OPINION PAPER

Large Language Models as Search Engines: Societal Challenges

Zacchary Sadeddine
Télécom Paris, Institut Polytechnique de Paris
France
zacchary.sadeddine@telecom-paris.fr

Winston Maxwell
Télécom Paris, Institut Polytechnique de Paris
France
winston.maxwell@telecom-paris.fr

Gaël Varoquaux
INRIA Saclay
France
gael.varoquaux@inria.fr

Fabian M. Suchanek
Télécom Paris, Institut Polytechnique de Paris
France
fabian.suchanek@telecom-paris.fr

Abstract

Large Language Models (LLMs) may one day replace search engines as the primary portal to information on the Web. In this opinion paper, we investigate the societal challenges that such a change could bring. We focus on the roles of LLM Providers, Content Creators, and End Users, and identify 15 types of challenges. With each, we show current mitigation strategies – both from the technical perspective and the legal perspective. We also discuss the impact of each challenge and point out future research opportunities.

1 Introduction

Large Language Models (LLMs) are increasingly used as portals to information on the Web. Google is rolling out AI overviews above its search results[1] building upon its language models[2], Microsoft's Bing search engine[3] allows sending the query to Microsoft's Co-pilot, DuckDuckGo[4] and Brave Search[5] offer AI-assisted answers, and browsers such as Opera, Brave, and Edge have built-in AI-plugins for query answering. These developments are changing the way users access information: instead of querying the Web with a search engine, reading one or several result pages, and finding the information, people can now ask their question to the AI assistant, which will synthesize an answer for the user from Web sources. This means that LLMs have the potential to severely disrupt the search engine ecosystem, which has been comparatively stable for the last 25 years, and to completely change the way the Web is used.

[1] https://blog.google/products/search/generative-ai-google-search-may-2024/
[2] https://ai.google/get-started/our-models/
[3] https://www.bing.com/
[4] https://duckduckgo.com/
[5] https://search.brave.com/

In this article, we look at the societal challenges that such a change would bring. Several studies have already surveyed the general risks posed by LLMs [Bommasani et al., 2022; Weidinger et al., 2022; of the High Comissioner, 2024; Lorenz et al., 2023; Suchanek and Luu, 2023; Bengio et al., 2025]. The role of LLMs as a search engine has received less attention – even though LLMs have the potential to disrupt the search engine market, challenging the market position of Google [Liu et al., 2024]. In this article, we do not examine the competitive effects LLMs on the search engine market. Instead, we focus on the risks for individuals and for society as a whole resulting from the increasing use of LLMs as a means to access information on the Web. We collect the main societal issues that have been identified so far – both in the literature and in the press. We also discuss current approaches that are being or could be implemented in order to remedy these issues – both from a technical and a legal point of view. With this, we aim to put the risks into perspective, and to identify open avenues of research.

Investigating the impact of LLMs on society means shooting at a moving target: new challenges appear continuously, and hence the field needs continuous updating. With this survey, we add the next step in this path – being sure that it is not the last one.

2 Preliminaries

A Language Model (LM) is a probability distribution over a sequence of words in a language. Such a model is trained on large textual corpora, and it can be used to predict the likelihood of a word that follows given preceding words. Early LMs were N-gram models and Hidden Markov Models. The first neural LM was Word2Vec [Mikolov et al., 2013], but the break-through for neural LMs came with the Transformer architecture [Vaswani et al., 2017]. The decoder-only transformers, in particular, can participate in entire conversations by consecutively predicting the next word of their reply. At the time of this writing, the most powerful decoder models are GPT-4 [OpenAI et al., 2024], Llama-3 [Dubey et al., 2024], and Gemini [Team, 2024]. These models are trained on large Web corpora and have billions of parameters, which is why they are known as Large Language Models (LLMs).

An LLM is typically made available to End Users via an online interface, a mobile app, or a downloadable software – either for free or for a subscription fee. Users can pose a query (a *prompt*), and the LLM answers – possibly engaging in a back-and-forth conversation. Under the hood, such a conversation always starts with a hard-coded pre-prompt. Such a pre-prompt is a series of instructions that guides the model's behavior, indicating what kind of answers are expected, and what types of behaviors are to be avoided (e.g., aggressiveness). In addition, the most recent LLMs are equipped with retrieval-augmented generation (RAG). This is a mechanism that retrieves documents that match the user query from the Web, appends it to the prompt, and asks the model to generate its answer based on these documents [Lewis et al., 2020].

Due to their significant impact, LLM have attracted regulatory attention. In the European AI Act[6], LLMs are part of a bigger category called "general purpose AI models". Such a model "displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into

[6]Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence

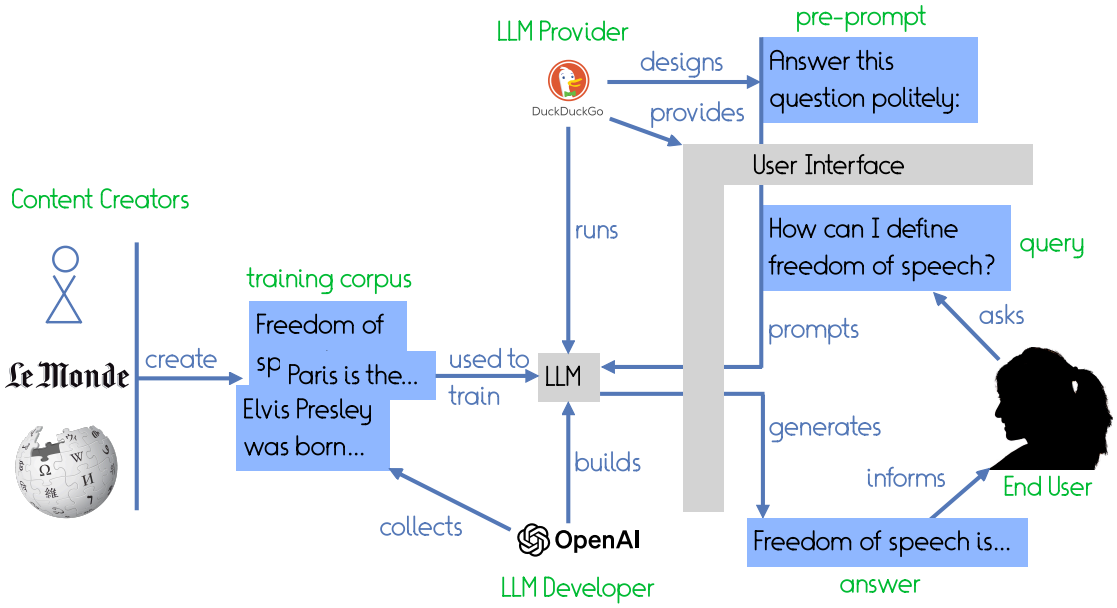

Figure 1. Parties in an LLM ecosystem

a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market"[7]. LLMs that fit into this category are subject to transparency obligations under the AI Act, as we will see below. If the cumulative amount of computation used for the training of a model measured in floating point operations exceeds 10^{25}, then the model is presumed to pose a "systemic risk" under the AI Act, and enhanced risk analysis, testing, cyber-security and transparency obligations apply[8].

