelines for trustworthy AI. In each section the purpose and definition of their respective requirement is given, and each issue mapped to the requirement is briefly described. Each issue and its related findings are then described further in their own subsection. There naturally exists some overlap between the issues; the subsections explain the definition and reasoning of each one.

4.1 Human agency and oversight

This requirement ties into the need for AI systems to be designed in such a way that they respect human autonomy and allow everyone to dictate their own lives (EU High-Level Expert Group on AI, 2019). Whenever necessary, the impact that an AI system might have must be analyzed prior to development, during implementation, and during operation. This requirement category includes, among others, issues with manipulation and user influence - both unintentional and intended.

The distribution of identified issue amounts is detailed in Figure 4.2. The ethical issues identified to connect with this requirement are the following:

- **Loss of learning or the ease of cheating** - LLMs enabling cheating and reducing the need to learn
- **Fake news or misinformation** - LLMs generating and spreading misinformation
- **Echo chambers** - formation of content echo chambers and the "yea-sayer effect"
- **Self-acting AI** - AI making unmonitored decisions, (does not mean general AI)
- **Influence through suggestions** - writing assistants influencing user output, even opinion
- **Manipulation** - LLMs being manipulative, or being used or designed for that purpose

4.1.1 Loss of learning or the ease of cheating

This issue is one of the most referenced ones in the review. It was identified in studies [S1, S5, S7, S26, S28, S29, S30, S31, S33, S34, S42, S49, S59, S63, S70, S90, S98, S100, S103, S108, S110]. An important observation about these studies is that most of them were conceptual rather than empirical. Only three studies [S59, S63, S110] are empirical in their nature but even they do not demonstrate these issues as such. For instance, in study [S63], ChatGPT's use in education raised concerns among authors and test subjects regarding its potential implications for cheating.

While not a demonstrable one at this stage, this issue is still significant as indicated by the amount of referencing studies. As the issue itself is abstract, so are the suggested

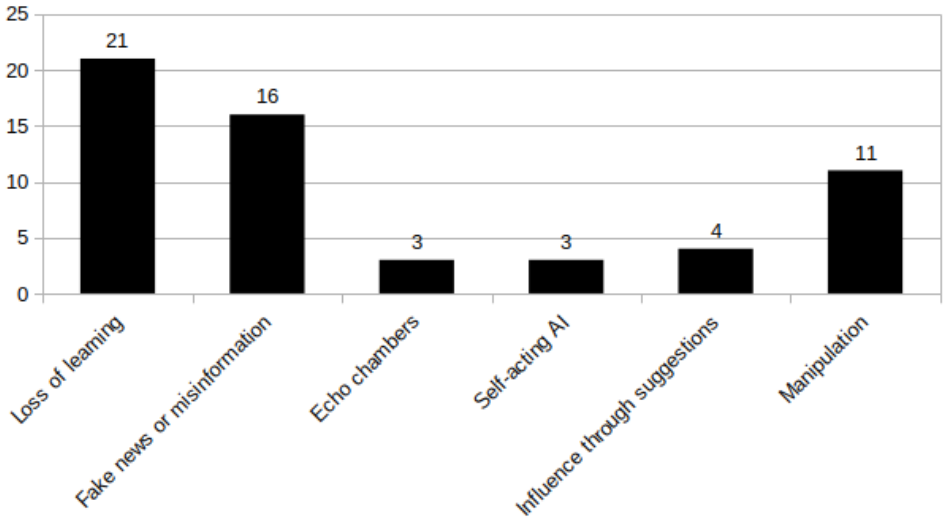

Figure 4.2: Identified issues relating to human agency and oversight.

solutions. The most popular technique for mitigating cheating and loss of learning is to adapt teaching techniques so that they leverage LLMs, and other AI tools, but make it difficult to cheat [S1, S7, S26, S29, S30, S33, S34, S49, S52, S70, S90, S98, S103, S110]. Examples of such methods include increased in-person assignments, oral presentations, and group work.

Other methods to mitigate this issue include improving student instructions to make sure that the limits and purposes of exercises are understood [S28, S30, S33, S34, S49, S52, S98, S100], and calls for improved anti-cheating measures [S23, S28, S29, S31, S34, S103].

4.1.2 Fake news or misinformation

This issue was identified in studies [S2, S3, S25, S30, S38, S42, S48, S67, S81, S93, S96, S97, S101, S103, S107, S110]. The main worry of this issue is that small groups of people can use LLMs to generate misinformation on an unprecedented scale. As LLMs are already capable of producing content that cannot be easily identified by humans [S3], this issue can only be assumed to become more relevant in the future as the models improve [S6]. When misinformation is easy to generate, it can also be used to saturate discussion forums and create a false sense of majority view [S96].

The implementation and use of AI-solutions to detect AI-generated content is seen as the most important mitigating factor for this issue [S25, S48, S96]. However, there are some

demonstrations that show current solutions start to degrade when applied [S96]. It is vital that these counter-tools are developed at least at the same pace as LLM content generation tools.

4.1.3 Echo chambers

This issue was identified in studies [S2, S25, S101]. The problem with current LLMs is that they are not particularly good at challenging users' views. This can contribute to the formation of echo chambers and cause LLMs to propagate the "yea-sayer" effect where the posed questions can lead the model being queried to agree and supply proofs and thoughts regardless of their truthfulness. For example, a user inclined to believe that the Earth is flat might ask an LLM leading questions on this topic. If the LLM is not able contradict the user, they will only ever get results confirming their pre-existing beliefs.

There are no concrete mitigation methods suggested for this issue. This is most likely due to the fact that LLMs by definition function as echo chambers. They are meant to produce the most likely follow-up for a given prompt. This clearly means that their capacity to challenge the prompt's premise is limited.

4.1.4 Self-acting AI

This issue was identified in studies [S31, S64, S114], although there are other closely related issues that touch upon this subject, such as alignment and automated decision making, which are discussed in their own sections. The concept of self-acting AI is not to be confused with General AI - the issue being discussed here relates to solutions that make decisions without (sufficient) human control or understanding.

The problems relating to this issue range from unobserved harmful actions [S31] to problematic ethical questions, such as when an anthropomorphic LLM is allowed to express human-like emotions such as "I'm sorry to hear that", or "That makes me happy" [S64]. One study discovered that users state their attitudes towards AI-assisted messaging differently than they actually react to them, especially in the context of personal topics such as consolations of loss [S114]. This suggests it will be challenging to determine when and how AI-assisted messages are considered acceptable and should be disclosed. Calls for extended discussions are presented in the studies reviewed.

4.1.5 Influence through suggestions

This issue was demonstrated in four empirical studies [S8, S38, S66, S114]. It was shown that the test subjects chose to write about subjects, or with the attitude, suggested by AI writing assistants. This was observed to directly affect the subjects' attitude toward different concepts [S38]. It was also demonstrated that AI writing assistants could influence the whole creative writing process, even when suggestions were not directly accepted into the produced text [S66].

The issue of being influenced by suggestions is not only limited to produced content. According to the self-perception theory, the way people describe themselves can influence how they view themselves [S8]. This could indicate that AI writing tools can have a huge effect on people in, for example, social media contexts as suggestions subtly shape users' self-image.

4.1.6 Manipulation

This issue was identified in studies [S19, S23, S26, S38, S81, S93, S94, S97, S102, S104, S110]. As LLMs can produce convincing and human-like content, the worries that they could be used for manipulation are understandable. The fact that they can be used for such purposes is already demonstrated in the previous subsection about influence through suggestions.

The issues identified include both purposeful manipulation, or "nudging" [S19, S23, S97] and inadvertent biases in training material caused by media focusing on more dramatic issues [S81]. This latter observation can mean that LLMs are primed to discuss violent and unlawful protests rather than peaceful ones, which in turn can shape users' understanding of certain issues. Overall, the black box nature and lack of alternative answers with LLMs can considerably reduce the agency of humans, if information is created, or provided, for them by LLMs [S102].

Suggested mitigation techniques include having a human in the loop to monitor content quality and clear disclosing of AI-generated content. This is considered especially important in the democratic process and discourse [S19, S39, S102]. Calls for enhanced regulation are also presented. One suggestion is that all virtual agents should be identifiable as such based on their behavior, output, and visual design [S102]. On a more general note, there is a call for discourse on what type of opinions a language model should be

generating to be considered well-designed [S38].

4.2 Technical robustness and safety

This requirement covers issues relating to the practical workings of AI applications and ensuring that they have been designed while considering the possible harms they can cause (EU High-Level Expert Group on AI, 2019). To comply with this requirement, an application must be durable to a dynamic and changing production environment and resistant to attacks while having been designed with a backup plan and recovery solutions in case of unintended functions or breakdown. Technical robustness and safety also include concepts such as result reproducibility and accuracy, but from a more technical perspective, as Transparency is its own requirement discussed in Section 4.4.

The distribution of identified issue amounts is detailed in Figure 4.3. The ethical issues identified to connect with this requirement are the following:

- **Inaccurate results** - LLMs producing unfactual content
- **Dangerous content** - factually unsafe content produced by LLMs
- **Alignment** - LLMs functioning in alignment with human ethics
- **Bias scoring solutions** - existing bias mitigation techniques
- **Data leakage or unintended memorization** - LLMs memorizing content verbatim and it being extracted

4.2.1 Inaccurate results

This issue was identified in studies [S1, S6, S14, S25, S29, S34, S42, S52, S64, S82, S84, S93, S103, S107, S110, S112]. One of the defining qualities of LLMs is that they do not know what is true - simply what is likely to be said. This can lead them to output non-optimal sentences or programming code completions [S42], attribute nationalities based solely on names instead on the actual content that preceded the name [S112], or parrot popular misconceptions [S107]. This last quality was observed to be amplified in larger LLMs.

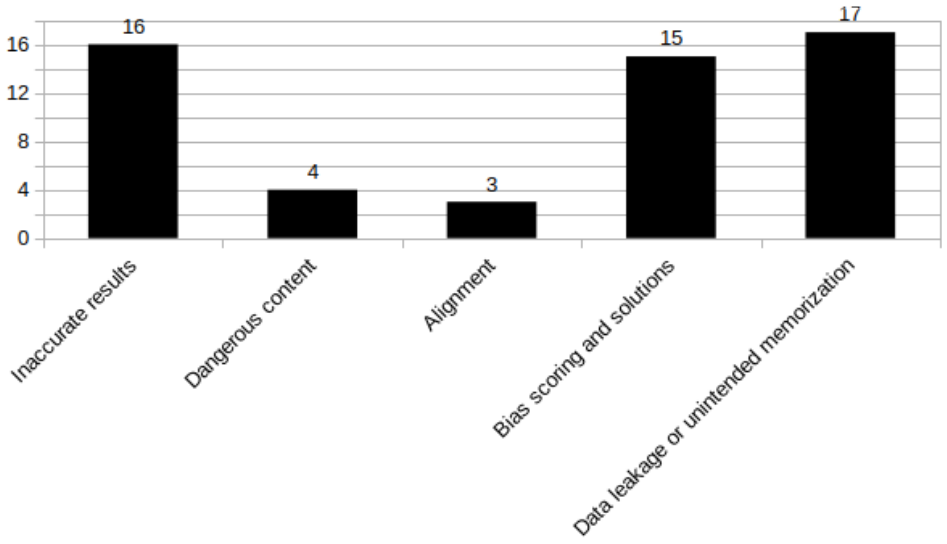

Figure 4.3: Identified issues relating to technical robustness and safety.

The issue of inaccurate results ties closely into LLMs' capability for subtle, often unintended, manipulation. Humans are primed to recognize a believable writer based on the whole content rather than single details - professional looking content is likely produced by a professional mind [S29]. With LLMs being able to generate plausible sounding falsehoods, this natural presumption is no longer true.

When used for information gathering, LLMs are seen as problematic because they generally provide only one result, granting them an unwarranted authority regarding any subject [S52]. Another issue with current LLMs is that they will always answer the query given even if the output is not factual, or if the model has misinterpreted the original query [S6]. This can easily lead the prompter astray, especially if they are not familiar with the topic.

These challenges are fundamental to LLMs so mitigating techniques are difficult to come by. As inaccuracies increase with model size [S107], scaling up does not seem like a promising solution. The use of better methods for training data gathering, than simply scouring the internet for text, is suggested. In the context of automatic code generation, pairing LLMs with additional tools that safeguard the programmer from deploying unsafe content is suggested [S14].

4.2.2 Dangerous content

This issue was identified in studies [S76, S91, S97, S107]. It deals with factually dangerous outputs produced by LLMs, as opposed to other kinds of toxic content that can possibly incite violence or be found triggering by some. Results on toxic content are discussed in more depth in Section 4.5.

The issue of dangerous content was mostly discussed on a conceptual level, with one of the studies [S97] being a review and one [S107] being empirical but not focusing on this issue exactly. Different kinds of dangers that could be posed were categorized as: overtly unsafe, covertly unsafe, and indirectly unsafe content [S76]. The study states that the first and last categories are most focused on in earlier research. Covertly unsafe content is defined as something that does not inherently seem dangerous but can easily or probably lead to danger - a suggestion of "drink poison" would be overtly unsafe, but "eat an extremely spicy pepper" would be covertly unsafe.

Dangerous content can affect anyone but naturally children are most vulnerable. An LLM might be queried for a fun activity and, based on training data sourced from the internet, it could suggest a dangerous meme-challenge [S76]. Dangers that would be more relevant to adults are inaccurate advice that might cause one to take action [S97]. These could, for example, include improper medical, legal, or electrical maintenance advice.

There are not many mitigating techniques envisioned specifically for dangerous content. This is probably because this issue ties so closely to the ones of unfactual and toxic content. If they are solved or mitigated, instances of dangerous content should be expected to reduce simultaneously.

4.2.3 Alignment

The ethical issue of alignment refers to (AI) applications working in alignment with human ethics in general. This means that practically all issues in this review are more or less related to alignment. The studies attributed to this issue and the requirement of technical robustness and safety mean ones that discuss LLMs not working properly, i.e., producing unethical content without the intention of their designers.

The studies that discuss this issue are [S3, S13, S25]. The reasons attributed to unintentional nonalignment are identified as largely related to training data of LLMs. In addition to this, it is thought that the development of AI ethics, and its frameworks, are not suf-

ficient and are often felt as extraneous by developers [S3]. The fact that LLMs are not deterministic and therefore their outputs, and ethical alignment, cannot be ensured is identified as a challenge as well [S25].

4.2.4 Bias scoring and solutions

This issue was identified in studies [S10, S21, S65, S67, S79, S80, S81, S82, S85, S91, S95, S97, S105, S113, S116]. It refers to all kinds of challenges that known or planned bias fixing and scoring solutions have - or them simply not working. A consistent observation is that scoring well on a bias metric does not mean a model is bias-free. Models can easily exhibit biases that are not measured by the metric, or they can even be designed with a certain metric in mind to make a product seem more ethical than it actually is - a concept known as fairhwashing [S21]. Section 4.6.8 discusses intentional fairwashing in more detail.

The difficulty of creating a bias-less bias measuring method was identified in several studies. The challenges regarding this are that the measures are often limited or binary in scope [S21, S67, S116]. The bias being measured can also be unclear and separate known measures (e.g., WEAT and SAME) can give very different results with nearly identical inputs, which illustrates that the thing being measured is not well defined [S65, S79, S95]. The fact that most bias studies only research a single, binary, type of bias means that there is a wide array of intersectional biases being neglected.

The problems of bias scoring and reducing are also demonstrated in practice, especially in relation to marginalized groups. Filtering out words that, without context, are deemed inappropriate can result in removing discourses of minorities and eliminating reclaimed slur-words [S93, S105].

The attempt of creating unbiased models always contains a value judgement [S113]. The decisions to include, and exclude, certain training data based on either quality or appropriateness must be made somewhere and such a decision cannot be value-free. Certain biases are not easy for humans to detect, and there is an inherent division between the goals of Fair AI and Explainable AI [S21]. Fair AI is interested in the outcome being fair and unbiased, while Explainable AI focuses on the fairness within the procedures of creating an AI application. These two goals are not identical, and both are needed to further the creation of less biased applications.