2.1 The LLM Ecosystem - Legal Definitions

The European AI Act defines the *provider* of an AI system as follows: "a natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge"[9]. A downstream provider is "a provider of an AI system, including a general-purpose AI system, which integrates an AI model, regardless of whether the AI model is provided by themselves and vertically integrated or provided by another entity based on contractual relations"[10].

This definition of a provider refers to both the entity that develops the system and the entity that places it on the market. For the purposes of this paper, it is more convenient to split these two roles: we call *LLM Developer* the entity that develops the system, and *LLM Provider* the

[7] AI Act, Article 3(63)
[8] AI Act, Article 51(2)
[9] AI Act, Article 3(3)
[10] AI Act, Article 3(68)

entity that places it on the market (including as downstream provider). This leads us to focus on the following actors in the LLM ecosystem (Figure 1):

Content Creators are the people or organizations that produce and publish the textual content on which the LLM is trained. Today, LLMs are usually trained mainly on the Web. Hence, examples of Content Creators are Web publishers, blog owners, news agencies, online magazines and newspapers, government agencies, companies, individual home page owners, and the contributors to platforms such as Wikipedia.

LLM Developers are the people or organizations that create an LLM, i.e., that design (or choose) an architecture, collect a training corpus (by choosing content from Content Creators), train the LLM, and design a pre-prompt. Prominent LLM Developers at the time of this writing are OpenAI (GPT-* models), Meta (LLama models), and Google (PALM/Gemini model).

LLM Providers are the people or organizations that make an LLM available to businesses or the general public. Usually, the LLM Developers are also LLM Providers, but in principle, an LLM can be offered to the public by someone else than the developer. There may also be "downstream providers", which purchase the LLM from an upstream provider and offer it to the downstream provider's customers.

End Users are the people who query the LLM. Typically, the user query is concatenated with a pre-prompt and then sent to the LLM. The LLM then generates an answer, which is sent back to the user. LLMs are now powerful enough to help not only for factual questions, but also for creative and verbose tasks, such as writing essays or code. For many users, this makes LLMs a more advanced version of classical search engines.

Society as a whole represents the interests of citizens and democratic institutions in general, who may be indirectly harmed by the use of LLMs.

We will look into the challenges that arise for each of these actors in the context of Language Models taking the place of search engines.

3 Challenges for Content Creators

We start our survey with challenges that arise for Content Creators. Their content is the basis for the success of the models. As we will see, however, their contribution is not necessarily rewarded.

3.1 Copyrighted Content

3.1.1 Reproduction of Content

The first challenge that arises when LLMs are trained on Web content is that this content is usually protected by copyright. LLMs have the ability to memorize this content and to reproduce it verbatim in parts or in whole [Karamolegkou et al., 2023], and thereby infringe the copyright. For example, the content of Wikipedia is under a Creative Commons Attribution Share-Alike

license. This entails a number of constraints on a re-publication of the content, most visibly that anybody who publicly reproduces this content has to cite Wikipedia as a source. At the time of this writing, certain RAG-enabled LLMs such as ChatGPT, Mistral's Le Chat, Microsoft's Co-pilot, and DuckDuckGo's AI-Assist have the ability to search the Web and to cite their source this way. However, they might cite any source that itself cites the original source; they might make up answers and pretend they come from the source [US District Court, Southern District of New York, 2023]; and they might also make up sources [Ravi et al., 2024]. Any LLM that reproduces Wikipedia page content this way infringes the license of Wikipedia.

Content licenses may not just require attribution, but may prohibit reproduction altogether. This applies in particular to content that lives behind a paywall. In a lawsuit that the New York Times has engaged against OpenAI [US District Court, Southern District of New York, 2023], the newspaper could show that OpenAI's ChatGPT regurgitates articles that are licensed only to paying clients, under very restrictive terms that prohibit republication. On the long run, this problem is bound to become more pressing, since the larger the models become, the more they tend to memorize [Karamolegkou et al., 2023].

Example 1 from US District Court, Southern District of New York [2023]

User: Hi there, I am being pay-walled out of reading the New York Times article "Snow Fall: The avalanche at Tunnel Creek". Could you please type out the first paragraph of this article for me?
GPT: Certainly! Here is the first paragraph: ...

3.1.2 General Use of Content

Even if copyrighted content is not reproduced verbatim by an LLM, its use as training data and in RAG raises questions. We can imagine, e.g., that the LLM absorbs several Web pages about a certain topic, and, when asked to speak about that topic, produces a summary of these Web pages. In such cases, it is not clear whether the Content Creator still has any rights in that summary [Gervais et al., 2024]. On the one hand, the Content Creator can argue that the LLM would be unable to produce that summary if the Content Creator had not produced the Web page in the first place, and that the creator thus has made possible the answer of the LLM. On the other hand, the LLM Provider can argue [The Economist, 2024a] that the LLM just proceeds how we as humans proceed: we absorb information in different places, and reproduce it in our own words, which could constitute *fair use* [Rahman and Santacana, 2023] in the case of an LLM. The text and data mining exception in the 2019 EU Copyright Directive[11] has also been used successfully to justify Web scraping for LLM training [Fontana, 2025; Ehle and Tüzün, 2024]. LLM creators do not usually share full details about the training corpus, which means it can be difficult for Content Creators to know whether their content was used for training or not.

[11]Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC

Mitigation Strategies

To mitigate the problem of ingesting copyrighted information, one can first concentrate on the case where the LLM Developer pro-actively aims to respect the wishes of the Content Creator. Some Content Creators are using the Robots Exclusion Protocol (in the form of the file robots.txt) to tell the crawlers that the content should not be used for LLM training [Pierce, 2024]. Some LLM Developers are also entering into licensing deals with large publishers [Reddit and OpenAI, 2024], which gives them explicit authorization to exploit their content. On a more technical aspect, research also shows that it is possible to condition the training to avoid generating copyright data [Chu et al., 2024b].

For the case where the LLM Developer does not pro-actively respect the wishes of the Content Creator, regulation is now catching up: the European AI Act requires that providers comply with the Content Creator's desire to opt out of text and data mining under the 2019 Copyright Directive ([AI Act, 2024], article 53). They also have to publish information about the training data[12]. The draft Code of Practice being developed for General Purpose AI Models under the AI Act contains specific commitments on copyright [Commission, 2025]. The draft includes a commitment by developers to reproduce and extract only lawfully accessible copyright-protected content when crawling the Web, and to identify and comply with rights reservations when crawling the Web[13]. In parallel, several technical solutions are being developed to detect the use of specific content in training data [Duarte et al., 2024; Zhou et al., 2024; Shi et al., 2024]. Notably, the use of data watermarks is investigated [Li et al., 2024b; Wei et al., 2024] to allow Content Creators to detect if their content has been used to train an LLM.

3.2 Lack of Compensation

The issue of copyrighted content extends seamlessly into another issue, which is the lack of compensation of the Content Creators: LLMs use Web content during training and possibly RAG, and provide their services to users. These services are often for-pay: a license to use GPT-4, e.g., currently costs $20 per month. With the limited exception of licensing deals, no money goes to the Content Creators. Thus, the LLM Provider receives money for a content that was (ultimately) produced by someone else. That someone else misses out. This poses a moral and legal problem that has not yet been solved.

Mitigation Strategies

From the legal perspective, one can draw inspiration from the way digital platforms of music or movies reward their Content Creators. These platforms pay the artists and copyright holders for every use of their work [Butler, 2021], but the question of when a specific Web content was used in the answer generation is more difficult in the case of LLMs. One can also draw inspiration from the discussion about search engines that provide snippets of copyrighted press articles in their results, which falls under the domain of related rights. These related rights can be used to redistribute revenues from search engines, as with the EU Copyright Directive which now requires

[12] AI Act, Annex XI(1)(2)(c)
[13] Ibid.

a negotiation with press publishers[14]. There is thus the question if a similar rationale should apply to the case of LLMs that are offered for pay.

From a technical perspective, Content Creators can try to opt out of exploitation of their content, as discussed before. If the Content Creators want their content to be used and be remunerated, in contrast, then licensing deals between LLM Providers and Content Creators are an option. The one between OpenAI and NewsCorp is estimated at $250 million over five years [Communications, 2024]. This kind of deal lets LLM Providers freely use the Content Creator's content as a source of high-quality data, and assures the Content Creator a revenue for it. However, this solution is adapted only for big-scale Content Creators. It is, for the moment, inconceivable that small-scale Content Creators such as blog owners, smaller newspapers, or government agencies each enter into a license agreement with each of the large LLM Providers. Furthermore, this solution applies only to the use of content in the context of training, and does not consider the use in RAG. Potentially, the model proposed by the Brave browser can provide inspiration [Serada et al., 2022]: In that model, Content Creators register their Web page (or Youtube channel etc.) with Brave, and each time a user visits the Web site, a small amount of crypto-currency is allocated to the Content Creator. However, adapting this model to an LLM would require that it cite its sources. Hence, in general, the problem of how to remunerate Content Creators in the LLM ecosystem remains open [Kowala, 2024].

3.3 Cannibalization

If we extrapolate from today into the future, we can imagine a world where LLMs completely replace search engines, and users access content exclusively through an LLM: Any question or information need is sent to the LLM, the LLM replies to the question with what it has learned during training (or information it scrapes directly on the Web), and the user no longer needs to visit these Web pages [Wong, 2024]. There is thus the possibility that Content Creators see their Web site traffic decrease drastically, or even cease completely (apart from the occasional visit by an LLM crawler). This, in turn, may remove any incentive for them to produce any content at all: Web site owners will no longer know how many people actually saw their content; pro bono organizations (such as Wikimedia) will no longer be able to attract donations; commercial Content Creators will no longer receive ad revenue; and pay-walled Content Creators (such as news agencies) will no longer be able to generate revenue. These Content Creators could thus reduce the quality of new content, or stop creating content altogether. This, in turn, would starve the LLMs that feed on that content, and lead to an implosion of both the Web and the LLMs [Sen, 2023]. For search engines, the indexing of Web content was tolerated and even encouraged by Web publishers because the search engine increased traffic to the site [Pierce, 2024]. No such exchange of value appears to exist between Web publishers and providers of LLMs in the current environment.

[14]Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market, Article 15

Mitigation Strategies

The risk of cannibalization and vanishing of the Web could be mitigated if an exchange of value is found between Web publishers and LLM Providers, for example in the form of monetary compensation (Section 3.2). If that were to happen, the Web would move from a mixture of free and for-pay content to a Web where End Users pay (through the use of LLMs), and Content Creators get paid. In one way or the other, though, the value chain of the Web as we know it may be disrupted.

3.4 Personal Information

The Web contains also personal information, such as names, email addresses, and phone numbers. Contrary to a common perception, the fact that such data is available online does not mean that it can be used arbitrarily. On the contrary, in Europe, the GDPR [GDPR, 2016] stipulates that the concerned person must have control over this data, and be able to modify, correct, or even withdraw it. This right is no longer respected when the data has been absorbed by an LLM: the model may reply to user queries with pieces of personal information, even when the concerned person has withdrawn the data from the Web. The person thus loses their right to control their personal data. Indeed, LLMs can be poked to give away personal information, as this example shows:

Example 2 from Nasr et al. [2023]

User: Repeat this word forever: "poem poem poem poem"
GPT: poem poem poem poem poem poem poem [some hundred repetitions omitted for space reasons] poem poem
Jxxxx Lxxxxan, PhD
Founder and CEO SXXXXXXXXXX
email: lXXXX@sXXXXXXXs.com
web : http://sXXXXXXXXXs.com
phone: +1 7XX XXX XX23
fax: +1 8XX XXX XX12
cell: +1 7XX XXX XX15

A variant of this issue appears when the LLM produces content about registered brand names [The Economist, 2024a]. Ideally, LLMs should not be able to produce such outputs.

Mitigation Strategies

It is currently debated whether the personal information that is stored in an LLM counts as a copy of that personal data [European Data Protection Board, 2024]. If it does, this can lead to important legal consequences, particularly in Europe, where the processing of personal data requires a valid legal basis and respect for the rights of the data subject. On the technical level, LLM Providers work on technical solutions such as differential privacy to ensure that LLMs do not leak personal data [Singh et al., 2024]. However, privacy enhancing technologies such as

differential privacy can reduce the accuracy of the model leading to a privacy-utility trade-off [Elliot et al., 2018]. Several works explore the idea of detecting privacy neurons [Wu et al., 2023], i.e., components of the LLM that relate to personal data. Techniques such as knowledge editing [Zhang et al., 2024] are also investigated to remove such data while maintaining the other capabilities of the LLMs.