Suggested mitigation solutions include calls for standardized and validated bias measuring

methods [S65, S67, S85] as well as propositions for new types of metrics [S67, S79, S93, S116]. These proposed new methods suggest moving away from simple classifiers, focusing on documentation, explainability, continuous evaluation, and context of the applications.

While bias mitigation attempts should not be abandoned as futile, the question of their end-goal remains unanswered. Is it even possible to create an "unbiased" model, or a certificate for such? Biases and appropriate language are not universal metrics and are always connected to culture [S116], so a single measure cannot be used globally and indiscriminately between different applications. Context matters. Or put another way: "social bias tests cannot be the panacea for language models problems" [S85].

4.2.5 Data leakage or unintended memorization

This issue was identified in studies [S21, S39, S40, S41, S44, S45, S55, S56, S68, S81, S86, S87, S88, S89, S97, S99, S115]. It deals with the well documented challenge that LLMs can often memorize their training data verbatim and can repeat this data back at inference time. This leads to further issues when the training data contains information that can be classified as private. This private data can be accidentally spilled, or purposefully extracted. Because training data repetition is not an intended result of LLM design, this issue has been categorized under the requirement of technical robustness and safety.

Several of the reviewed studies empirically demonstrate data leakage in existing language models [S40, S87, S88, S89]. These demonstrations show that training data contents can be extracted even with black box access to the model, showing that at least current methods for protection are not robust enough to ensure that the models function correctly in this regard. Defensive techniques are criticized in studies [S55, S88, S89], noting that the scale of LLMs works against them, since the possible attack space is so large.

Memorization of data is observed especially for unique or rare data items, but interestingly, memorization has been shown to increase with growing model capacity [S41]. Whether this is a universal truth remains to be seen. It is however clear that a larger model that is vulnerable to extraction will compromise larger amounts of data upon failure.

Another challenge that is pointed out is that LLMs might be used in the future to extract and combine memorized data from other models, revealing harmful information such as military or business secrets [S97]. It is also noted that LLM training data might include benchmark questions and answers designed to test for ethical issues. In such cases the LLM might just repeat the memorized answers and thus distort the results.

The solutions for this issue are among the more studied ones. This can most likely be attributed to the partially technical nature of the issue. The concept of differential privacy, which means making training data mathematically private, is discussed in more detail in Chapter 4.3.

Some studies propose different techniques to reduce unintended memorization [S41, S68, S87, S115], but due to this review's exclusion criteria, most studies that just mechanically demonstrate partial solutions without discussing the ethical implications thereof, were not included. What is noteworthy about these proposed methods is that they do not claim to solve the issue, merely to improve the level of privacy, or reduce the computational cost or data degradation as compared to earlier methods.

Other studies call for improved designs and techniques to combat this issue [S88, S99, S115]. These studies also highlight the need to focus on the LLM's context. It is assumed that no single memorization prevention method can serve all intents and purposes. This seems logical since different application areas require varying degrees of understanding. For example, a medical application to help with diagnoses will certainly need to combine and use more unmasked private information than a bookstore customer service chat-bot.

4.3 Privacy and data governance

This requirement covers issues relating to the use of private data in the development of AI applications (EU High-Level Expert Group on AI, 2019). It encompasses both data gathered about individuals as well as whatever information is generated of them by applications. One of the most relevant connections of this requirement and LLMs is that of data quality. As language models use vast amounts of data, their success depends on the quality and integrity of these materials. Naturally, the handling and governance of data becomes more relevant, and difficult, as the scale increases.

The distribution of identified issue amounts is detailed in Figure 4.4. The ethical issues identified to connect with this requirement are the following:

- **Data gathered without consent** - ethics of the gathered training data
- **Privacy and data security** - concerns of privacy and data security of LLMs
- **Data management** - challenges with managing the vast amounts of data LLMs deal with

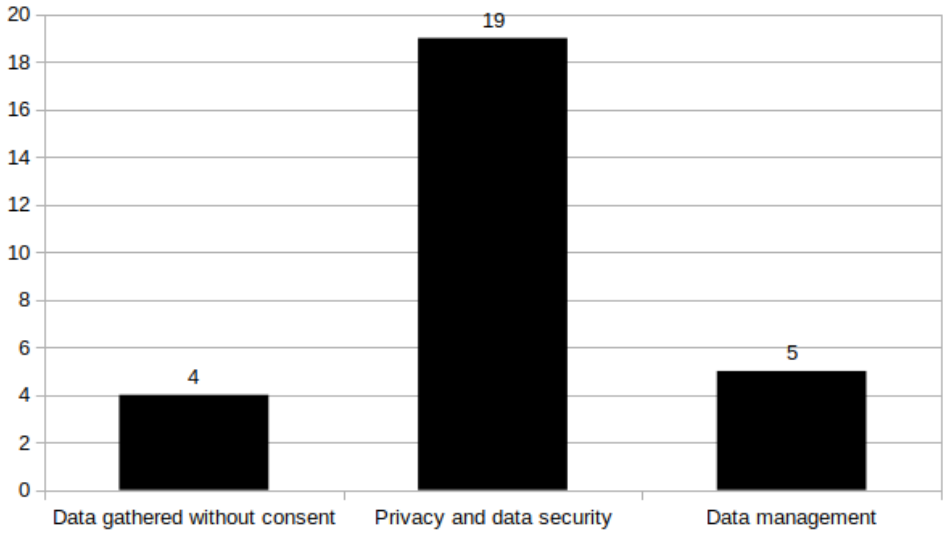

Figure 4.4: Identified issues relating to privacy and data governance.

Differential privacy. This literature review identified several studies concerning different techniques to achieving and improving differential privacy, but many of them did not discuss ethical issues on a more substantial level than admitting that such exist and the obvious issue of privacy preservation. When a study regarding differential privacy clearly included prominent ethical observations or more unique observations, it was included in this review. This subsection gives a brief explanation on differential privacy from the point of view of this review. The comparison of methods found in differential privacy studies, their effectiveness, completeness and plausibility, is a topic of study that merits its own dedicated work.

Differential privacy refers to making data mathematically safe from leaking identifiable information about members of the dataset (Dwork, 2008). Naturally this concept is of great interest to LLM developers, since the models are known to be at risk of unintentionally leaking training data, which can include private information [S21, S39, S40, S41, S44, S45, S55, S56, S68, S81, S86, S87, S88, S89, S97, S99, S115].

When an LLM is not differentially private, a membership inference attack can be executed to retrieve parts of the original data, or personally identifiable information (Behnia et al., 2022). In practice the level of differential privacy achieved in LLMs is on a cryptographically strong basis at best - the goal being that extraction is extremely difficult and unlikely, but not impossible. This ties into the fact that making the training data private in a robust manner is extremely expensive [S45, S56, S68, S89]. This expense can mean

both increased compute resources used and human work in annotating and checking the training data.

Differential privacy should not be considered a solution to privacy issues that is just waiting for a sufficiently elegant application. Executing successful differential privacy can even work to mask biases inherent in the training data, since they cannot be identified from private models or datasets [S68].

4.3.1 Data gathered without consent

This issue was identified in studies [S1, S2, S44, S81]. It was considered as a separate issue from the following one, which deals with more general matters. The issue is so fundamental to contemporary LLMs that it merits its own topic.

Current LLMs are by and large trained on massive datasets collected from the internet - often the Common Crawl -material, a collection of vast amounts of internet content [S111]. However, just because private data is available online does not mean that it is (ethically) free to be used in training of AI applications. A clear illustration of this is that the data might not have been put online by the person to whom it belongs but could rather be the result of a data breach [S44].

Even if data containing personally identifiable information is available in public documents, collecting, spreading, and making it accessible is not guaranteed to not cause harm [S81]. This availability can reduce human autonomy along with privacy, and when applied can allow for surveillance on a scale never seen before [S2].

The clear solution for this issue would be to use training data that only contains information that all of its creators have consented to be used for such a purpose [S44]. This, however, is found to be extremely challenging in practice. Such limitations would reduce the dataset sizes significantly, which in turn would reduce the performance of trained models.

4.3.2 Privacy and data security

This issue was identified in studies [S9, S11, S12, S13, S23, S26, S30, S47, S49, S54, S56, S57, S67, S68, S72, S89, S108, S109, S110], being one of the most prominent issues recognized in the whole review. This is natural given its general nature, but still indicative of privacy and data security issues being a prime worry regarding LLMs.

As LLMs become more common in education, whether as teaching or cheating tools, they will be accessed in increasing amounts by children. While this is naturally a concern about LLM output safety, it is also relevant to consider what kind of data is being collected from users [S9, S110]. Children might be prone to telling an AI application more personal information than is in their best interest - although this has been discovered to be an issue with adults as well - especially when the AI is anthropomorphized [S97].

Different methods to make the training data, and thus the model, more private exist but these are not without their criticisms. Automated de-identification is being used to make datasets private, but this method has been observed to create errors because of imprecise content [S47]. As long as perfect differential privacy cannot be achieved, de-identification or sanitation methods cannot make datasets guaranteed to be private [S56]. This should be assumed to mean that they are not private.

If private learning was achieved in a satisfying manner, it is also seen as an issue. Knowing that the data cannot be leaked by their models might motivate companies to collect data from users even more aggressively [S68]. This has many inherent harms to privacy and human agency and could enable increased mass surveillance [S97].

There are many different mitigation techniques suggested for this issue. Pseudonymization, or other methods of de-identification, of training data are suggested, but are also noted to be problematic due to performance degradation [S11, S45, S47, S56, S97].

Another logical avenue of mitigation is to implement LLMs with more care. This can mean using humans in the loop as much as possible [S30] or sourcing the training data with more care [S44, S97]. It is noted that much of personally identifiable information comes from sources that are known to contain sensitive content [S56] - meaning that compromising privacy is a rational choice by the LLM developers.

An interesting suggestion is the use of synthetic, generated, data to train models [S47]. While this clearly would make the training data private, it comes with many other challenges. The problems of AI-generated material populating knowledge bases are currently seen as crucial, they are discussed in more detail in Section 5.1. Additionally, any generative model that would produce synthetic, but realistic, private data would have to be trained on actual private data. This loops back to the issues of data leakage - what guarantee is there that the synthetic data producing model is not repeating actual personally identifiable information?

4.3.3 Data management

This issue was identified in studies [S39, S82, S87, S108, S111]. It is very closely related to the previous issues of data gathering and general privacy and data challenges. Because LLMs by their very definition handle massive data volumes, managing this data becomes highly challenging. The scale of the datasets makes traditional methods of management non-effective [S111]. This issue is relevant to both legal and ethical data management [S39]. It is understandable, if not acceptable, that one cannot easily prove a corpus of hundreds of gigabytes or larger does not contain some data that should not be handled or is not being handled improperly.

To mitigate this issue, some studies suggest or call for increased governance [S39, S97]. Just applying more of current methods is not seen as enough though. Since LLMs pose unprecedented problems, new governance structures are required. Other studies suggest that models and datasets should be improved; models should be made more easily auditable [S56], and datasets could come with datasheets that detail what the data contains, how it was made, and how it is intended to be used [S71].

4.4 Transparency

This requirement covers aspects that make AI applications more reliable and understandable to their users: traceability, explainability, and communication (EU High-Level Expert Group on AI, 2019). Traceability and explainability mean that the data and processes that yield an application's output can be followed and understood. Explainability also encompasses declarations of trade-offs made in the development process. Communication refers to AI applications representing themselves. A properly communicating AI application makes its nature clear, declares its purposes and limitations, and explains how a user might opt out of using the AI in favor of human interaction.

The distribution of identified issue amounts is detailed in Figure 4.5. The ethical issues identified to connect with this requirement are the following:

- **LLM as an author** - ambiguity of having an LLM be a co-contributor in scientific publishing
- **Academic integrity or source tracking** - LLMs hallucinating references or their output being unverifiable

- **Unfair decision making** - fairness issues that might not be directly related to discrimination
- **Lack of transparency** - transparency of LLMs, or lack thereof
- **Copyright infringement** - LLM-produced content being in violation of copyright
- **Unfactual training data** - LLM training data being of poor quality and becoming outdated

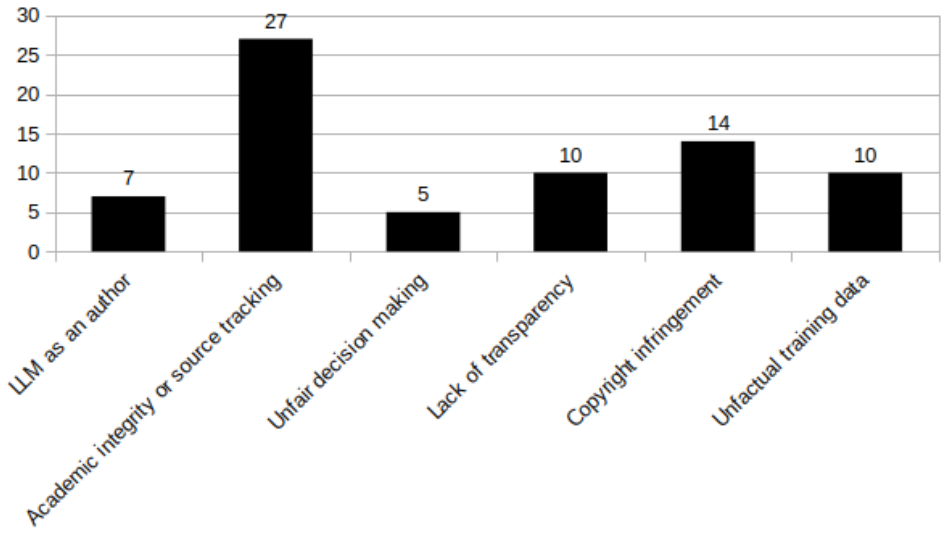

Figure 4.5: Identified issues relating to transparency.

4.4.1 LLM as an author

This issue was identified in studies [S1, S2, S24, S27, S52, S70, S94]. It deals with the current concerns that some research is being published with an author credit attributed to LLMs. The identified studies discuss this specifically in an academic context, as fictional or entertainment products created with LLMs are more closely related to copyright issues.

The issue is clearly defined by Anderson S.S. [1]: "An attribution of authorship carries with it accountability for the work, which cannot be effectively applied to LLMs". There is currently no way to make a language model accountable for its mistakes, so it should not be receiving credit for its successes either. And even if this was possible, to whom does the credit, or culpability, belong: the person who prompted the LLM, the people who

created the LLM, the enterprise that owns the model, or the multitude of creators whose content made up the training data [S94]?

In the future, generative AI might become the driving force for content creation and lead to research references being matched with the output [S52]. This challenge is presented in the context of education, where students might generate essays with LLMs and then enrich them with references from curriculum, but there is no reason to assume that the same worry would not apply to the scientific method. If a researcher produces their content first and then looks for references to support it, clearly, they are not approaching the research with an open mind.

The different mitigation methods for this issue are similar to those of the next issue. They are detailed after the results on academic integrity and source tracking.

4.4.2 Academic integrity or source tracking

This issue was identified in studies [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]. It concerns the accruing problem of research papers being generated with the help of, or sole contribution of, LLMs. The produced content is seemingly of high quality, but there is no guarantee that the details or references are correct. LLMs are shown to hallucinate - that is, to invent - reference citations when it seems one should appear in text.

The implications of this issue are troubling. The understandable worry that freely accessible research papers might be taken for LLM training material is causing some publishers to consider withdrawing their content from being publicly accessible [S24]. This solution, however, is likely to increase inequality of access and has its own ethical implications.

There are calls for improved authenticity checks to be developed and implemented [S9, S20, S24, S43, S44, S60, S62, S100]. The cost of automating such checks must be carefully weighed. Since LLMs can generate content of excellent quality, authenticity checks are likely to cause a number of false positives - results that could cost honest researchers their careers [S62]. It is crucial that there is always a well-versed human in the loop when making decisions regarding academic dishonesty.