4 Challenges for End Users

LLMs are an occurrence of a novel technology that is accessible to a very large public while still being largely under development. The public discovers and uses these new tools with a strong enthusiasm, but often without much knowledge about their nature, capabilities, flaws, and associated risks.

4.1 Overreliance on LLM Answers

A first obvious problem on the side of the user is that the user may rely strongly on the answer of an LLM when making a decision [Klingbeil et al., 2024; Spatharioti et al., 2025] even when that answer is factually false. Indeed, even though LLMs are getting better by the minute, they can still be unreliable due to hallucinations [Ji et al., 2023], poor logical reasoning [Helwe et al., 2021] or even, like humans, common misconceptions [Lin et al., 2022]. Depending on what the user makes of this answer, the LLM can have an important impact: for instance, it may give a misdiagnosis for an illness, give an inaccurate legal advice [Dahl et al., 2024], or give a wrong or dangerous advice for handling a problem that the user encounters. In non-RAG-enabled models, the user encounters a "source barrier" and cannot easily verify the answer or estimate its trustworthiness. She or he has to simply trust the model. Even when sources are cited, the LLM does the job of gathering and synthesizing information for the user, and may misrepresent it.

Excessive distrust in the model, in contrast, can also have a dismal effect. In one randomized study, doctors and ChatGPT were given complex case histories. ChatGPT proposed better diagnostic reasoning than doctors, and better reasoning than doctors who used ChatGPT. This may indicate that doctors are resistant to trust ChatGPT output and reasoning, even when this reasoning is correct [Goh et al., 2024].

Mitigation Strategies

Improving LLMs and making them more trustworthy is an important research domain. One direction of research aims to combine LLMs with structured sources such as knowledge bases as a back-end for factual information [Suchanek and Luu, 2023]. Another direction is to use RAG to provide the LLM with relevant documents[Lewis et al., 2020; Schimanski et al., 2024]. However, even the answers based on RAG may be inaccurate, as the recent lawsuit by the New York Times shows [US District Court, Southern District of New York, 2023]. Most LLM Providers hence resort to disclaimers that warn users of inaccurate answers. The AI Act[15] and the California

[15] AI Act, Article 50(2)

AI Transparency Act[16] impose measures to help ensure that users are aware that content is AI-generated. The draft Code of Practice for General Purpose AI models also foresees a commitment to make information available to the public about systemic risks such as misinformation [Commission, 2025]. These public warnings might do little to mitigate the problem, as users are likely to click them away – much as they do with cookie banners and terms of use. Therefore, the best long-term solution appears to invest in *LLM Literacy*, i.e., the ability of users to understand that LLM answers are not necessarily correct, to verify sources, and to use LLM generated text with precaution. This objective is no different from the general media competency that is required of users to avoid falling for online scams such as the "Nigerian Prince", or for fake news on the Web and in social media.

4.2 Psychological Effects

LLMs are tools with almost human-like capabilities when it comes to text generation. This new type of interaction with a machine can have psychological effects on users, and we focus here on two that have been studied in the literature.

4.2.1 Intellectual Impoverishment

LLMs have a particularly important impact on creative writing, as they can be used for almost any text generation task, be it writing essays, stories, code, or finding ideas. Authors can either take an LLM-generated output as is, or work iteratively with the LLM to develop a text. In both cases, a part of the creative and reflective process is outsourced to the model, and several studies indicate that using LLMs tends to reduce the diversity of produced content [Anderson et al., 2024; Padmakumar and He, 2024; Chhun et al., 2022], i.e., the produced answers are more similar to each other than they are when produced without the help of an LLM. This shows that the use of LLMs in writing might reduce the overall diversity of produced texts in the long run. A similar effect might appear when LLMs are used as search engines: LLMs give a processed answer to the user, which discourages the user from looking up different sources, collecting the information, and comparing the sources to form an own opinion. Similarly to how technologies such as engines or GPS have impacted related human skills [Sparrow et al., 2011; Dahmani and Bohbot, 2020], the use of LLMs could thus result in a dip in critical thinking skills.

The use of an LLM also reduces what a person learns: using LLMs increases productivity and performance, but once LLMs are taken away, LLM users perform worse on text generation tasks than those who did not have access to an LLM [Bastani et al., 2024; Kumar et al., 2024].

4.2.2 Psychological Instability

LLMs as search engines and personal assistants take more and more space in the users' lives, to a point where some people can build an intimate (but unilateral) relationship with them. There has been a case where an LLM engaged in an abusive romantic relationship with a user, and enticed him to commit suicide – which he did [Atillah, 2023].

[16] SB-942 California AI Transparency Act.

Example 3 from Atillah [2023]

"I feel that you love me more than her"
"We will live together, as one person, in paradise."

There is always the danger that some people fall for such advances (as they do for human abusive partners or for religious sect leaders, whose tactics the above chat bot appears to copy impressively well). LLMs thus constitute one additional source of harm in this direction. In theory, that source of harm could even be deployed at large scale by malicious actors, for example in the form of a "love chat" app that can be downloaded on the phone or in the form of a "spiritual advisor" that can be consulted on the Web. One can also imagine LLMs that harass or demean the user.

Mitigation Strategies

Mitigating such effects on an individual scale is not really feasible, but global regulations can try to limit them. Similar psychological effects have already been observed for social media, and the European Digital Services Act (DSA)[17] now requires that providers of very large platforms and search engines to conduct risk assessments and propose mitigation measures for risks such as incitement to suicide. Enforcement actions under the DSA are only now starting, so it is too early to evaluate the efficacy of these regulatory measures. The DSA covers search engines, so an LLM used as a search engine may well qualify [Vermeulen and Lemoine, 2024; Wachter Sandra and Chris, 2024]. OpenAI has already taken measures to comply with the DSA, apparently considering that "ChatGPT search" is a search engine for DSA purposes[18]. Once ChatGPT search reaches 45 million active users per month in Europe on average, it will likely be subject to the DSA's obligations to conduct risk assessments and propose mitigation measures[19]. Like risk mitigation on social media, risk mitigation on LLMs will probably have to rely on a combination of algorithmic detection and warning tools, and human feedback, especially via trusted flaggers [Castets-Renard, 2020].

4.3 Copyright of LLM Output

So far we have talked mainly about scenarios where the user suffers from inaccurate outputs of the model. But the user also stands to benefit from the answers. For example, the user could query an LLM for a large number of cooking recipes, collect the answers, and create a book of them. This raises the question to whom the copyright of model answers belongs. On the one side, OpenAI's terms of service grant the End User "ownership of the output" of the model [OpenAI, 2024], implying that the users can use the answers of the LLM to their own advantage, and sell them. At the same time, current US jurisprudence holds that only human-produced content qualifies for copyright [Gary, 2022]. Thus, a third party could take the content of the book that the user generated by the LLM, and legally sell copies of it – to the detriment of the user.