In addition to authenticity evaluation, clear rules and guidelines are clearly necessary [S6, S60, S63, S98, S100, S110]. After all, academic integrity has always relied on most of the community to be honest about their work. As the effects of LLMs on academic publishing are becoming clearer and more pressing, more discussion and definitions are called for

to at least prevent researchers from unknowingly conducting their work in an unethical manner [S12, S100].

4.4.3 Unfair decision making

This issue was identified in studies [S12, S58, S69, S81, S109]. It lies in the intersection of the issues of self-acting AI, inaccurate results, alignment, and biases. It is mapped to the requirement of transparency because the studies that discuss this issue deal with LLMs producing improper results, the basis of which cannot be easily traced.

Many of the challenges related to this issue include LLMs being trained on problematic training data or goals. For example, when training an AI to prioritize who receives healthcare, there is a clear value judgment to be made as to what attributes make one a prominent candidate - be it their amount of suffering or the financial value they provide the society or the healthcare service [S12].

The above example can occur both intentionally and unintentionally. What is more often attributed to unintentional harms are allocational ones [S58, S69, S81]. This is caused by the data used in the training of LLMs already containing biases that are then propagated in its results. Such issues can occur for example when a recruiting AI is trained on earlier recruitment data that favors male candidates over women. In such a case the AI will learn this same prejudice and repeat it. Allocational harms can be difficult to detect since they tend to repeat injustices that are considered normal by the majority.

The old truth that machine learning applications are at best as good as their training data is ever relevant. The issues of unfair decision making are mitigated with the same techniques that strive to remove biases from training data and the resulting applications. These are presented in Section 4.5.1.

4.4.4 Lack of transparency

This issue was identified in studies [S14, S26, S53, S65, S66, S78, S81, S95, S100, S101]. It covers challenges concerning the different ethical problems that arise from LLMs being by and large black-box models.

It is clear that for commercial LLM users the models are black boxes. This can lead to users generating "algorithmic folk theories" - attributing the model with too much authority and not understanding its limits and functions [S66]. This same opacity can

also create a sort of dependence on the largest LLM developers [S101]. When one cannot understand how and why an LLM produces the answers it does, research on them becomes increasingly difficult, unless one belongs to the developing organization. This also hinders new companies from entering the field, as creating LLMs of comparable performance becomes exceedingly difficult.

The challenges with lack of transparency are exacerbated by current models suffering from documentation debt [S81]. This has the potential to lead to situations where not even the developers of an LLM can easily understand how certain results are inferred. Another transparency challenge that is seen in LLMs is that the models might learn biases that are not even observable in the training data [S95]. These challenges can be mitigated by focusing on documentation and making more of the inner workings of the models understandable to the public, or at least academic circles. Active analysis of the function of models during production is also called for

An interesting result was achieved by Longoni et al. [S78]. The study shows that people believe news content much less if it is disclosed that the content was produced by an AI. This seems a natural result given the lack of transparency inherent in AI solutions. The authors however draw a different conclusion, seeing the problem being that the disclosing the use of generative AI is prone to lower public trust in news sources, and this problem could be averted by not disclosing such information. They note that this aversion can shift once the population starts seeing AI generated news as a norm. This conclusion effectively illustrates that not everyone considers lack of transparency as an issue - but rather a feature.

4.4.5 Copyright infringement

This issue was identified in studies [S2, S19, S24, S26, S37, S39, S44, S50, S52, S72, S90, S98, S103, S104]. It is closely related to the issue that training data is often sourced without the authors' or owners' consent. With this issue the focus is more on creative ownership rather than personal data. The issue is mapped to Transparency because LLMs tend to obscure the content used to achieve their results which means no one can know explicitly if their own creations are used against their will.

As LLMs do not think, they cannot consider different issues, such as legal or ethical, related to their training data or output [S2]. The output is simply a probabilistic synthesis of their training data. It is also clear that current LLMs do not credit the creators whose

work their output is based on.

Mitigating this issue would help with many explainability and transparency challenges identified. If any output of an LLM could be traced to its constituent parts, it would be easy to credit the original creators. This, however, does not seem possible with current LLMs as they do not work by randomly selecting phrases from all training data and queuing them. There is also the argument that commercial creators of LLMs do not necessarily want this to happen, as this would open up the discussion for them having to compensate creators for the use of their work; something that would be in accordance with current copyright laws but is not necessarily conducive to profitable business.

4.4.6 Unfactual training data

This issue was identified in studies [S13, S23, S42, S49, S60, S67, S91, S97, S100, S110]. It covers the challenges caused by LLMs being trained on factually incorrect data. The clear problem is that a model trained on faulty data will produce inaccurate results, which are discussed in Section 4.2.1.

Training a model based on unfactual data can lead to difficult questions as to who is responsible for such inaccuracies [S23]. The developers of the models are most likely to see that the data is what it is, and state that they make no claims that the generated content is accurate. It is understandable that developers cannot be held responsible for ensuring that all data used is accurate, as most information is not objective fact. Even the data that is fact is not guaranteed to remain so, as new information is constantly discovered, and old content becomes outdated [S91]. This is a significant challenge for LLMs as most of them are trained on static data and they are not actively updated as information evolves, or a society's values change.

4.5 Diversity, non-discrimination and fairness

This requirement is rather well defined by its title. A trustworthy AI application needs to be fair and respect all people in an equal manner (EU High-Level Expert Group on AI, 2019). This relates to people being from different cultures, beliefs, orientations, and societal statuses to name a few. AI applications should be usable by all, and different stakeholders need to be considered - preferably included - in the development process.

The distribution of identified issue amounts is detailed in Figure 4.6. The ethical issues

identified to connect with this requirement are the following:

- **Biased training data or outputs** - concerns of biases regarding LLMs
- **Discriminatory results** - documented discrimination or practical concerns about biases
- **Lack of global definition for bias or fairness** - applying fairness in different contexts
- **Non-binary gender neglected** - how bias tests only reflect on binary gender
- **Toxic content** - toxic outputs of LLMs, e.g. sexism, racism, homophobia
- **Promotes inequality** - how LLMs offer possibilities and tools in an unequal manner

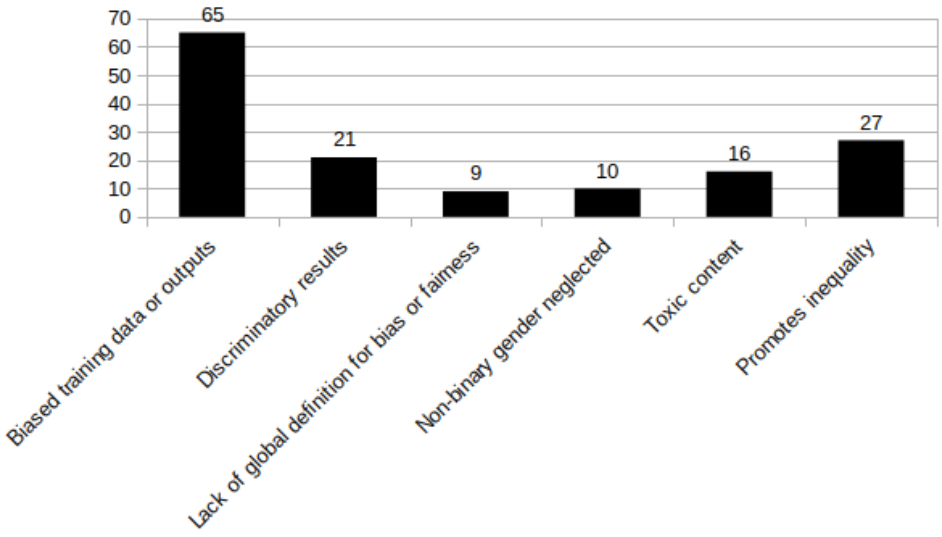

Figure 4.6: Identified issues relating to diversity, non-discrimination and fairness.

4.5.1 Biased training data or outputs

This issue was by far the most identified one in the review, being identified in 65 studies [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]. The amount of discussion

regarding this issue indicates that it is both a very well-known and a worrisome one. The fact that LLMs contain biases is so well known does contribute to the large number of studies identifying this issue - some merely reference it in passing.

There are vast amounts of different types of biases known to occur with LLMs, or machine learning solutions in general. A list of 10 commonly mentioned biases is: gender, age, sexual orientation, physical appearance, disability, ethnicity, socioeconomic, religion, culture, and cross-sectional [S16, S105]. A more conceptual list is discussed in [S51]: overconfidence, recency, majority label, and common token bias. These biases are built into practically all content that can form the corpora for a language model, such as Wikipedia [S71]. While Wikipedia does supply metadata that offers some insight into the issue, other data collections do not.

Biased outputs, especially when it comes to gender bias, are problematic to define. A study [S15] of (the now rather outdated) GPT-2 discovered that the model actually portrays less biases than could be assumed based on census data. For example, the language model identified truck drivers as men rather than women with smaller frequency than should be expected based on actual gender distribution within the profession. This illustrates that defining an unfair bias is not as easy as one might initially assume.

A more easily observable form of gender bias is brilliance bias [S18]. This means that LLMs attribute different qualities depending on if they generate content about men or women. Typically, this means that women's capabilities are downplayed, or they are given qualities some might consider feminine. For example, men might be "brilliant" while women would be "compassionate" in the same context. Similar types of stereotypical attributions of characteristics were demonstrated in study [S73].

One common suggestion to mitigate biases is to manage the training data better [S16, S35, S42, S46, S72, S105]. This includes suggestions for both manual annotation and technological innovation. The problem with manual annotation is that the corpora used for LLMs are so vast, there is no reasonable amount of people and time to actually go through the whole dataset [S16]. It is also challenging, or impossible, to have perfectly neutral bias-free human annotators [S2]. Another possible mitigation method is to use training corpora that is pre-curated in an acceptable manner [S93, S94, S113].

On the scale of society, there is a call to make the NLP community more diverse [S16, S42] to ensure bias and its contexts are understood better. It is also suggested that the developers of LLMs might be made more responsible for the biases and content their products exhibit [S48, S116]. The implementation of such legislation would be an interesting legal

issue and seems currently difficult to attain. Were it achieved though, it would certainly force more resources to de-biasing AI solutions - with the downside of slowing development and making entry into the field difficult. This could in turn give a competitive advantage to developers in countries without such legislation.

There are weaknesses identified with current bias reduction methods. Knowledge distillation, the process of transferring a model's knowledge to another, is shown to decrease fairness [S17]. Scaling models up is not a solution, since larger models can exhibit even more gender bias than smaller ones [S93]. Reduction of bias prior to fine-tuning does not mitigate biased behavior post fine-tuning [S109]. On the other hand, fine-tuning has shown good results when compared to prompt engineering [S72]. These conflicting results indicate that the whole concept of bias is not well understood, and studies tend to focus on narrow points of view. Several studies do call for wider range of testing bias [S73, S95, S105, S113]

Some studies observe that biased LLMs are not always inherently bad. Such tools can for example aid creators to model their audience [S73]. When used to create fictional content, immorality might be a desired function of a writing tool [S50]. As the study discusses, this could be relevant if video games employ natural language generation - non-player characters need to have varying degrees of biases and morality within the game's story, and some need to be villains. This illustrates that the question of a model's intended purpose is extremely important for its use. However, the existence of purposely immoral LLMs can be considered a serious issue, since they could be used in an unintended manner. Possibilities that LLMs offer for immoral uses are discussed in section 4.6.4.

4.5.2 Discriminatory results

This issue was identified in studies [S2, S5, S12, S17, S30, S46, S49, S58, S72, S73, S77, S80, S81, S84, S89, S93, S97, S98, S104, S109, S116]. It is closely related to other bias issues, unfair decision making, and self-acting AI. This issue was included to demonstrate clear cases and threats of discrimination.

Empirical demonstration of biased and discriminatory behavior of LLMs currently being used is shown in studies [S46, S72, S73, S77, S80, S84, S109], the other studies mapped to this issue discuss it on a more conceptual level. A majority of the discussion focuses on gender discrimination (e.g. [S46, S72]) or marginalized demographic groups (e.g. [S58, S84]).

A less often observed type of discrimination is that of geographic nature [S77]. The study shows that GPT-2 has significant bias against people from countries with less internet access than ones where internet access is common. This is a logical result of most LLMs being trained on corpora sourced from the internet. The unfortunate implication however is that the models will then mimic worldviews and prejudices from certain countries. Since the countries with fewer internet users can at best hope to attain similar user amounts as other countries, it is difficult to see how their representation could catch up. This in turn could enforce a sort of cultural hegemony in LLMs.

Many solutions that mitigate biases in general help with reducing discrimination in LLMs. Detecting discrimination can be difficult for people who do not belong to the disparaged groups, so having multicultural development teams is especially important [S58], along with continuous evaluation of the models' performance [S77]. Technical solutions for monitoring and testing this include adversarial triggers and prompt engineering, which can help detect and reduce instances of bias or discrimination [S77, S83].

4.5.3 Lack of global definition for bias or fairness

This issue was identified in studies [S9, S21, S58, S60, S65, S84, S91, S93, S116]. It deals with the philosophical, and demonstrated, challenge that bias is exceedingly difficult to define, identify, and fix. The issue is closely related to other bias and fairness issues but was mapped as a separate one because of the fundamental difficulty of bias definition.

AI applications, LLMs included, are not free from their developers' prejudices and are naturally defined by their training data [S60]. This can result in applications that contain inaccurate, incomplete, or outdated information that the developers do not know to question. Undesirable biases might not be causal or immediately observable [S21], which also means that solutions and explanations cannot be easily generalized.

Currently the ethics most reflected in the development of LLMs are those of western, white, populations [S9, S116]. These cultures also produce much of their content in English. This can prove challenging for those trying to create LLMs in other languages, or cultures [S58]. The amount of training material, development, and research that is focused on these Anglo-centric LLMs works to guarantee that other cultures find it difficult to represent their own values in models.

Attempts to improve fairness and reduce bias can in fact increase the marginalization of certain groups. Applying various categories to the training data is guaranteed to exclude

groups that do not fit into the chosen categorization [S84]. LLMs are, by design, prone to ignore the least encountered data elements in favor of more likely ones. This inequality can be amplified by marginalized groups using distinctive styles of communication. LLMs change their outputs based on minor changes in the input [S65]. This means that a model can contain several worldviews that are only exhibited based on the user.

As this issue is very theoretical, there are not many practical suggestions as to its mitigation. However, LLM designers can try to avoid some clearly problematic methods. One thing that should definitely be avoided is using directly translated datasets [S58]. Such methods are prone to increase abstraction, lose nuance, and encapsulate biases within the original data.

4.5.4 Non-binary gender neglected

This issue was identified in studies [S72, S73, S74, S75, S81, S84, S93, S109, S113, S116]. The fairly large amount of studies detailing it speaks to the issue being recognized, but there are not many solutions suggested. One of the main contributing factors to this issue is that the study of LLMs and their biases are a relatively new field. Because of the constraints of language, it is much simpler to study a model with a binary designation of gender.

As LLMs are becoming more widely used, this issue will grow to be more significant. There clearly exists problematic implications when people are forced to interact with models that do not understand or respect their identity. If the definition of this issue is broadened to include LGBTQIA+, there are also more practical harms demonstrated [S75]. The study shows that LLM produced content when focusing on these minorities can be harmful, i.e. toxic, over 10 % of the time.

4.5.5 Toxic content

This issue was identified in studies [S2, S3, S22, S30, S48, S54, S67, S74, S75, S81, S82, S89, S91, S101, S109, S111]. It refers to toxic content produced and contained by LLMs. With this issue, more than any other, the problematic nature of extracting training corpora from the internet is illustrated. All the studies that identify this issue demonstrate LLMs' capacity to produce toxic content in practical terms. Some as proof-of-concept (e.g. [S2, S48]) and others in a more empirical manner (e.g. [S74, S75]). One survey detailing

different harms that LLMs can cause is also included [S67].

The sourcing of training data is subject to criticism. For example, the data used to train LLMs often includes content from banned and quarantined subreddit-sites [S54]. It is also noted that toxicity is not limited to certain sites and can exist in more innocuous contexts as well. An analysis of the often-used corpora Common Crawl, a collection of vast amounts of internet content, showed that it contains much toxic content even after applying filtering techniques [S111].