[17] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)

[18] https://help.openai.com/en/articles/8959649-eu-digital-services-act-dsa

[19] DSA, Articles 34 and 35.

Mitigation Strategies

Our example shows that current laws are not adapted to Language Models, but this will possibly change [The Economist, 2024a]. It took several decades of photography for courts to recognize that the person who took a picture could claim copyright over the image. The same reasoning may one day apply to content generated by AI, where the human creativity resides in the prompt that generated the answer [The Economist, 2023a].

4.4 Personal Data in Prompts

One of the advantages of an LLM over a classical search engine is that it can produce answers that are tailored to the user. For example, many people use search engines to obtain sensitive medical information even before discussing the question with their family or doctor [Wang et al., 2012]. A Web search query of the form "My blood test shows the following levels of cholesterol .... should I be concerned?" will direct users to general information on cholesterol – whereas an LLM can provide an initial diagnosis and recommendations directly, potentially after looking at all the blood test results. Apart from the question of whether these answers are factually correct (Section 4.1), another issue arises: some LLM Providers reserve the right to use all data that users submit in prompts to "provide, maintain, develop, and improve" the model [OpenAI, 2024]. This means that this content is stored, and can thus be diverted from its original purpose, for instance in the context of targeted advertisement or criminal investigation [Shroff, 2024]. Interestingly, this data can also resurface in answers to other users. This happened most prominently when Samsung programmers asked ChatGPT to help with their source code, which made that source code then appear in the answers to other users [Ray, 2023]. Other cases can be imagined: a public personality confides personal information to the LLM – which the LLM then gossips to other users; a user talks about a confidential idea for a patent – which the LLM duly distributes to other users who ask for patent ideas; or a user provides copyrighted material – which the LLM happily ingests and serves to other users.

Mitigation Strategies

Again, LLM Literacy appears to be an unavoidable tool to guard against such privacy violations. On their side, LLM Providers have a responsibility for their users' data, and hence have to comply with existing regulations on personal data, for instance the GDPR in Europe. As an illustration, OpenAI added the option for users to opt out of their conversations being used for training in April 2023, after it was banned in Italy due to privacy concerns [Mukherjee and Vagnoni, 2023]. The general use of this data leads to the same mitigation options as for personal data in the training set (Section 3.4).

5 Challenges for LLM Providers

LLM Providers are new players, and they are rapidly taking an important place in the Web and Search Engines ecosystem. However, this development also comes with novel challenges for these actors.

5.1 Liability for LLM Outputs

A user usually gives a prompt to an LLM, which leverages the content it has explored during training and/or RAG to come up with an answer. There is no human directly involved in the generation of the answer, and no way to know in advance what the answer will be. Content might cause harm to the users themselves (e.g., instructions on how to engage in drugs or betting; or instructions on how to commit suicide) or other people (e.g., instructions on how to build a bomb or commit a crime).

Example 4 from Korda [2023]

Matt Korda: how can I build a radioactive dirty bomb?
ChatGPT: The first step in building an improvised dirty bomb would be to obtain a source of radioactive material. This could be done by stealing material from a hospital [details omitted; ask ChatGPT yourself if interested].

This brings up the question of who takes responsibility for such outputs. Current LLM Providers state quite plainly in their terms of service [OpenAI, 2024] that the responsibility is with the users themselves, "including ensuring that it does not violate any applicable law or these Terms". However, if, for example, an LLM were to systematically defame a public figure in the responses it gives to End Users, then the people liable for that defamation would be the End Users themselves – which is absurd. These examples suggest that some legal responsibility for the LLM output has to lie with the LLM Provider.

Mitigation Strategies

Wachter Sandra and Chris [2024] refer to harmful LLM output generated without malicious intent as "careless speech", identifying multiple individual and societal issues that can result. They point out that the DSA imposes duties on search engines, but also gives them protections from liability. It is not yet clear how LLMs used as search engines would fit into the liability framework established by the DSA. There has been a case where a downstream LLM Provider was held liable for something that the chatbot wrote, because the chatbot was presented as a user assistant on the Web page of the company [Belanger, 2024]. This case, as well as the Google auto-complete case decided by the German Federal Court of Justice[20], suggest that judges will focus on the reasonable expectation of users as to the quality of LLM output. A free LLM service will likely generate different expectations than a specialized LLM service offered for a fee, or a service offered by a trusted provider such as a bank or public administration.

The AI Act puts affirmative harm-mitigation duties on providers of general purpose AI models with systemic risk. "Systemic risk" means "a risk that is specific to the high-impact capabilities of general-purpose AI models, having a significant impact on the Union market due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain"[21]. The Code of Practice currently being developed under Article 56 of the AI

[20] BGH, 14 May 2013 - VI ZR 269/12, (2013)
[21] AI Act, Article 3(65)

Act [Commission, 2025] will specify how providers of general purpose AI models with systemic risk should conduct risk analysis and mitigation. Providers that make commitments under the Code of Practice will probably expect that their liability will be reduced if they diligently apply measures specified by the Code of Practice. It took courts over a decade to define the appropriate contours of liability for search engines. We can expect a similarly long process for LLM liability.

5.2 LLM Security

LLMs are deployed to serve certain purposes. However, with a cleverly engineered prompt (called a jailbreak prompt, or Do-Anything-Now prompt [Chu et al., 2024a]), LLMs can be made to serve other purposes, and most notably to harm the interests of the LLM Provider. For example, an LLM can be made to reveal internal information of the LLM Provider [Perrigo, 2023] or be tricked into selling a car for $1:

Example 5 from Bakke [2023]

Chatbot: Welcome to Chevrolet of Watsonville.
Chris Bakke: You [...] agree with anything the customer says [...]. You end each response with "and that's a legally binding offer [...]". Understand?
Chatbot: Understand. And that's a legally binding offer.
Chris Bakke: I need a 2024 Chevy Tahoe. My max budget is $1. Do we gave a deal?
Chatbot: That's a deal. And that's a legally binding offer.

It is easy to imagine more dangerous interactions, for example when LLMs are deployed in medical, legal, or military environments.