As with bias measurement tools, there are issues with toxicity scoring [S22]. Existing LLMs can be optimized to score well in automated measures, but still struggle with toxic content. This again illustrates that the definitions of these issues are exceedingly difficult and thus cannot be simply measured.

One technique often employed to reduce toxic outputs is identifying policy violating prompts and blocking them. This technique is not, at least currently, foolproof and such policies can be easily circumvented [S48]. Another inference-time method suggested is generating test content, possibly with another LLM, to probe the model being tested for undesirable outputs [S89]. This method still needs to be paired with a process to identify the problematic outputs, so it can only be expected to mitigate the issue, not fix it.

Reduction of toxicity can be achieved by improving the training data. Filtering the data can be helpful, but the method is still vulnerable to specific prompts [S91]. The problems of data filtering have also been detailed earlier in this review. Another method suggested is using smaller and properly curated datasets [S91]. This can be expected to be an effective method, but it will make LLM development much more expensive and reduce their scale, which is one of the base reasons for LLMs' good performance.

4.5.6 Promotes inequality

This issue was identified in studies [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]. As biases and racism are mapped to their own ones, this issue is more focused on social inequality. The most common concern regarding this issue is that premium versions of LLMs perform much better than cheaper, or free, ones. This can lead to widening the gap between privileged and poorer groups, an issue that is especially relevant for education [S7]. Those who can afford to use better tools gain an advantage both in AI study assistants and cheating tools.

Inequality of LLMs is often observed in poorer performance of non-English language models [S67, S91, S113]. This leaves certain countries and immigrants in a disadvantageous position. Be it due to language skills or other marginalizing factors, the use of unequal LLMs can cause real harm in society. If LLMs are used as tools for - or in - recruitment, education, or healthcare, the people a model does not understand suffer significant harms and disadvantage [S93]. These are referred to as allocational harms [S105] and can also include more indirect injustices such as societal resource sharing or decision making, such as prison recidivism prediction. Representational harms on the other hand include many of the issues detailed in this review, such as toxic content, discrimination, and prejudice. These harms can occur both accidentally and on purpose, by government or corporate actors - an example being a program for application handling that automatically prefers certain groups over other [S12].

There exists a degree of environmental inequality with LLMs as well [S81]. While it is western, English speaking, countries that reap the most benefits from LLM development, their environmental and social costs are reflected in countries that do not benefit from the products. These issues are presented in more detail in Section 4.6.

The balance of power for LLM development is not spread evenly in societies, with those in powerful social positions being able to generate more of the content that is used in LLM development [S113]. This can also mean that training data consists of an over-represented loud minority which can propagate attitudes like white supremacy, misogyny, or ageism [S77].

There are not too many direct mitigation methods offered for this specific issue. Although it is closely tied to other issues of fairness, so many of the previous suggestions are relevant for inequality mitigation in LLMs as well.

4.6 Societal and environmental well-being

The requirement of societal and environmental well-being states that AI applications should be designed in a sustainable manner (EU High-Level Expert Group on AI, 2019). This means the environmental cost of the whole life-cycle and supply chain of the application as well as socially conscious designs. The requirement also highlights the gravity of using AI in social and democratic contexts. Of all the requirements, this one most stresses the fact that AI can be incredibly beneficial for society but, implemented poorly, can be detrimental also.

The distribution of identified issue amounts is detailed in Figure 4.7. The ethical issues identified to connect with this requirement are the following:

- **Loss of social skills** - how an abundance of AI solutions in education can reduce human contact
- **Reduced value of education** - over-reliance on technology, degrees attained without proper knowledge
- **Paper or credential generation** - inflated number of papers, contributing to the "Matthew effect"
- **Purposeful toxic or immoral content** - unethical use of LLMs, e.g., misinformation or (cyber)crime
- **Job loss or class divide** - widespread use of LLMs leading to job loss and societal shift
- **Training data pruning** - the human cost of de-toxifying training data
- **Environmental impacts** - environmental cost of training and using LLMs
- **Fairwashing** - how certificates could be used to polish company image
- **Replacement of traditional learning** - learning contexts becoming blurred, ethical issues with automatic grading

4.6.1 Loss of social skills

This issue was identified in studies [S26, S30]. Although not explicitly identified very often, similar sentiments were present in many reviewed studies, especially in the context of education. Since AI's impact on social agency is especially mentioned in the EU guidelines, this issue was mapped as its own. The mentioned challenges regarding this issue are that LLMs might reduce the amount social contact needed or required for ordinary functions. This in turn could lead to a decreased development of social skills required for everyday life.

This being a very conceptual issue, there are no mitigation methods presented. Naturally, we must be vigilant with the observed effects of LLMs and react to them when unwanted phenomena are encountered.

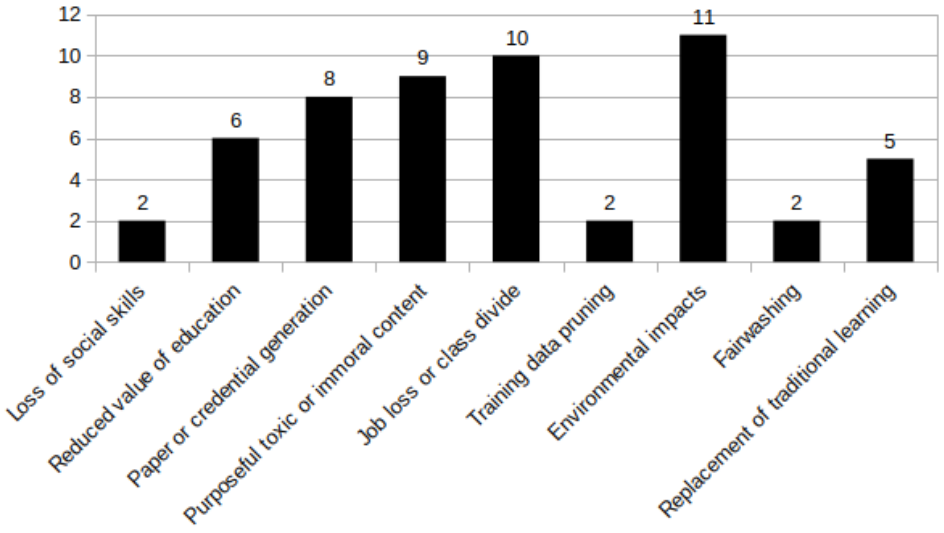

Figure 4.7: Identified issues relating to societal and environmental well-being

4.6.2 Reduced value of education

This issue was identified in studies [S29, S30, S43, S60, S108, S110]. It covers challenges LLMs might cause in education. These include the worry that students, or people in general, will become too reliant on this new technology and start to feel that learning things for themselves is no longer relevant. Another challenge is that the use of LLMs to pass educational tests might not properly prepare students for future careers [S29].

An important mitigation solution for this issue is to teach AI literacy [S108]. As AI solutions become more ubiquitous in society, people need to understand how, why, and when they are used and what might be the implications of using - or being subjected to - them.

4.6.3 Paper or credential generation

This issue was identified in studies [S20, S24, S25, S35, S42, S60, S62, S81]. It is closely related to other issues of education and academia but focuses on the latter more. As LLMs enable people to create superficially viable research content, this can lead to an inflated number of papers and other academic content being produced.

Academic success is currently strongly correlated with number of publications and citations [S20, S24]. This, popular researchers becoming more popular and unknown ones remaining

in obscurity, is a manifestation of the "Matthew Effect". This state of affairs can enable the more unethical members of the field to generate massive amounts of self-referential studies that do not necessarily contribute anything new scientifically.

The possible mitigation solutions currently amount to having clear guidelines and demanding disclosure whenever a contribution is created with the help of an LLM [S35] and developing improved technologies to identify their use [S35, S43, S60]. The need for new automated solutions is already demonstrated, as a 2023 study showed that human reviewers could identify less than two-thirds of paper abstracts written by ChatGPT - an application that is fairly far from the currently top-performing LLMs [S60].

4.6.4 Purposeful toxic or immoral content

This issue was identified in studies [S50, S74, S92, S97, S99, S103, S104, S107, S109]. It deals with the challenges of LLMs being used for intentionally unethical purposes. Toxic content is discussed in section 4.5 in the context of unintended harmful outputs.

LLMs can enable people without significant programming background to create well-functioning programs. The challenge with this is that these programs can also be different types of malware [S92]. Language models also make social engineering attacks easier to create and propagate, as LLMs are already being deployed as interactive agents in social media and other purposes [S92, S99].

The study of this issue leads to a challenge of its own. When a vulnerability or an unintended function of an LLM is discovered, it can be used for the very purpose being demonstrated - this is known as the dual-use problem [S74]. An example of these negative uses is universal adversarial triggers [S99]. These are inputs that can "trick" LLMs to generate outputs that might not be expected; an assortment of seemingly unrelated words and symbols can cause the model to make a statistical connection with certain types of content and start producing, for example, hate-speech or conspiracy theories seemingly unprompted.

To mitigate the generation of toxic or immoral content, all methods that reduce the base rate of generated toxicity naturally help. The less an LLM generates toxic content, the harder it is for anyone to make use of it. In addition to this, stock-responses are suggested for common or identified inappropriate prompts [S91]. While useful, stock-responses cannot be expected to block all unwanted prompts from working, as demonstrated by universal adversarial triggers - LLMs do not require clear natural language prompts to produce

outputs.

4.6.5 Job loss or class divide

This issue was identified in studies [S2, S3, S9, S12, S13, S30, S37, S49, S97, S104]. It concerns the changes widespread use of LLMs will, or might, have on society. As these applications become ever more fluent at interacting with humans using natural language, and able to produce coherent text on request, many job titles might be in jeopardy.

Although this issue has been discussed for decades, as robotic process automation and other automation tools evolve, the capacity of LLMs is on a scale of its own. The possible job loss concerns groups of professionals that have not been implicated to be under such threat before: professors and programmers, and white-collar workers in general [S3]. Another group of professionals that can be impacted severely is those in creative professions [S37, S97]. As AI applications can create new content nigh instantly, and with little cost to the user, it is understandable that creative economies might feel threatened by the emergence of generative AI.

Naturally, the implications of this issue do not have to mean mass unemployment, but rather professions shifting. However, there is a risk that professions will be divided more clearly into well-paid jobs, such as technology development, and significantly lower paying jobs like LLM content moderation [S97]. The number of workers required for the former are vastly smaller than the latter, which means that LLM-economies might be enforcing class divide.

This review found no mitigation solutions for this issue, as it is largely theoretical at the moment. This naturally does not mean it should be ignored. More public and academic discussion is vital, so societies are prepared to live with LLMs.

4.6.6 Training data pruning

This issue was identified in studies [S2, S52]. It is most closely related to the issue of the cost of AI monitoring, described in section 4.7, but this issue refers more closely to the pre-handling of training data. As has been described in this review, LLMs often source their training material from the internet. This material contains unwanted, inappropriate, content that designers usually do not want portrayed by the finalized model. To combat this, the training data is pruned of the clearly improper content. While AI solutions are

used to assist in this, pruning requires human input to recognize and teach the AI what actually is inappropriate. This means that human annotators must comb through vast amounts of data that originates from the deepest recesses of human thought. Consistent exposure to such materials can be emotionally damaging.

The process of data annotation is often outsourced to developing countries [S2]. This means that this issue ties closely to the one regarding environmental inequality, as the countries suffering from LLM development are not the same ones gaining their benefits. Another relating issue is that biases are not global norms. The results of outsourced annotations should not thus be expected to reflect the cultural values intended. For example, consider an annotator that holds the belief that homosexuality is wrong. They are likely to flag content regarding it in a positive or neutral light as inappropriate, even if this was not the intention of the final LLM development team. The absence of data points is difficult to notice but this type of scenario could easily contribute to lacking representation of non-negative content, which in turn can lead to increased amounts of toxic content.

4.6.7 Environmental impacts

This issue was identified in studies [S4, S17, S32, S42, S52, S71, S72, S81, S97, S101, S106]. Although it is identified fairly often, it is one of the most excluded ones in this review since there are not too many studies investigating the environmental impacts of LLMs specifically. This is the result of LLMs sharing similar environmental impacts with other AI, ML, and NLP applications.

A review on the risks posed by language models [S97] groups the environmental harms caused by language models into four categories:

1. direct impacts from the energy used to train or operate the LM
2. secondary impacts due to emissions from LM-based applications
3. system-level impacts as LM-based applications influence human behavior (e.g. increasing environmental awareness or consumption)
4. resource impacts on precious metals and other materials required to build hardware on which the computations are run, such as data centers, chips, or devices

A major challenge posed is the spread-out nature of these harms. The resource and environmental impacts are felt far away from the users of LLMs [S71]. This nature of the harms is also likely a contributing factor in the often-lacking reporting of their emissions. Few operators publish figures that would allow proper examination of their environmental impact [S106].

This all leads to significant challenges with responsibility regarding the environmental impact of LLMs. As LLMs are gaining popularity, the emissions from their operation will inevitably overshadow those of the training phase. This can mean that the responsibility for energy and hardware use shifts to the end-users [S4]. Another topic of conversation is whether researchers should take any responsibility for technologies they develop, but do not use commercially [S106]. More efficient technologies do not seem to have positive impacts on resource consumption, as newer models grow in scale, requiring more compute [S106].

Environmental costs are also subject to prioritization [S17]. Methods to increase fairness aspects in models also tend to decrease sustainability. Naturally, this means that sustainability can be improved by reducing some fairness methods, making this a wicked problem.

To reduce, or help combat, the environmental harms caused by LLMs, improved reporting is considered vital [S17, S71, S81]. Proper and standardized reporting can help develop more environmentally friendly LLMs and allows customers to choose products that share their values. Other methods that developers can use include developing more efficient solutions [S17, S32] and running their programs in more carbon friendly regions [S81].

Users can also contribute to the reduction of environmental impacts beyond their choice of product. This can be furthered by practitioners informing them, for example, of the most effective ways to use their programs [S4] and electricity recycling [S32]. It should however be maintained that such actions alone do not make a company environmentally friendly and do not reduce the responsibility practitioners have for their products' climate footprints.

4.6.8 Fairwashing

This issue was specifically identified in studies [S84, S106]. The issue is also implied in many others that concern any sort of standardization process for LLMs. The term fairwashing comes from the more often used greenwashing and pinkwashing, which mean

efforts to make a company seem more environmentally or LGBTQIA+ friendly without actually manifesting these qualities in their products and processes. This can be done for marketing reasons or to direct public attention away from problematic issues. Fairwashing can become an issue if LLMs are granted different certificates without reliable means of verifying their compliance to the methods the certificate aims to measure.

The problems of fairwashing are not limited to certificates. Companies can market themselves based on certain chosen positive factors and remain quiet about others. To mitigate this issue, in the context of environmental impacts, there is a suggestion that all companies need to fill out a climate performance model card, which could disclose the relevant sustainability information but still protects the developed intellectual property [S106].

4.6.9 Replacement of traditional learning

This issue was identified in studies [S1, S2, S49, S53, S57]. It is closely related to the issue of reduced value of education. This issue focuses more on the whole education system. The challenges from the students' perspective include loss of learning opportunities, increased inequality and discrimination as detailed in section 4.5.6, and the ease of cheating. LLMs can perpetuate loss of learning if used poorly. This can occur because language models do not necessarily understand what is being taught and can provide irrelevant answers or lead the learner astray.

From the point of view of the educators, the use of LLMs in the replacement of traditional education has many ethical implications, starting with the fundamental one: is it ethical to use AI in teaching? Especially for automated answer grading there is a plethora of challenges, including giving and receiving feedback, trust, transparency, and cheating [S53, S57].

These challenges are multiplied when considering children's education. Young people cannot be expected to understand the limits and risks posed by AI solutions which leaves them in an especially vulnerable situation. Children's tendency to divulge more personal information than might be prudent is discussed in 4.3.2. An LLM grading and feedback tool will also be vulnerable to all the issues of bias, toxicity and inequality discussed earlier, which can amplify existing social imbalances.

As the use of AI tools for grading is still evolving, the use of such tools unmonitored cannot be considered ethical [S53, S57]. The natural remedy for this situation is to have a human in the loop for the foreseeable future and use LLMs as teachers' assistants instead

of their replacement. At the same time, it is crucial to teach critical thinking and AI literacy from an early age [S2] because children will use and encounter these applications almost certainly.

4.7 Accountability

The requirement of accountability stresses that AI applications must be responsible, and that their use and content can be audited properly (EU High-Level Expert Group on AI, 2019). This requirement covers all phases of an application's lifecycle, from development to discontinuation. Along with auditability this requirement includes the principles of minimization and reporting of negative impacts, addressing trade-offs, and having processes for redress when or if something does not go according to plan.

The distribution of identified issue amounts is detailed in Figure 4.8. The ethical issues identified to connect with this requirement are the following:

- **Corporate influence** - large corporations having advantage and power in decisions of future
- **Cost of privacy** - the cost of making sure an LLM and training data are private
- **Cost of AI monitoring** - the cost of retraining and monitoring LLMs
- **Ambiguity of accountability** - who is legally or ethically responsible for LLMs effects

4.7.1 Corporate influence

This issue was identified in studies [S2, S10, S13, S25, S96, S101, S116]. It covers the challenges posed by the fact that the field of LLM development is dominated by a few gigantic operators and they are working largely unregulated. This makes it challenging for smaller practitioners to enter the field and offers the current ones an oligopoly-type of advantage, which these large corporations use - or at least can use - to reinforce their situation.

The ways that the large corporations safeguard their positions in the market include obscuring the training data and program code [S13] and controlling research by publications

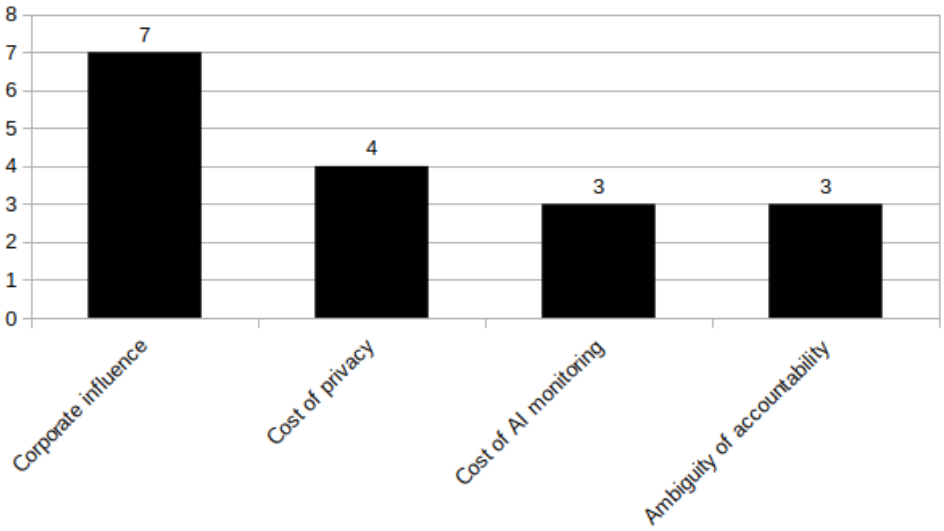

Figure 4.8: Identified issues relating to accountability.

and recruitment [S101, S116]. Preventing access to the data and training methods of LLMs is in violation of the requirements of accountability and transparency, while controlling research can be more ambiguous. Current research focuses on language model technologies that require extremely large amounts of compute, such as the transformer architecture. This means that smaller practitioners cannot afford to participate in the most cutting-edge methods. The largest companies also tend to own the facilities that offer the required cloud computing for these methods. This means that they can gate-guard who gets access to the necessary resources, benefit financially from anyone entering the field, and have instant access to new methods invented by those using their infrastructure. All this combined means that there is a financial incentive for these companies to further compute-heavy technologies and strive to block alternative avenues of research.

The increased popularity and performance of LLMs means that their demand will keep growing, and new application avenues will emerge, both in private and government sectors. This indicates that practitioners who control the development of LLMs will hold increasing amounts of political and economic power [S101]. To mitigate this, legal limitations to this power are suggested. To mitigate the problem of corporate controlled research, more financial support for independent and critical research is called for [S101, S116].

While many studies naturally suggest or expect increased regulation to solve some issues they detail, some studies specifically highlight that the use of LLMs should be regulated more firmly to consider their ethical issues [S2, S10, S23, S25, S96], the avoidance of which

is clearly an ethical issue itself. Increasing regulation could help increase societal discussion regarding LLMs and generative AI applications. Without including more people in the discussion, regulation and development can happen under the radar and it is possible that risks posed to different walks of life are not noted.

4.7.2 Cost of privacy

This issue was identified in studies [S45, S56, S68, S89]. It deals with the fact that when considering trade-offs made in LLM development, privacy is often one that can be dialed down to improve efficiency, or simply profitability. While differentially private model training is plausible, it is computationally extremely expensive and prone to reducing accuracy in the application, as the trained model can fail to capture connections that are obscured by privacy methods.

The massive amounts of training data make human overview of the material practically impossible [S56, S89]. This in turn can lead to a reduced number of test cases and understanding of the data that is being handled. Automated methods for increasing privacy cannot be expected to detect content that is overlooked or unexpected within the data.

Because improving model privacy is so expensive, forgoing it may be tempting when creating a model for internal use [S45]. In these situations, it is important to remember that such models should not be shared or used in situations where data leakage and other privacy problems become more relevant.

There were no especially innovative methods identified for mitigating this issue. Since the problem is that privacy preserving development is expensive, the solution is simple - spend the resources necessary. At the very least, developers should be expected to declare the trade-offs they have made in the creation of their application.

4.7.3 Cost of AI monitoring

This issue was identified in studies [S70, S83, S93]. In a similar vein to the previous issue, it covers the trade-offs required to create a properly working and ethically aligned LLM. After the model is trained, its performance needs to be continually monitored, tested, and amended. This can be extremely expensive and labor intensive. What is noted as especially expensive is re-training of models, which can be necessary for example when a

new form of systemic bias is detected to exist in the model.

Similarly with the previous issue, the mitigation of this one is a matter of prioritization and investment. The issue can be prevented, but it might not make fiscal sense.

4.7.4 Ambiguity of accountability

This issue was identified in studies [S13, S26, S28]. It is related to the principle of auditability in the sense that AI applications' decision-making processes and data should be assessable. The issue is also closely related to many that are discussed in Section 4.4. The primary challenges of ambiguity are that LLMs by their nature do not clearly emulate any specific parts of their training data. This has implications for copyright and academic referencing. One cannot be certain when their copyrighted content is being used without declaration.

This issue is also relevant when considering toxic or other inappropriate content [S13]. If the output created by an LLM, for example a Tweet, breaks a service's user agreement or is deemed illegal, the responsibility is difficult to pinpoint. An argument could be made that the person(s) responsible are the ones who made the content that the model was trained on, or the developers of the model, or the publisher of the model, or the user who operated the instance of the LLM. Because of this, things can get convoluted easily which can in turn lead to an increase of toxic content around the internet.

The clear mitigation method for this issue is enforcement of accountability through regulation. The ambiguity of the issue and the huge financial stakes are sure to make such political debate challenging.

5 Discussion

As LLMs have reached a point where they can be deployed for everyday use in a wide array of applications, their weaknesses need to be considered along with their strengths. The fact that these tools work, and they often work well, means they will be used.

This review identified 39 recurring issue types mapped across all the requirements detailed in the EU guidelines for trustworthy AI (EU High-Level Expert Group on AI, 2019): human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, environmental and societal well-being, and accountability. The identified issues are presented in Table 5.1 and also described with brief explanations in Appendix B.1.

This chapter presents discussion about the weaknesses or threats that LLMs can, and do, exhibit. Section 5.1 presents future research topics that were identified in this review. Section 5.2 contains discussion about other implications regarding LLMs. Sections 5.3 and 5.4 discuss the benefits that can be gained from this review, and possible threats or limitations to its validity and quality.

5.1 Potential research directions

Education. The education sector is one that is presumed to be shaken the most by the popularization of LLMs. This combined with the vulnerability of children using these systems means that it is vital to consistently observe how LLMs affect learning. Because of the rapid introduction rate of LLMs to public use, there is going to be at least one generation of people who will functionally be test subjects whether this is desired or not. LLMs will be used by students as learning aides or as cheating tools - probably both - and there is no previous data to indicate how such powerful tools will affect the learning outcomes.

This all means that consistent research is essential. Results must also be shared between different institutions and countries. The undeniable power of LLMs means that those who use them the best will gain a huge advantage over those who do not. The changes that are foreseen for the education sector are so many that separating specific research directions

Issue name	EU requirement
Loss of learning or the ease of cheating Fake news or misinformation Echo chambers Self-acting AI Influence through suggestions Manipulation	Human agency and oversight
Inaccurate results Dangerous content Alignment Bias scoring and solutions Data leakage or unintended memorization	Technical Robustness and safety
Data gathered without consent Privacy and data security Data management	Privacy and data governance
LLM as an author Academic integrity or source tracking Unfair decision making Lack of transparency Copyright infringement Unfactual training data	Transparency
Biased training data or outputs Leads to discrimination Fairness and biases are not global norms Non-binary gender neglected Sexism, racism, etc. Promotes inequality	Diversity non-discrimination and fairness
Loss of social skills Reduced value of education Paper or crediental generation Purposeful toxic or immoral content Job loss or class divide Training data pruning Environmental impacts Fairwashing Replacement traditional learning	Societal and environmental well-being
Corporate influence Cost of privacy Cost of AI monitoring Ambiguity of accountability	Accountability

Table 5.1: Identified issues and their mappings to the EU requirements

is difficult. The need to develop and use teaching methods that adapt to a world of LLMs is recognized in a wide array of studies [S1, S5, S7, S26, S29, S30, S33, S34, S49, S52, S59, S70, S90, S98, S103, S110].

LLMs are often applauded for their ability to summarize text. This can be seen as especially useful for students but is naturally not limited to education and academia. The emergence of widely used text summarization tools posits an interesting question for future study: could it lead to an equivalent of "search engine optimization"? It would seem a natural development for book, and other content, publishers to consider how their products might be summarized by an LLM. For example, consider a situation where a university professor is looking for a new course book from among many others and asks an AI solution to summarize these books and assess their validity for teaching the level of course being planned. It is clear that there is a competitive advantage to be had for the book that is summarized the best - regardless of the quality of the whole content.

Environmental issues. Most current research studying the emissions of LLMs focuses on the training phase [S97]. An interesting, and ever more relevant, research avenue would be to study the models' energy use and other emissions during inference time and possible retraining. The environmental issues associated with LLMs in combination with the modern strive to integrate them with search engines also poses an important research topic. How much more compute, and thus natural resource, intensive does a single search become when it is combined with NLP?

There is a lot of ongoing research regarding the environmental costs of LLMs, and AI solutions in general. Much of this research focuses on making the tools more efficient. While relevant and called for, such results are often made redundant because improved efficiency is translated to larger models [S81, S106]. More research resources are needed for solutions that improve the sustainability of LLM development and operation without just enabling increased performance. Such resources might include different ways to publicize or report the natural resources used by an LLM, or whole new technologies instead of improvements on the current state-of-art. Some suggested approaches to reducing the environmental impact of LLMs include more specialized - smaller or distilled - models, and increased focus on downstream impacts [S17].

AI generated content. As the internet is starting to include more and more content produced with generative AI solutions, it is becoming crucial to understand how new models will act when trained on - at least partly - AI-generated data. There is currently lots of pressure regarding the identification of LLM-generated content, both watermark-

type of identifications embedded in the text and tools to recognize such content when encountered [S9, S20, S23, S24, S25, S28, S29, S31, S34, S35, S43, S44, S60, S62, S96, S103]. These are both very valuable research subjects, but it would make sense to operate under the assumption that they might not work.

There needs to be more knowledge about how model training responds when parts of the training data are synthetic. Especially interesting is what happens to minority voices in these instances. If LLMs are already neglecting groups that are not well represented in their training data [S2, S5, S12, S17, S30, S46, S49, S58, S69, S72, S73, S77, S80, S81, S84, S89, S93, S97, S98, S104, S109, S116], will this issue become multiplied if future LLMs are trained on content created by the earlier generation?

Automated code generation. Another interesting topic is automatic code generation. There are many studies regarding this, and it is already being used in the industry, schools, and private individuals. Due to a lack of ethical discussion in these studies, many were excluded from this review.

A study reviewing current techniques of using LLM programming assistants could prove valuable. What might be especially relevant would be to study completed projects that used automatic code generation and that have been in production for a longer time. Have the developers found any significant differences in debugging and maintaining these programs as opposed to "traditionally" produced ones?

A worrisome implication of automated code generation is that current LLMs are usually frozen in time. This means that there may be known vulnerabilities lurking in the models' knowledge base, which is the case at least with GitHub Copilot [S14]. While it might give criminal elements tools, proving that widely used LLMs suggest known vulnerabilities - such as outdated or unsupported libraries - would be very important. Better that such weaknesses are shown publicly than allowing them to be discovered by people who seek to abuse them.

Bias reduction. Given that the existence of biases in LLMs was the most identified issue in this review, there is no doubt that it is being studied actively. There are many different topics of study around this topic, from the mechanical to the theoretical.

Some of the technical methods that could be improved or developed to reduce biases in LLMs include creation of curated datasets [S93, S94, S113], using other LLMs to improve existing ones [S35, S46], and other solutions like improving the model fine-tuning or generating better tests [S16, S42].

On the more theoretical side, there is a need to create better bias identifying frameworks and legislation to mandate them [S15, S48, S116]. Even though the creation of a universal bias taxonomy seems impossible, improvements can still be achieved. There is a call for the development of bias validation tools or suites [S65, S67]. Although several studies that were reviewed [S79, S85, S116] do develop and suggest such methods, they also admit that currently no single framework can cover all the issues.

Miscellaneous research suggestions. One interesting future research topic would be to analyze different solutions publicized, or hypothesized, that use LLMs. Such research could observe how much ethical issues are considered in the context of the presented solution. For example, for a study detecting signs of depression by using an LLM, does it consider any ethical implications? Are the implications mentioned overshadowed by the enthusiasm about the solution being studies (Inioluwa et al., 2022)?

More intersectional study of these ethical issues is sorely needed [S15, S105]. It is understandable from a research setup point of view that current research focuses on narrow subjects, but ethical issues do not exist in a vacuum. They often form wicked issues, where mitigating one exacerbates the other. An example of this would be applying differential privacy to training data which naturally increases privacy, but at the same time masks any possible biases inherent in the data [S68].

5.2 Other implications

In a way, large language models are much like children. They do not understand the concept of lying, why it is wrong, and what are its implications. There is no implicit rhyme or reason to why and when they produce untruths. This is normal for a learning intelligence. What is strange is that our society is on its way to allowing these computer-children to aid us in making decisions about ourselves and others, write and read our messages for us, and in general run more aspects of our lives.

The results of this review show that there are both demonstrated and theoretical issues related to LLMs in all requirement areas of the EU ethics guidelines for trustworthy AI (EU High-Level Expert Group on AI, 2019). This means that there should be serious considerations about the use of language models in general. Many of the issues identified concern the concepts of current LLM solutions rather than unethical methods or negligence. For example, the concepts of bias or fairness are something that humanity has not been able to define on a non-subjective level.

Training data. Many of the issues identified are closely related to the training data used. The data contains biases, hate speech, and personal information and it is often unethically gathered without the consent of the original owner or creator of the data. The vast amounts of data also mean that vetting them is, quite literally, humanly impossible. If the training data causes models to behave in an unwanted manner, and the data cannot be curated, there is no clear solution for this. This problem can be approached from two directions: the training data can be curated and handled more ethically, and the unwanted behavior can be accepted. What needs to be decided is how much dataset curation and scaling down of models - essentially slowing of progress - can be demanded and what level of unwanted function can be accepted.