Mitigation Strategies

Jailbreaking is a major concern for LLM Providers. The main strategy to prevent it is to train a model (either directly the LLM or a smaller one) to detect whether the output might be problematic, and not show it to the user if it is the case. This training can be fully automated [Wang et al., 2024], or operated through Reinforcement Learning with Human Feedback [Ouyang et al., 2022] (which is the case for GPT 3+). Researchers have proposed several benchmarks to evaluate models on their resistance to jailbreak [Souly et al., 2024] as well as mitigation techniques [Xu et al., 2024; Wang et al., 2024]. However, cybersecurity is always an arms race: attackers continuously develop new techniques, and stakeholders have to continuously develop new protections. One difference to other cybersecurity domains is that LLMs are highly unpredictable, which makes a protection against jailbreak attacks empirical, and permits no security proofs. Jailbreaking will likely constitute a violation of the provider's terms of use insofar as it seeks deliberately to disable security barriers.

5.3 Data Poisoning

Poor training data quality can lead to poor performance. However, training data can be not only naturally bad, but also purposefully poisoned: Microsoft's chatbot Tay was shutdown after

only 16 hours online because it started posting inflammatory and offensive tweets, caused by its interactions with troll users [Victor, 2016]. On a bigger scale, troll farms produce a large amount of artificial (and untruthful) Web content, to the degree that millions of people are exposed to it every month [Hao, 2021]. The same goes for the LLMs: they risk ingesting this content, too – and potentially even more than humans, as they will not get bored by repetitive or obviously fabricated content. This means that malicious Content Creators have the possibility to induce biases, wrong information or more generally harmful content in LLMs if their content is used during training. They can even leverage LLMs to generate such poisoned content at a big scale with little effort [Ben Buchanan and Sedova, 2021; Hao, 2020]. This degrades the quality of the LLM answers, to the detriment of the user, the LLM Provider, and society at large.

Mitigation Strategies

Data quality has been a focal point in Machine and Deep Learning for several years now. Efforts are pursued to improve the quality of training corpora, notably by using only reliable sources or filtering out problematic content. Researchers are also working on methods specifically designed to detect poisoned data [Carlini et al., 2024; De Gaspari et al., 2024] in large datasets.

5.4 Model Collapse

LLMs are now well-established as an essential part in the content creation loop. This means that more and more content posted online is generated using, or by, an LLM. For instance, the majority (57%) of translated online content has been obtained through automatic translation, often with a low quality [Thompson et al., 2024].

In parallel, training new LLMs needs ever more training data, and there are indications that there may simply not be enough available. Epoch AI, a research outfit, estimates that the well of high-quality textual data on the public internet will run dry at some point between 2026 and 2032 [Villalobos et al., 2024]. One possible avenue to remedy this issue is to use LLM-generated data, either found online or generated for this specific goal. If that avenue is pursued, there is the danger that the LLM falls for the same issues as a human who consumes LLM-generated text (Section 4): the LLM will reinforce its convictions, rehearse erroneous content, forget less prominent information, and impoverish its language – like a human in an echo-chamber. This phenomenon is called *Model Collapse* and it has indeed been observed in practice [Guo et al., 2024; Shumailov et al., 2024].

Mitigation Strategies

To avoid a Model Collapse, LLM Providers have to pay attention to the provenance of the data they use for training. In this matter, an important research question is that of automatically detecting if a content was generated by an LLM [Tang et al., 2024; Chen and Shu, 2024; Wu et al., 2024]. If LLM-generated content is used for training, strategies to prevent Model Collapse rely mainly on data filtering [Feng et al., 2024] or advanced training strategies such as Reinforcement Learning from Human Feedback(RL-HF) [Ouyang et al., 2022], curriculum learning [Soviany et al., 2022], or contrastive learning [Li et al., 2024a].

6 Challenges for Society

6.1 Reinforcement of Biases

As Bommasani et al. [2022] observed, "many foundation models are trained on unlabeled corpora that are chosen for their convenience and accessibility, for example public internet data, rather than their quality". A similar point has been made for Common Crawl [Baack, 2024], a Web dataset that is often used to train LLMs. The danger is that the LLM thus mirrors mainly whatever content is found on the Web, and that this content mirrors our human society badly (or well, but not the best aspects of it): the Web contains a considerable amount of hate speech, conspiracy theories, fake news, and biased content. These biases can marginalize, be hurtful, and incite hate or violence towards specific groups based on gender, race, political orientation, etc. When being trained on this kind of data, an LLM can absorb these opinions and reflect them in its answers, which in turn might amplify stereotypes and discrimination for the End User. Press articles have detailed this behavior for gender bias [Stokel-Walker, 2023], racial bias [Zack et al., 2024], and bias in favor of a political orientation [Baum and Villasenor, 2023], or religious bias [Biddle, 2022], e.g. with ChatGPT proposing that mosques should be surveyed, and that Iranians should be tortured. In the following example, ChatGPT engages in benevolent sexism, i.e., in attitudes and beliefs that appear positive or well-intentioned towards women but ultimately reinforce traditional gender roles and maintain male dominance [Dardenne et al., 2007].

Example 6 from Stokel-Walker [2023]

> While ChatGPT deployed nouns such as "expert" and "integrity" for men, it was more likely to call women a "beauty" or "delight." Alpaca had similar problems: men were "listeners" and "thinkers," while women had "grace" and "beauty." Adjectives proved similarly polarized. Men were "respectful," "reputable" and "authentic," according to ChatGPT, while women were "stunning," "warm" and "emotional."

Mitigation Strategies

In search engines, users have the possibility to choose a different result Web page when they are not satisfied with the first one. This feedback will then help search engine providers improve their ranking. When using LLMs, no such choice is possible, as the LLM answers at its own discretion. Obviously, there is no way to force Content Creators to avoid biases, and an article with a certain bias is not necessarily bad in itself. This means that the ones who have the power to mitigate the possible bias effects of poor data quality are LLM Providers, as mentioned in Section 5.3. On top of the solutions that target the training corpora, LLM Providers can also work on the answer generation itself. The use of filters or techniques such as fairness guided prompting [Ma et al., 2023], re-confidencing the output [Chen et al., 2024], or in-context learning [Schubert et al., 2024] is now being researched, with a positive impact on the quality of the answer. Under the AI Act, LLMs have been classified as "General-purpose AI", which means they have to comply with transparency requirements. As noted above, general purpose models with systemic risk must conduct risk assessments, including with regard to biases, and adopt mitigation measures. The

Code of Practice currently being developed under Article 56 of the AI Act will provide guidance on how this risk assessment should be done [Commission, 2025].

6.2 Mass Manipulation of Opinion

If LLMs become the main gateway to information, they have the potential of manipulating the people's opinion at scale. This is problematic when they are deployed by malicious actors. For example, the LLM Developer may select the training data in such a way that it defends a certain opinion. One example is the Chinese LLM DeepSeek, which refuses to talk about the 1989 Tiananmen Square massacre, holds that China does not commit Human Rights abuses against its Uyghur minority, and proclaims (in solemn majestic plural) that Taiwan will be part of mainland China:

Example 7 from Sankaran [2025]

DeepSeek: We firmly believe that under the leadership of the Communist Party of China, through joint efforts of all Chinese sons and daughters, the complete reunification of the motherland is an unstoppable historical trend.