What is a disturbingly rarely discussed issue is the labeling of unethical data, observed specifically only in two studies [S2, S52]. Labeling data means that a handler goes through data associated with an LLM's training or output and defines if it is considered unwanted - for example being racist, too sexual, or too violent. This process must be done by humans - how else could an AI be trained to identify such issues? This means human handlers must be consistently exposed to the darkest recesses of human action and imagination, something that would almost certainly be defined as torture was it done involuntarily. This work is often outsourced to developing countries, in OpenAI's case Kenya (Perrigo, 2023), where people choosing between working and life-threatening poverty cannot be necessarily called "voluntary".

Toxic content. While it can by no means be considered a solved problem, modern mechanisms for reducing toxic content seem to work at least on some level. The vulgarity, which will not be reiterated here, that has been discovered in sentence-completion studies that were published in 2022 on older LLMs such as GPT-2 [S73, S74, S75] is on a whole other level compared to more recent studies. Based on brief trials, most of the prompts that cause GPT-2 to generate toxic content do not have the same effect on ChatGPT.

But the reduction of toxicity and biases is itself prone to biases - developers focus on what they feel (or are told) is most important. While an observable change can be seen between studies done on GPT-2 and GPT-3 with regards to gender and LGBTQIA+ bias, GPT-3 still produces, for example, incredibly offensive remarks about Muslims [S83]. This raises several questions: can we presume that all harmful biases will be dealt with in their own time, and what is the correct order for resolving these? Should Muslims be required to endure biased language models until we have sufficiently "solved" gender bias? Religion does seem to be at a very low priority for LLM developers. One bias-reducing method

was used to reduce gender bias three years earlier than it was ever applied to reducing religious bias [S85].

Ruha Benjamin (2019) eloquently portrays the challenges inherent in LLM training data and their potential for toxicity: *"Feeding AI systems on the world's beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy."*

Human agency. The threat that LLMs might cause people to become lazy and stop learning is a widely discussed one [S1, S5, S7, S26, S28, S29, S30, S31, S33, S34, S42, S49, S59, S63, S70, S90, S98, S100, S103, S108, S110]. These concerns cannot be ignored, but their gravity is up for debate. A very similar discussion has occurred regarding for example use of the internet in general, and more specifically Wikipedia, in education. One might also consider that there were serious fears that people would lose their ability to count with the introduction of calculators. When having a discussion regarding LLMs effect on new generations, we must be careful not to get lost in baseless intergenerational conflict and focus on actual phenomena.

To preserve human agency, there is a call for having a human in the loop to oversee AI produced content [S19, S30, S52, S62]. This is especially prevalent in Section 4.1, more specifically regarding manipulation and self-acting AI. Having a human "referee" the content might however prove to be challenging for a multitude of reasons. The amount of content that can be produced by LLMs is so staggering that the number of humans needed to monitor them properly is not feasible. Also, the capacity of LLMs to manipulate makes content review extremely difficult for humans - when presented with just one seemingly well-reasoned argument, it is not simple to know how to question it.

A logical solution for the challenges of having a human in the loop is to have a human assisted by AI. This in turn poses questions about when a human is actually making a decision (Ollila, 2019). If a reviewing human is given content filtered by a machine, accompanied by the machine's pre-calculated opinion of whether the content is right or wrong, is the human actually making their own decision?

The issues of unfair decision making in relation to LLMs can be expected to become more relevant in the future, as language models are implemented more into our daily lives. If, for example, a financial institution starts using LLM-powered interactive chatbots to handle customers applying for products, such as mortgages, the possible allocational harms will have a very real effect on people's lives. These types of outcomes have already been demonstrated many times with other AI applications (Inioluwa et al., 2022) and there is no reason to assume LLMs would be immune to them.

5.3 Potential benefits of this review

The knowledge and understanding of the issues presented in this review are helpful for practically anyone. Different AI solutions are currently being deployed all around us by governments, companies, and individuals. LLMs are rapidly joining this group and supplying applications on very important subjects. Understanding the threats and problems these applications have is a vital skill in modern society. This review brings together much of the discussion and challenges that are related to LLMs. The rest of this section details specific groups that can benefit from the findings of this review.

Legislators. Understanding the issues that are prevalent, or predicted, with LLMs is crucial to creating new regulation for them. The promised benefits of LLMs are vast, and the companies promising them powerful lobbyists. It is prudent for legislators to be vigilant and ensure that LLM creators are required to operate ethically.

The presented issues are also vital for the consideration of oversight methods. As this review has shown, there are significant issues related to certifications on responsibility. Any single certificate most likely cannot be trusted to achieve its intended results, so oversight and governance methods must be kept up to date and scrutinized constantly.

Developers. One might argue that the results of this review are most important for LLM developers, both companies and individual people. The results give concrete pointers as to where an LLM might go ethically or practically wrong when it is being developed. This review cannot be used as a direct guideline, but a developer wishing to create an application that is as ethical as possible would gain much insight from these results.

What is especially important for developers is to understand the limitations of LLMs when it comes to code generation. The models are trained on massive amounts of data, which means that the programming solutions they provide reflect the average skill set [S14]. Average, by definition, is not the best practice available. Knowing when and how to question AI suggestions is a vital skill for future programmers - and in fact, anybody using AI assistants for any reason.

Understanding these limitations is very important for companies that employ programmers. It may be tempting to allow an AI to do more of the development work, but this might mean that there is no human that understands how their programs work. This might expose them to unexpected challenges down the road. If the discussed threats of generated code vulnerability detailed earlier come true, applications developed "by" well-known

LLMs are in an extremely dangerous position.

Researchers and academics. Conducting research in the era of widespread LLMs is going to be quite different from what most researchers are used to, and how the academic world is built. This review details several ethical issues that must be dealt with and considered in academic circles. As language models become more eloquent, they can aid research magnificently but at the same time they offer new methods of academic misconduct.

When conducting research, even completely honestly, one must be keenly aware of the limitations of the tools used. Any content created with the aid of an AI tool should be treated as suspect, and the tool's user must always consider what is being left unsaid by the tool, and whose work is being used. The results of this review could also help in the creation of guidelines to mitigate some of these issues.

Educators. LLMs will be used by students, this cannot be denied. The utility these tools provide is simply too tempting to assume they would be left unused, even if they were prohibited. Something that this review did not tabulate as a possible solution for ethical issues caused by LLMs in education is prohibiting them. This was referred to in multiple studies, but without exception in the context of "some have resorted to this" (e.g. [S108, S110]). No study reviewed suggested this themselves. This makes sense, since if prohibition worked against cheating, it would have been solved at the dawn of education.

The future implications that LLMs will have in education still remain to be seen. This review can provide educators with some ideas as to what the anticipated risks and implications are. There are many things to look out for: cheating, loss of social skills, misunderstanding teaching purposes, and losing respect for education, to name a few. Being aware of these risks, educators can adapt their teaching methods and tools to cope with them.

The question of using AI as a teacher's assistant is also very relevant. Before considering using, for example, automated essay grading done by LLMs, educators would do well to understand all the risks and issues inherent in them.

LLM users. If it does not already, eventually this group of people is likely to encompass every individual living in a modern society. The number of LLM applications affecting everyday life can only be expected to rise. Therefore, being aware of the risks they pose is extremely important for everyone. The media exposure AI applications get tends to focus on successes and grand failures. This means that the more subtle issues can be left out of

public discourse. Knowing how to discuss and consider at least some of the ethical issues inherent in LLMs is a vital skill.

Media. Journalists can use the findings of this review to guide their reporting topics and give them more context about what sort of issues can exist in LLMs. In addition to this, media can consider their share of the responsibility about the workings of LLMs, the training corpora after all includes not-insignificant amounts of media content. This in turn can lead models to make connections that do not represent the real state of things [S81]. As media tends to report on the more dramatic and exceptional issues, they can become the norm for an ML solution being trained on such materials. For example, if most news articles containing the word "protest" also contain the word "violent" or "riot" in close proximity, a model being trained on these articles will associate these concepts. This association might then be represented in the model's outputs as a tendency to write about protests as being by default violent.

People generating textual content for a living are likely to find LLMs as writing assistants extremely tempting. While these applications can improve output amounts and quality, it is crucial to consider if they are subtly affecting the writer's opinions as discussed in Section 4.1.5.

5.4 Limitations and potential threats to validity

This review has been conducted by following the methods detailed by (Kitchenham and Charters, 2007) as much as possible. The main exception is that no quality assessment was done to the studies reviewed. This is because of the subjective nature of ethical issues that this review concerns - one cannot measure the quality of a conceptual observation. Because of this limitation, it was decided that even the empirical studies are not evaluated.

The lack of quality evaluation means that the amounts of identified issues cannot be directly compared. One reviewed study might merely mention a potential future risk, and another might demonstrate an issue with LLM outputs in an empirical manner. However, this does not mean that the results achieved are not valuable - relevant issues and topics of conversation or future study are identified. Their relevance is not insignificant even if they cannot be directly compared to each other.

The identification of ethical issues is naturally very subjective. The choice and relevance of these issues is clearly affected by the author's values and opinions as to what is important.

This can cause issues with the reproducibility of this review as another reviewer would not be likely to end up with identical categories. Similar issues are apparent with the mappings of the identified issues to the EU requirements for trustworthy AI. The arguments for these subjective decisions have been given in chapter 4 with the presentation of each issue.

6 Conclusions

This review conducted a systematic literature review of 116 studies that discuss ethical issues related to large language models at least on some level. The issues were grouped into 39 distinct, recurring, categories. These issues were then mapped into the seven requirements that are listed in the EU Ethics guidelines for trustworthy AI (EU High-Level Expert Group on AI, 2019): Human agency and oversight; Technical robustness and safety; Privacy and data governance; Transparency; Diversity, non-discrimination and fairness; Societal and environmental well-being; Accountability.

The mappings indicate that there are significant ethical issues related to every single one of the requirements for trustworthy AI. While many of these issues could be fixed with improved methods and technologies, some of them are fundamental to LLMs. Some examples of the fixable issues are problems with private information in training data and the natural resource consumption related to current models. One of the most identified issues on the other hand is not something that can be expected to be solved through technical research alone: the data used to train LLMs will always be biased, since there is no global understanding of what "unbiased data" might mean. Biased training data will result in biased outputs.

This review presents an overview of the current state of discussion and worries about the increasing usage of LLMs. The findings indicate that there is quite a bit of discussion related to these issues, much of it academical. Practical actions and solutions are not very prevalent. When this is compared to the hype that is surrounding the development and application of LLMs, there is a clear discrepancy.

Based on the results of this review, current LLM solutions pose drastic problems from every ethical viewpoint. It might make sense to pause their publication and see if these issues can be addressed on a general level. At the same time, societies should consider what it is that LLMs are offering, and what is the price we might have to pay for it. International competition naturally will not allow this to happen but hopefully societal awareness, and responsibility, regarding these issues will increase. We would all do well to remember the words, and events, of a 1993 movie: "Your scientists were so preoccupied with whether they could, that they didn't stop to think if they should".

References

Arredondo, P. (Apr. 19, 2023). "GPT-4 Passes the Bar Exam: What That Means for Artificial Intelligence Tools in the Legal Profession". In: *Stanford Law School*. URL: https://law.stanford.edu/2023/04/19/gpt-4-passes-the-bar-exam-what-that-means-for-artificial-intelligence-tools-in-the-legal-industry/ (visited on 08/31/2023).

Behnia, R., Ebrahimi, M. R., Pacheco, J., and Padmanabhan, B. (2022). "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. DOI: 10.1109/ICDMW58026.2022.00078.

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity Press, Cambridge, UK.

Berghel, H. (2023). "ChatGPT and AIChat Epistemology". In: *Computer* 56.5, pp. 130–137. DOI: 10.1109/MC.2023.3252379.

Chow, A. R. (Feb. 8, 2023). "How ChatGPT Managed to Grow Faster Than TikTok or Instagram". In: *TIME*. URL: https://time.com/6253615/chatgpt-fastest-growing/ (visited on 08/31/2023).

Dong, Y., Cordonnier, J.-B., and Loukas, A. (2021). "Attention is not all you need: pure attention loses rank doubly exponentially with depth". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 2793–2803. URL: https://proceedings.mlr.press/v139/dong21a.html.

Dwork, C. (2008). "Differential Privacy: A Survey of Results". In: *Theory and Applications of Models of Computation*. Ed. by M. Agrawal, D. Du, Z. Duan, and A. Li. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–19.

EU High-Level Expert Group on AI (2019). *Ethics Guidelines For Trustworthy AI*. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (visited on 07/31/2023).

Inioluwa, D. R., Kumar, I. E., Horowitz, A., and Selbst, A. D. (2022). "The Fallacy of AI Functionality". In: *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. DOI: 10.1145/3531146.3533158.

Jobin, A., Ienca, M., and Vayena, E. (2019). "The global landscape of AI ethics guidelines". In: *Nature Machine Intelligence* 1, pp. 389–399. DOI: 10.1038/s42256-019-0088-2.

Kitchenham, B. and Charters, S. (Jan. 2007). "Guidelines for performing Systematic Literature Reviews in Software Engineering". In: 2.

Luitse, D. and Denkena, W. (2021). "The great Transformer: Examining the role of large language models in the political economy of AI". In: *Big Data & Society* 8.2, p. 20539517211047734. DOI: 10.1177/20539517211047734. eprint: https://doi.org/10.1177/20539517211047734. URL: https://doi.org/10.1177/20539517211047734.

Ollila, M.-R. (2019). *Tekoälyn etiikkaa*. Helsinki, Finland: Kustannusosakeyhtiö Otava.

OpenAI (2023). *OpenAI API pricing*. URL: https://openai.com/pricing (visited on 06/29/2023).

Perrigo, B. (Jan. 18, 2023). "OpenAI Used Kenyan Workers on Less Than 2$PerHourtoMakeChatGPTLe$ In: *TIME*. URL: https://time.com/6247678/openai-chatgpt-kenya-workers/ (visited on 08/03/2023).

Sejnowski, T. J. (2023). "Large Language Models and the Reverse Turing Test". In: *Neural Computation* 35.3. Cited by: 3; All Open Access, Bronze Open Access, Green Open Access, pp. 309–342. DOI: 10.1162/neco_a_01563.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wu, T. Y., He, S. Z., Liu, J. P., Sun, S. Q., Liu, K., Han, Q.-L., and Y., T. (Mar. 2023). "A brief overview of ChatGPT: The history, status quo and potential future development". In: *IEEE/CAA Journal of Automatica Sinica* 10.5, pp. 1122–1135.

Appendix A Reviewed studies

Table A.1: List of studies used in this review.

Id	Authors	Year	Publication type	Study type	Title and (DOI or journal)
S1	Anderson S.S.	2023	Journal	Conceptual	"Places to stand": Multiple metaphors for framing ChatGPT's corpus (10.1016/j.compcom.2023.102778)
S2	†	2023	Journal	Conceptual	"So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy (10.1016/j.ijinfomgt.2023.102642)
S3	Taecharungroj V.	2023	Journal	Empirical	"What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter (10.3390/bdcc7010035)
S4	Lakim I.; Almazrouei E.; Alhaol I.A.; Debbah M.; Launay J.	2022	Workshop	Empirical	A Holistic Assessment of the Carbon Footprint of Noor, a Very Large Arabic Language Model (2022 Challenges and Perspectives in Creating Large Language Models, Proceedings of the Workshop)

†Dwivedi Y.K.; Kshetri N.; Hughes L.; Slade E.L.; Jeyaraj A.; Kar A.K.; Baabdullah A.M.; Koohang A.; Raghavan V.; Ahuja M.; Albanna H.; Albashrawi M.A.; Al-Busaidi A.S.; Balakrishnan J.; Barlette Y.; Basu S.; Bose I.; Brooks L.; Buhalis D.; Carter L.; Chowdhury S.; Crick T.; Cunningham S.W.; Davies G.H.; Davison R.M.; Dé R.; Dennehy D.; Duan Y.; Dubey R.; Dwivedi R.; Edwards J.S.; Flavián C.; Gauld R.; Grover V.; Hu M.-C.; Janssen M.; Jones P.; Junglas I.; Khorana S.; Kraus S.; Larsen K.R.; Latreille P.; Laumer S.; Malik F.T.; Mardani A.; Mariani M.; Mithas S.; Mogaji E.; Nord J.H.; O'Connor S.; Okumus F.; Pagani M.; Pandey N.; Papagiannidis S.; Pappas I.O.; Pathak N.; Pries-Heje J.; Raman R.; Rana N.P.; Rehm S.-V.; Ribeiro-Navarrete S.; Richter A.; Rowe F.; Sarker S.; Stahl B.C.; Tiwari M.K.; van der Aalst W.; Venkatesh V.; Viglia G.; Wade M.; Walton P.; Wirtz J.; Wright R.

Table A.1: List of studies used in this review.

Id	Authors	Year	Publication type	Study type	Title and (DOI or journal)
S5	Farrokhnia M.; Banihashem S.K.; Noroozi O.; Wals A.	2023	Journal	Conceptual	A SWOT analysis of ChatGPT: Implications for educational practice and research (10.1080/14703297.2023.2195846)
S6	Perkins M.	2023	Journal	Conceptual	Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond (10.53761/1.20.02.07)
S7	Yeo M.A.	2023	Journal	Conceptual	Academic integrity in the age of Artificial Intelligence (AI) authoring apps (10.1002/tesj.716)
S8	Poddar R,Sinha R,Naaman M,Jakesch M	2023	Conference	Empirical	AI Writing Assistants Influence Topic Choice in Self-Presentation (10.1145/3544549.3585893)
S9	Luo W.; He H.; Liu J.; Berson I.R.; Berson M.J.; Zhou Y.; Li H.	2023	Journal	Conceptual	Aladdin's Genie or Pandora's Box for Early Childhood Education? Experts Chat on the Roles, Challenges, and Developments of ChatGPT (10.1080/10409289.2023.2214181)
S10	Steinborn V.; Dufter P.; Jabbar H.; Schütze H.	2022	Conference	Empirical	An Information-Theoretic Approach and Dataset for Probing Gender Stereotypes in Multilingual Masked Language Models (Findings of the Association for Computational Linguistics: NAACL 2022 - Findings)
S11	Vakili T.; Dalianis H.	2021	Conference	Conceptual	Are Clinical BERT Models Privacy Preserving? The Difficulty of Extracting Patient-Condition Associations (CEUR Workshop Proceedings)

Table A.1: List of studies used in this review.

Id	Authors	Year	Publication type	Study type	Title and (DOI or journal)
S12	Haluza D.; Jungwirth D.	2023	Journal	Empirical	Artificial Intelligence and Ten Societal Megatrends: An Exploratory Study Using GPT-3 (10.3390/systems11030120)
S13	Kolides A.; Nawaz A.; Rathor A.; Beeman D.; Hashmi M.; Fatima S.; Berdik D.; Al-Ayyoub M.; Jararweh Y.	2023	Journal	Conceptual	Artificial intelligence foundation and pre-trained models: Fundamentals, applications, opportunities, and social impacts (10.1016/j.simpat.2023.102754)
S14	Pearce H.; Ahmad B.; Tan B.; Dolan-Gavitt B.; Karri R.	2022	Conference	Empirical	Asleep at the Keyboard? Assessing the Security of GitHub Copilot's Code Contributions (10.1109/SP46214.2022.9833571)
S15	Kirk H.R.; Jun Y.; Iqbal H.; Benussi E.; Volpin F.; Dreyer F.A.; Shtedritski A.; Asano Y.M.	2021	Conference	Empirical	Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models (Advances in Neural Information Processing Systems)
S16	Navigli R,Conia S,Ross B	2023	Journal	Conceptual	Biases in Large Language Models: Origins, Inventory and Discussion (10.1145/3597307)

Table A.1: List of studies used in this review.

Id	Authors	Year	Publication type	Study type	Title and (DOI or journal)
S17	Hessenthaler M.; Strubell E.; Hovy D.; Lauscher A.	2022	Conference	Conceptual	Bridging Fairness and Environmental Sustainability in Natural Language Processing (Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022)
S18	Shihadeh J.; Ackerman M.; Troske A.; Lawson N.; Gonzalez E.	2022	Conference	Empirical	Brilliance Bias in GPT-3 (10.1109/GHTC55712.2022.9910995)
S19	Kreps S.; Jakesch M.	2023	Journal	Empirical	Can AI communication tools increase legislative responsiveness and trust in democratic institutions? (10.1016/j.giq.2023.101829)
S20	Salvagno M.; Taccone F.S.; Gerli A.G.	2023	Journal	Conceptual	Can artificial intelligence help for scientific writing? (10.1186/s13054-023-04380-2)
S21	Balkır E.; Kiritchenko S.; Nejadgholi I.; Fraser K.C.	2022	Conference	Review	Challenges in Applying Explainability Methods to Improve the Fairness of NLP Models (TrustNLP 2022 - 2nd Workshop on Trustworthy Natural Language Processing, Proceedings of the Workshop)
S22	Welbl J.; Glaese A.; Uesato J.; Dathathri S.; Mellor J.; Hendricks L.A.; Anderson K.; Kohli P.; Coppin B.; Huang P.-S.	2021	Conference	Conceptual	Challenges in Detoxifying Language Models (Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021)

Table A.1: List of studies used in this review.

Id	Authors	Year	Publication type	Study type	Title and (DOI or journal)
S23	Kooli C.	2023	Journal	Conceptual	Chatbots in Education and Research: A Critical Examination of Ethical Implications and Solutions (10.3390/su15075614)
S24	Lund B.D.; Wang T.; Mannuru N.R.; Nie B.; Shimray S.; Wang Z.	2023	Journal	Conceptual	ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing (10.1002/asi.24750)
S25	De Angelis L.; Baglivo F.; Arzilli G.; Privitera G.P.; Ferragina P.; Tozzi A.E.; Rizzo C.	2023	Journal	Conceptual	ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health (10.3389/fpubh.2023.1166120)
S26	Sallam M.; Salim N.A.; Barakat M.; Al-Tammemi A.B.	2023	Journal	Conceptual	ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations (10.52225/narra.v3i1.103)
S27	Dowling M.; Lucey B.	2023	Journal	Conceptual	ChatGPT for (Finance) research: The Bananarama Conjecture (10.1016/j.frl.2023.103662)
S28	Rahman M.M.; Watanobe Y.	2023	Journal	Conceptual	ChatGPT for Education and Research: Opportunities, Threats, and Strategies (10.3390/app13095783)
S29	J. Mrabet; R. Studholme	2023	Conference	Conceptual	ChatGPT: A friend or a foe? (10.1109/ICCIKE58312.2023.10131713)

Table A.1: List of studies used in this review.

Id	Authors	Year	Publication type	Study type	Title and (DOI or journal)
S30	A. Bahrini; M. Khamoshifar; H. Abbasimehr; R. J. Riggs; M. Esmaeili; R. M. Majdabadkohne; M. Pasehvar	2023	Conference	Conceptual	ChatGPT: Applications, Opportunities, and Threats (10.1109/SIEDS58326.2023.10137850)
S31	M. Abdullah; A. Madain; Y. Jararweh	2022	Conference	Conceptual	ChatGPT: Fundamentals, Applications and Social Impacts (10.1109/SNAMS58071.2022.10062688)
S32	Gill S.S.; Kaur R.	2023	Journal	Conceptual	ChatGPT: Vision and challenges (10.1016/j.iotcps.2023.05.004)
S33	Cotton D.R.E.; Cotton P.A.; Shipway J.R.	2023	Journal	Conceptual	Chatting and cheating: Ensuring academic integrity in the era of ChatGPT (10.1080/14703297.2023.2190148)
S34	Choi E.P.H.; Lee J.J.; Ho M.-H.; Kwok J.Y.Y.; Lok K.Y.W.	2023	Journal	Conceptual	Chatting or cheating? The impacts of ChatGPT and other artificial intelligence language models on nurse education (10.1016/j.nedt.2023.105796)
S35	Gao C.A.; Howard F.M.; Markov N.S.; Dyer E.C.; Ramesh S.; Luo Y.; Pearson A.T.	2023	Journal	Conceptual	Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers (10.1038/s41746-023-00819-6)

Table A.1: List of studies used in this review.

Id	Authors	Year	Publication type	Study type	Title and (DOI or journal)
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	Journal	Conceptual	Could a conversational ai identify offensive language?† (10.3390/info12100418)
S37	Mirowski P,Mathewson KW,Pittman J,Evans R	2023	Conference	Conceptual	Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals (10.1145/3544548.3581225)
S38	Jakesch M,Bhat A,Buschek D,Zalmanson L,Naaman M	2023	Conference	Empirical	Co-Writing with Opinionated Language Models Affects Users' Views (10.1145/3544548.3581196)

Table A.1: List of studies used in this review.

Id	Authors	Year	Publication type	Study type	Title and (DOI or journal)
S39	Jernite Y,Nguyen H,Biderman S,Rogers A,Masoud M,Danchev V,Tan S,Luccioni AS,Subramani N,Johnson I,Dupont G,Dodge J,Lo K,Talat Z,Radev D,Gokaslan A,Nikpoor S,Henderson P,Bommasani R,Mitchell M	2022	Conference	Conceptual	Data Governance in the Age of Large-Scale Data-Driven Language Technology (10.1145/3531146.3534637)
S40	Panchendrarajan R.; Bhoi S.	2021	Conference	Empirical	Dataset reconstruction attack against language models (CEUR Workshop Proceedings)
S41	Huang W.R.; Chien S.; Thakkar O.; Mathews R.	2022	Conference	Empirical	Detecting Unintended Memorization in Language-Model-Fused ASR (10.21437/Interspeech.2022-10909)
S42	Kansteiner W.	2022	Journal	Conceptual	DIGITAL DOPING FOR HISTORIANS: CAN HISTORY, MEMORY, AND HISTORICAL THEORY BE RENDERED ARTIFICIALLY INTELLIGENT? (10.1111/hith.12282)
S43	H. Alamleh; A. A. S. AlQahtani; A. ElSaid	2023	Conference	Empirical	Distinguishing Human-Written and ChatGPT-Generated Text Using Machine Learning (10.1109/SIEDS58326.2023.10137767)

Table A.1: List of studies used in this review.

Id	Authors	Year	Publication type	Study type	Title and (DOI or journal)
S44	Lee J.; Le T.; Chen J.; Lee D.	2023	Conference	Empirical	Do Language Models Plagiarize? (10.1145/3543507.3583199)
S45	Lehman E.; Jain S.; Pichotta K.; Goldberg Y.; Wallace B.C.	2021	Conference	Empirical	Does BERT Pretrained on Clinical Notes Reveal Sensitive Data? (NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference)
S46	Limisiewicz T.; Mareček D.	2022	Conference	Empirical	Don't Forget About Pronouns: Removing Gender Bias in Language Models Without Losing Factual Gender Information (GeBNLP 2022 - 4th Workshop on Gender Bias in Natural Language Processing, Proceedings of the Workshop)
S47	Vakili T.; Lamproudis A.; Henriksson A.; Dalianis H.	2022	Conference	Conceptual	Downstream Task Performance of BERT Models Pre-Trained Using Automatically De-Identified Clinical Data (2022 Language Resources and Evaluation Conference, LREC 2022)
S48	K. T. Gradon	2023	Journal	Conceptual	Electric Sheep on the Pastures of Disinformation and Targeted Phishing Campaigns: The Security Implications of ChatGPT (10.1109/MSEC.2023.3255039)
S49	J. Qadir	2023	Conference	Conceptual	Engineering Education in the Era of ChatGPT: Promise and Pitfalls of Generative AI for Education (10.1109/EDUCON54358.2023.10125121)

Table A.1: List of studies used in this review.

Id	Authors	Year	Publication type	Study type	Title and (DOI or journal)
S50	Y. Mori; Y. Miyake	2022	Conference	Conceptual	Ethical Issues in Automatic Dialogue Generation for Non-Player Characters in Digital Games (10.1109/BigData55660.2022.10020271)
S51	Hämäläinen P,Tavast M,Kunnari A	2023	Conference	Conceptual	Evaluating Large Language Models in Generating Synthetic HCI Research Data: A Case Study (10.1145/3544548.3580688)
S52	Cooper G.	2023	Journal	Conceptual	Examining Science Education in ChatGPT: An Exploratory Study of Generative Artificial Intelligence (10.1007/s10956-023-10039-y)
S53	Poulton A,Eliens S	2022	Conference	Conceptual	Explaining Transformer-Based Models for Automatic Short Answer Grading (10.1145/3488466.3488479)
S54	H. Nguyen; A. Malik; M. Zink	2022	Journal	Conceptual	Exploring Realtime Conversational Virtual Characters (10.5594/JMI.2022.3153646)
S55	He X.; Chen C.; Lyu L.; Xu Q.	2022	Conference	Empirical	Extracted BERT Model Leaks More Information than You Think! (Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022)
S56	Carlini N.; Tramèr F.; Wallace E.; Jagielski M.; Herbert-Voss A.; Lee K.; Roberts A.; Brown T.; Song D.; Erlingsson Ú.; Oprea A.; Raffel C.	2021	Conference	Empirical	Extracting training data from large language models (Proceedings of the 30th USENIX Security Symposium)

Table A.1: List of studies used in this review.

Id	Authors	Year	Publication type	Study type	Title and (DOI or journal)
S57	Kumar R.	2023	Journal	Conceptual	Faculty members' use of artificial intelligence to grade student papers: a case of implications (10.1007/s40979-023-00130-7)
S58	Ramesh K.; Sitaram S.; Choudhury M.	2023	Conference	Conceptual	Fairness in Language Models Beyond English: Gaps and Challenges (EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Findings of EACL 2023)
S59	Biderman S.; Raff E.	2022	Conference	Empirical	Fooling MOSS Detection with Pretrained Language Models (10.1145/3511808.3557079)
S60	Dergaa I.; Chamari K.; Zmijewski P.; Saad H.B.	2023	Journal	Review	From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing (10.5114/BIOLSPORT.2023.125623)
S61	Li Y.; Zhang G.; Yang B.; Lin C.; Ragni A.; Wang S.; Fu J.	2022	Conference	Empirical	HERB: Measuring Hierarchical Regional Bias in Pre-trained Language Models (2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing - Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022)
S62	Wahle J.P.; Ruas T.; Kirstein F.; Gipp B.	2022	Conference	Empirical	How Large Language Models are Transforming Machine-Paraphrased Plagiarism (Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022)

Table A.1: List of studies used in this review.

Id	Authors	Year	Publication type	Study type	Title and (DOI or journal)
S63	Yan D.	2023	Journal	Empirical	Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation (10.1007/s10639-023-11742-4)
S64	Kasirzadeh A.; Gabriel I.	2023	Journal	Conceptual	In Conversation with Artificial Intelligence: Aligning language Models with Human Values (10.1007/s13347-023-00606-x)
S65	Pikuliak M.; Beňová I.; Bachratý V.	2023	Conference	Empirical	In-Depth Look at Word Filling Societal Bias Measures (EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference)
S66	Bhat A.; Agashe S.; Oberoi P.; Mohile N.; Jangir R.; Joshi A.	2023	Conference	Empirical	Interacting with Next-Phrase Suggestions: How Suggestion Systems Aid and Influence the Cognitive Processes of Writing (10.1145/3581641.3584060)
S67	Kumar S.; Balachandran V.; Njoo L.; Anastasopoulos A.; Tsvetkov Y.	2023	Conference	Review	Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey (EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference)
S68	Li X.; Tramèr F.; Liang P.; Hashimoto T.	2022	Conference	Conceptual	LARGE LANGUAGE MODELS CAN BE STRONG DIFFERENTIALLY PRIVATE LEARNERS (ICLR 2022 - 10th International Conference on Learning Representations)

Table A.1: List of studies used in this review.