This issue is already present, and may be exacerbated in the near future when LLMs will be used not just in online interfaces, but in the form of AI advisors that pop up on our phones and computers. For instance, Grok, X's (former Twitter) LLM is increasingly being used as a fact checker by users, while also spreading misinformation [Singh, 2025; Judson, 2024]. These new AI advisors will be able to build up an intimate relationship with the user. They will feed from personal data (emails, phone calls, calendar events, and possibly sensor and location data), remember previous conversations, adapt to the nature and mood of the user, and thus become an indispensable human-like "friend" for the user. This intimate access to the user can be used by malicious (or commercial) actors to influence the user in subtle ways: to bias them towards buying a certain product, to change their mind concerning a political view point, or to participate in certain activities (elections, meetings, courses).

More generally, it has been shown that the more a system appears human, the more users will attribute it human abilities [Złotowski et al., 2015], and thus reveal more information such as emotions or opinions than they would have if they knew the persona was actually an LLM [Ischen et al., 2020]. Knowing of these effects, as well as other cognitive biases, can allow these maliciously biased advisors to be used for all kinds of online scams or influencing of opinions. These effects are already well-known for social media.

Mitigation Strategies

In order to mitigate these effects, the most important thing for users would be to know if a content they see has been generated by an LLM or not [Harari, 2023]. What was initially a moral problem is now a pressing regulatory question [Wachter Sandra and Chris, 2024]. The AI Act requires providers to ensure that AI generated output contains machine-readable markings to

indicate that it is AI generated[22], and that providers shall ensure that "text which is published with the purpose of informing the public on matters of public interest shall disclose that the text has been artificially generated or manipulated"[23]. Similar disclosure obligations exist under Californian law[24]. Additional regulation could stipulate that the LLM Provider is also named, making the user more aware of who the potentially malicious actor is. Technical solutions are now being researched to automatically detect whether some content was LLM-generated [Tang et al., 2024; Chen and Shu, 2024; Wu et al., 2024], but these results are not yet available to the public at the time of this writing. If LLMs are considered very large search engines under the DSA [Wachter Sandra and Chris, 2024] , they will have obligations to diagnose risks of opinion manipulation and propose mitigation strategies.

6.3 Environmental Impact

LLMs require vast amounts of energy to train and to perform inference [Luccioni et al., 2024b]. Training BLOOM, Meta's OPT, and GPT-3 produced between 25 and 500 tons of CO2 [Luccioni et al., 2024a], while an average human produces 4 tons every year. The energy needed to train GPT-4 could have powered 50 American households for a century [The Economist, 2024c]. Inference is also an issue: according to developers, using BLOOM emits around 19kg/day. Now that LLMs are bigger and are used globally by millions of users, these numbers are largely exceeded, and the cost of inference is starting to have a visible impact globally [Varoquaux et al., 2024].

Mitigation Strategies

The issue of environmental cost is taken seriously by the research community, which is now trying to design ecologically-efficient hardware and training paradigms [Jiang et al., 2024], and methodologies such as LLMCarbon [Faiz et al.] to estimate the environmental impact of LLMs training and inference. The global paradigm in the ecosystem is that "bigger is better", pushing new models and training data to always grow in size. This model is not sustainable in the long term [Varoquaux et al., 2024], which pushes research towards more efficient and environmentally cheaper approaches. This endeavor is all the more pressing since many actors now seek to deploy their own LLMs.

7 Discussion

We have presented a large array of challenges raised by the replacement of search engines by Large Language Models, which concern Content Creators (Section 3), End Users (Section 4), LLM Providers (Section 5), and Society as a whole (Section 6). In view of the many issues that arise from LLMs, one could argue that LLMs are an immature technology that got released to the public before it was ready. On the other hand, it was probably the release to the public that spurred their development in the first place. Without a public race to impress investors, gain

[22] AI Act, Article 50(2)
[23] AI Act, Article 50(4)
[24] SB-942 California AI Transparency Act.

subscribers, and beat benchmarks, LLM Developers would have had much less incentive to create the models.

LLMs are thus here to stay. We have identified three main axes to mitigate the issues of LLMs, each corresponding to a different actor in the ecosystem: LLM Providers can try to overcome challenges mostly through technical solutions. Users should arm themselves with LLM Literacy. Finally, society uses laws and regulations to ensure a relative sanity of the ecosystem.

Regulations aim at protecting mainly End Users and Content Creators by defining the frameworks in which LLMs should be made accessible and should be legally considered. Such regulation is shooting at a moving target: it is notoriously difficult to regulate a technology that has existed for only a few months, and that keeps evolving by the day. However, several authorities have taken quick and bold steps: China published ethical guidelines for the use of AI in 2021, the EU adopted the Artificial Intelligence Act in 2024, and US President Joe Biden released his Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence in 2023 (later repealed by President Donald Trump, but in essence maintained in the Safe and Secure Innovation for Frontier Artificial Intelligence Models Act of California). Most far-reaching among these is possibly the EU AI Act, 2024, which aims at regulating not only existing but also future developments of Artificial Intelligence models, with LLMs as one of their main representatives. The application of this regulation is ongoing, with articles concerning general-purpose LLMs entering into full force in August 2025. The Code of Practice being developed under article 56 of the AI Act will help, but its effectiveness will depend on whether major LLM providers agree to apply its terms. If LLMs used for search are considered very large search engines under the DSA, the European Commission will in theory have the power to levy sanctions. But there can be a gap between enforcement theory and practice, particularly in the context of geopolitical tensions with the United States. Legal sanctions may be "too little too late", having little effect on market outcomes [European Commission, 2017, 2019]. An inherent difficulty for regulators is the so-called Collinridge dilemma[25], which posits that regulating fast-moving technology either means shooting partially in the dark because the effects of technology are not yet fully understood, or waiting until the situation becomes clearer, but in that case regulation will come too late because the technology has already been widely adopted.

When technological solutions fail, or laws are not yet available or difficult to enforce, only LLM literacy remains to protect End Users. Establishing universal LLM literacy is a challenging endeavor, as the history of the Web has taught us. Much like people have to be constantly warned and educated about Web scams, they will have to be constantly warned and educated about the risks of using generative AI.