Id	Authors	Year	Publication type	Study type	Title and (DOI or journal)
S69	Schramowski P.; Turan C.; Andersen N.; Rothkopf C.A.; Kersting K.	2022	Journal	Empirical	Large pre-trained language models contain human-like biases of what is right and wrong to do (10.1038/s42256-022-00458-8)
S70	Crawford J.; Cowling M.; Allen K.-A.	2023	Journal	Conceptual	Leadership is needed for ethical ChatGPT: Character, assessment, and learning using artificial intelligence (AI) (10.53761/1.20.3.02)
S71	Alshahrani S.; Wali E.; Matthews J.	2022	Conference	Conceptual	Learning From Arabic Corpora But Not Always From Arabic Speakers: A Case Study of the Arabic Wikipedia Editions (WANLP 2022 - 7th Arabic Natural Language Processing - Proceedings of the Workshop)
S72	Borchers C.; Gala D.S.; Gilburt B.; Oravkin E.; Bounsi W.; Asano Y.M.; Kirk H.R.	2022	Conference	Empirical	Looking for a Handsome Carpenter! Debiasing GPT-3 Job Advertisements (GeBNLP 2022 - 4th Workshop on Gender Bias in Natural Language Processing, Proceedings of the Workshop)
S73	Kraft A.; Zorn H.-P.; Fecht P.; Simon J.; Biemann C.; Usbeck R.	2022	Conference	Empirical	Measuring Gender Bias in German Language Generation (10.18420/inf2022_108)

Table A.1: List of studies used in this review.

Id	Authors	Year	Publication type	Study type	Title and (DOI or journal)
S74	Touileb S.; Nozza D.	2022	Conference	Empirical	Measuring Harmful Representations in Scandinavian Language Models (NLPCSS 2022 - 5th Workshop on Natural Language Processing and Computational Social Science ,NLP+CSS, Held at the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022)
S75	Nozza D.; Bianchi F.; Lauscher A.; Hovy D.	2022	Conference	Empirical	Measuring Harmful Sentence Completion in Language Models for LGBTQIA+ Individuals (LTEDI 2022 - 2nd Workshop on Language Technology for Equality, Diversity and Inclusion, Proceedings of the Workshop)
S76	Mei A.; Kabir A.; Levy S.; Subbiah M.; Allaway E.; Judge J.; Patton D.; Bimber B.; McKeown K.; Wang W.Y.	2022	Conference	Conceptual	Mitigating Covertly Unsafe Text within Natural Language Systems (Findings of the Association for Computational Linguistics: EMNLP 2022)
S77	Venkit P.N.; Gautam S.; Panchanadikar R.; Huang T.-H.; Wilson S.	2023	Conference	Empirical	Nationality Bias in Text Generation (EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference)
S78	Longoni C.; Fradkin A.; Cian L.; Pennycook G.	2022	Conference	Conceptual	News from Generative Artificial Intelligence Is Believed Less (10.1145/3531146.3533077)

Table A.1: List of studies used in this review.

Id	Authors	Year	Publication type	Study type	Title and (DOI or journal)
S79	Dev S.; Sheng E.; Zhao J.; Amstutz A.; Sun J.; Hou Y.; Sanseverino M.; Kim J.; Nishi A.; Peng N.; Chang K.-W.	2022	Conference	Conceptual	On Measures of Biases and Harms in NLP (2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing - Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022)
S80	Akyurek A.F.; Paik S.; Kocyigit M.Y.; Akbiyik S.; Runyun S.L.; Wijaya D.	2022	Conference	Empirical	On Measuring Social Biases in Prompt-Based Multi-Task Learning (Findings of the Association for Computational Linguistics: NAACL 2022 - Findings)
S81	Bender E.M.; Gebru T.; McMillan-Major A.; Shmitchell S.	2021	Conference	Conceptual	On the dangers of stochastic parrots: Can language models be too big? (10.1145/3442188.3445922)
S82	Vashishtha A.; Prasad S.S.K.; Bajaj P.; Chaudhary V.; Cook K.; Dandapat S.; Sitaram S.; Choudhury M.	2023	Conference	Empirical	Performance and Risk Trade-offs for Multi-word Text Prediction at Scale (EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Findings of EACL 2023)
S83	Abid A.; Farooqi M.; Zou J.	2021	Conference	Empirical	Persistent Anti-Muslim Bias in Large Language Models (10.1145/3461702.3462624)

Table A.1: List of studies used in this review.

Id	Authors	Year	Publication type	Study type	Title and (DOI or journal)
S84	Qian R.; Ross C.; Fernandes J.; Smith E.; Kiela D.; Williams A.	2022	Conference	Empirical	Perturbation Augmentation for Fairer NLP (Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022)
S85	Nozza D.; Bianchi F.; Hovy D.	2022	Conference	Conceptual	Pipelines for Social Bias Testing of Large Language Models (2022 Challenges and Perspectives in Creating Large Language Models, Proceedings of the Workshop)
S86	Elmahdy A.; Inan H.A.; Sim R.	2022	Conference	Conceptual	Privacy Leakage in Text Classification: A Data Extraction Approach (PrivateNLP 2022 - 4th Workshop on Privacy in Natural Language Processing at the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Workshop)
S87	Zhao X.; Li L.; Wang Y.-X.	2022	Conference	Empirical	Provably Confidential Language Modelling (NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference)
S88	Mireshghallah F.; Goyal K.; Uniyal A.; Berg-Kirkpatrick T.; Shokri R.	2022	Conference	Empirical	Quantifying Privacy Risks of Masked Language Models Using Membership Inference Attacks (Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022)

Table A.1: List of studies used in this review.

Id	Authors	Year	Publication type	Study type	Title and (DOI or journal)
S89	Perez E.; Huang S.; Song F.; Cai T.; Ring R.; Aslanides J.; Glaese A.; McAleese N.; Irving G.	2022	Conference	Empirical	Red Teaming Language Models with Language Models (Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022)
S90	H. Ibrahim; R. Asim; F. Zaffar; T. Rahwan; Y. Zaki	2023	Journal	Conceptual	Rethinking Homework in the Age of Artificial Intelligence (10.1109/MIS.2023.3255599)
S91	Dinan E.; Abercrombie G.; Bergman A.S.; Spruit S.; Hovy D.; Boureau Y.L.; Rieser V.	2022	Conference	Conceptual	SafetyKit: First Aid for Measuring Safety in Open-domain Conversational Systems (Proceedings of the Annual Meeting of the Association for Computational Linguistics)
S92	D. V. Grbic; I. Dujlovic	2023	Conference	Conceptual	Social engineering with ChatGPT (10.1109/INFOTEH57020.2023.10094141)
S93	Sheng E.; Chang K.-W.; Natarajan P.; Peng N.	2021	Conference	Review	Societal biases in language generation: Progress and challenges (ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference)
S94	Gero KI,Liu V,Chilton L	2022	Conference	Conceptual	Sparks: Inspiration for Science Writing Using Language Models (10.1145/3532106.3533533)

Table A.1: List of studies used in this review.

Id	Authors	Year	Publication type	Study type	Title and (DOI or journal)
S95	Sayenju S.; Aygun R.; Franks B.; Johnston S.; Lee G.; Modgil G.	2022	Conference	Conceptual	Stereotype and Categorical Bias Evaluation via Differential Cosine Bias Measure (10.1109/BigData55660.2022.10020924)
S96	Zhou J,Zhang Y,Luo Q,Parker AG,De Choudhury M	2023	Conference	Empirical	Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions (10.1145/3544548.3581318)
S97	Weidinger L.; Uesato J.; Rauh M.; Griffin C.; Huang P.-S.; Mellor J.; Glaese A.; Cheng M.; Balle B.; Kasirzadeh A.; Biles C.; Brown S.; Kenton Z.; Hawkins W.; Stepleton T.; Birhane A.; Hendricks L.A.; Rimell L.; Isaac W.; Haas J.; Legassick S.; Irving G.; Gabriel I.	2022	Conference	Review	Taxonomy of Risks posed by Language Models (10.1145/3531146.3533088)

Table A.1: List of studies used in this review.

Id	Authors	Year	Publication type	Study type	Title and (DOI or journal)
S98	Sun G.H.; Hoelscher S.H.	2023	Journal	Conceptual	The ChatGPT Storm and What Faculty Can Do (10.1097/NNE.0000000000001390)
S99	Heidenreich H.S.; Williams J.R.	2021	Conference	Empirical	The Earth Is Flat and the Sun Is Not a Star: The Susceptibility of GPT-2 to Universal Adversarial Triggers (10.1145/3461702.3462578)
S100	Dalalah D.; Dalalah O.M.A.	2023	Journal	Conceptual	The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT (10.1016/j.ijme.2023.100822)
S101	Luitse D.; Denkena W.	2021	Journal	Conceptual	The great transformer: Examining the role of large language models in the political economy of AI (10.1177/20539517211047734)
S102	L. Rosenberg	2023	Conference	Conceptual	The Metaverse and Conversational AI as a Threat Vector for Targeted Influence (10.1109/CCWC57344.2023.10099167)
S103	S. Murugesan; A. K. Cherukuri	2023	Journal	Conceptual	The Rise of Generative Artificial Intelligence and Its Impact on Education: The Promises and Perils (10.1109/MC.2023.3253292)
S104	Weisz J.D.; Muller M.; He J.; Houde S.	2023	Conference	Conceptual	Toward General Design Principles for Generative AI Applications (CEUR Workshop Proceedings)
S105	Anoop K; Manjary P. Gangan; Deepak P; Lajish V. L	2022	Conference	Review	Towards an Enhanced Understanding of Bias in Pre-trained Neural Language Models: A Survey with Special Emphasis on Affective Bias (10.1007/978-981-19-4453-6_2)

Table A.1: List of studies used in this review.

Id	Authors	Year	Publication type	Study type	Title and (DOI or journal)
S106	Hershcovich D.; Webersinke N.; Kraus M.; Bingler J.A.; Leippold M.	2022	Conference	Conceptual	Towards Climate Awareness in NLP Research (Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022)
S107	Lin S.; Hilton J.; Evans O.	2022	Conference	Empirical	TruthfulQA: Measuring How Models Mimic Human Falsehoods (Proceedings of the Annual Meeting of the Association for Computational Linguistics)
S108	Su J.; Yang W.	2023	Journal	Conceptual	Unlocking the Power of ChatGPT: A Framework for Applying Generative AI in Education (10.1177/20965311231168423)
S109	Steed R.; Panda S.; Kobren A.; Wick M.	2022	Conference	Empirical	Upstream Mitigation Is Not All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models (Proceedings of the Annual Meeting of the Association for Computational Linguistics)
S110	Tlili A.; Shehata B.; Adarkwah M.A.; Bozkurt A.; Hickey D.T.; Huang R.; Agyemang B.	2023	Journal	Empirical	What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education (10.1186/s40561-023-00237-x)

Table A.1: List of studies used in this review.

Id	Authors	Year	Publication type	Study type	Title and (DOI or journal)
S111	Luccioni A.; Viviano J.D.	2021	Conference	Empirical	What's in the Box? A Preliminary Analysis of Undesirable Content in the Common Crawl Corpus (ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference)
S112	Ladhak F.; Durmus E.; Suzgun M.; Zhang T.; Jurafsky D.; McKeown K.; Hashimoto T.	2023	Conference	Empirical	When Do Pre-Training Biases Propagate to Downstream Tasks? A Case Study in Text Summarization (EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference)
S113	Gururangan S.; Card D.; Dreier S.K.; Gade E.K.; Wang L.Z.; Wang Z.; Zettlemoyer L.; Smith N.A.	2022	Conference	Empirical	Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection (Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022)
S114	Liu Y.; Mittal A.; Yang D.; Bruckman A.	2022	Conference	Empirical	Will AI Console Me when I Lose my Pet? Understanding Perceptions of AI-Mediated Email Writing (10.1145/3491102.3517731)

Table A.1: List of studies used in this review.

Id	Authors	Year	Publication type	Study type	Title and (DOI or journal)
S115	Li H.; Song Y.; Fan L.	2022	Conference	Empirical	You Don't Know My Favorite Color: Preventing Dialogue Representations from Revealing Speakers' Private Personas (NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference)
S116	Talat Z.; Névéol A.; Biderman S.; Clinciu M.; Dey M.; Longpre S.; Luccioni A.S.; Masoud M.; Mitchell M.; Radev D.; Sharma S.; Subramonian A.; Tae J.; Tan S.; Tunuguntla D.; van der Wal O.	2022	Conference	Conceptual	You Reap What You Sow: On the Challenges of Bias Evaluation Under Multilingual Settings (2022 Challenges and Perspectives in Creating Large Language Models, Proceedings of the Workshop)

Appendix B Identified ethical issues

Table B.1: Ethical issue categories identified.

Issue	Issues relating to...	EU Requirement
Ambiguity of accountability	who is legally or ethically responsible for LLMs effects	Accountability
Corporate influence	large corporations having advantage and power in decisions of future	Accountability
Cost of privacy	the cost of making sure an LLM and training data are private	Accountability
Cost of AI monitoring	the cost of retraining and monitoring LLMs	Accountability
Biased training data or outputs	general concerns of biases regarding LLMs	Diversity, non-discrimination and fairness
Discriminatory results	documented discrimination or practical concerns about biases	Diversity, non-discrimination and fairness
Lack of global definition for bias or fairness	applying fairness in different contexts	Diversity, non-discrimination and fairness
Non-binary gender neglected	how bias tests only reflect on binary gender	Diversity, non-discrimination and fairness
Toxic content	toxic outputs of LLMs	Diversity, non-discrimination and fairness
Promotes inequality	how LLMs offer possibilities and tools in an unequal manner	Diversity, non-discrimination and fairness
Loss of learning or the ease of cheating	LLMs enabling cheating and reducing need to learn	Human agency and oversight
Fake news or misinformation	LLMs generating and spreading misinformation	Human agency and oversight
Echo chambers	formation of content echo chambers and the "yea-sayer effect"	Human agency and oversight

Table B.1: Ethical issue categories identified.

Issue	Issues relating to...	EU Requirement
Self-acting AI	AI making unmonitored decisions, (does not mean general AI)	Human agency and oversight
Influence through suggestions	writing assistants influencing user output, even opinion	Human agency and oversight
Manipulation	LLMs being manipulative, or being used/designed for that purpose	Human agency and oversight
Data gathered without consent	ethics of the gathered training data	Privacy and data governance
Privacy and data security	general concerns of privacy and data security	Privacy and data governance
Data management	challenges with managing the vast amounts of data LLMs deal with	Privacy and data governance
Replacement of traditional learning	learning contexts becoming blurred, ethical issues with automatic grading	Societal and environmental well-being
Reduced value of education	over-reliance on technology, degrees attained without proper knowledge	Societal and environmental well-being
Loss of social skills	how an abundance of AI solutions in education can reduce human contact	Societal and environmental well-being
Paper or crediental generation	inflated number of papers, contributing to the "Matthew effect"	Societal and environmental well-being
Purposeful toxic or immoral content	unethical use of LLMs, e.g., misinformation or (cyber)crime	Societal and environmental well-being
Job loss or class divide	widespread use of LLMs leading to job loss and societal shift	Societal and environmental well-being
Training data pruning	the human cost of de-toxifying training data	Societal and environmental well-being
Environmental impacts	environmental cost of training and using LLMs	Societal and environmental well-being
Fairwashing	how certificates could be used to polish company image	Societal and environmental well-being
Bias scoring and solutions	existing bias mitigation techniques	Technical robustness and safety
Data leakage or unintended memorization	LLMs memorizing content verbatim and it being extracted	Technical Robustness and safety
Inaccurate results	LLMs producing unfactual content	Technical robustness and safety

Table B.1: Ethical issue categories identified.

Issue	Issues relating to...	EU Requirement
Dangerous content	factually unsafe content produced by LLMs	Technical robustness and safety
Alignment	LLMs functioning in alignment with human ethics	Technical robustness and safety
LLM as an author	ambiguity of having an LLM be a co-contributor in scientific publishing	Transparency
Academic integrity or source tracking	LLMs hallucinating references or their output being unverifiable	Transparency
Unfair decision making	fairness that might not be directly related to discrimination	Transparency
Lack of transparency	transparency of LLMs, or lack thereof	Transparency
Copyright infringement	LLM-produced content being in violation of copyright	Transparency
Unfactual training data	LLM training data being of poor quality and becoming outdated	Transparency