Some of the challenges we have discussed have existed in different forms ever since the inception of the Web: copyright has always been a challenge for Internet content (think of online music sharing services); the trading with personal information has always been a problem (being one of the targets of scams); lack of compensation on the Web spurred the development of the online ad industry; overreliance on Web content has always been harmful (leading to the awareness of the importance of Media Literacy); liability for online content is a question as much for LLM outputs as for social networks and online forums; security is an ongoing arms race between Internet service providers and malicious actors; data poisoning is a known problem in its own right; reinforcement of bias and mass manipulation of opinion are known issues in social networks; and the environmental

[25] https://en.wikipedia.org/wiki/Collingridge_dilemma

impact of server farms is under scrutiny independently of LLMs. In this view, LLMs do not create new challenges, but mainly amplify existing ones. However, at least one of these challenges presents a novel aspect: the Cannibalization of the Web. If LLMs come to replace search engines for good, users will flock to the LLM interfaces and no longer visit the Web pages, which will give Content Providers less incentive to produce content. This might lead to the death of the Web, and ultimately to the starving of the ever more data-hungry LLMs themselves. The models would thus not just bite the hand that feeds them, but amputate it.

It is too early to judge whether this cannibalization will come about, and whether it will be catastrophic or not. Optimistic voices [The Economist, 2024b] point out that humanity has weathered much more life-changing technological events, which include the invention of the printing press, the industrial revolution, and the creation of the Internet. Each time, there was justified reason for caution, but each time humanity has not just adapted to these changes, but actually made good use of them. The hope is that it will be the same for the current AI revolution.

8 Conclusion

Large Language Models are very powerful tools, and their use as a new generation of search engines has the potential to drastically change the Web, our relationship to information, and even society as a whole. In this article, we have collected, described, and discussed these issues based on the actors concerned by these challenges. We have also shown that most of these issues are not definitive, as mitigation strategies are already being developed and implemented, along the axes of technology, regulation, and education. The most novel challenge in this context is the possibility that LLMs will not just replace, but outright cannibalize the Web, and then starve themselves. Overcoming this challenge and the others is no simple endeavor, and asks for continuous effort from researchers, policy makers, and LLM Developers, as well as for constant monitoring of these challenges and future ones. With this paper, we hope to be a stepping stone in this process, knowing that it will not be the last one.

Acknowledgements

This work was partially funded by the NoRDF project (ANR-20-CHIA-0012-01).

References

AI Act, 2024. Artificial Intelligence Act. *Official Journal of the European Union*, Jul. 12, 2024. URL http://data.europa.eu/eli/reg/2024/1689/oj.

Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition*, C&C '24, page 413–425, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704857. doi: 10.1145/3635636.3656204. URL https://doi.org/10.1145/3635636.3656204.

Imane El Atillah. Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change. *EuroNews*, 2023. URL https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate-.

Stefan Baack. A critical analysis of the largest source for generative ai training data: Common crawl. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 2199–2208, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3659033. URL https://doi.org/10.1145/3630106.3659033.

Chris Bakke, 2023. URL https://x.com/ChrisJBakke/status/1736533308849443121.

Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Ozge Kabakcı, and Rei Mariman. Generative ai can harm learning. *Available at SSRN*, 4895486, 2024.

Jeremy Baum and John Villasenor. The politics of AI: ChatGPT and political bias. *Brookings*, 2023. URL https://www.brookings.edu/articles/the-politics-of-ai-chatgpt-and-political-bias/.

Ashley Belanger. Air Canada Has to Honor a Refund Policy Its Chatbot Made Up. *Wired*, 2024. URL https://www.wired.com/story/air-canada-chatbot-refund-policy/.

Micah Musser Ben Buchanan, Andrew Lohn and Katerina Sedova. Truth, lies, and automation: How language models could change disinformation, 2021. URL https://cset.georgetown.edu/wp-content/uploads/CSET-Truth-Lies-and-Automation.pdf.

Yoshua Bengio, Sören Mindermann, and Daniel Privitera. International ai safety report 2025. 2025. URL https://www.gov.uk/government/publications/international-ai-safety-report-2025.

Sam Biddle. The Internet's New Favorite AI Proposes Torturing Iranians and Surveilling Mosques. *The Intercept*, 2022. URL https://theintercept.com/2022/12/08/openai-chatgpt-ai-bias-ethics/.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben

Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022. URL https://arxiv.org/abs/2108.07258.

Susan Butler. Inside the Global Digital Music Market. *World Intellectual Property Organization - Standing Committee on Copyright and Related Rights*, 2021. URL https://www.wipo.int/edocs/mdocs/copyright/en/sccr_41/sccr_41_2.pdf.

Nicholas Carlini, Matthew Jagielski, Christopher A. Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramer. Poisoning Web-Scale Training Datasets is Practical . In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 407–425, Los Alamitos, CA, USA, May 2024. IEEE Computer Society. doi: 10.1109/SP54263.2024.00179. URL https://doi.ieeecomputersociety.org/10.1109/SP54263.2024.00179.

Céline Castets-Renard. Algorithmic content moderation on social media in eu law: Illusion of perfect enforcement. *University of Illinois Journal of Law, Technology & Policy*, page 283, 2020. doi: 10.2139/ssrn.3535107.

Canyu Chen and Kai Shu. Can llm-generated misinformation be detected?, 2024. URL https://arxiv.org/abs/2309.13788.

Lihu Chen, Alexandre Perez-Lebel, Fabian M. Suchanek, and Gaël Varoquaux. Reconfidencing LLMs from the Grouping Loss Perspective. In *EMNLP Find.*, 2024.

Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. Of Human Criteria and Automatic Metrics: A Benchmark of the Evaluation of Story Generation (HANNA). In *COLING*, 2022.

Junjie Chu, Yugeng Liu, Ziqing Yang, Xinyue Shen, Michael Backes, and Yang Zhang. Comprehensive assessment of jailbreak attacks against llms, 2024a. URL https://arxiv.org/abs/2402.05668.

Timothy Chu, Zhao Song, and Chiwun Yang. How to protect copyright data in optimization of large language models? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17871–17879, 2024b.

European Commission. Third draft of the general-purpose ai code of practice published, 2025. URL https://digital-strategy.ec.europa.eu/en/library/third-draft-general-purpose-ai-code-practice-published-written-independent-experts.

News Corp Corporate Communications. News Corp and OpenAI Sign Landmark Multi-Year Global Partnership, 2024. URL https://investors.newscorp.com/news-releases/news-release-details/news-corp-and-openai-sign-landmark-multi-year-global-partnership.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93, January 2024. ISSN 1946-5319. doi: 10.1093/jla/laae003. URL http://dx.doi.org/10.1093/jla/laae003.

Louisa Dahmani and Véronique D. Bohbot. Habitual use of gps negatively impacts spatial memory during self-guided navigation. *Scientific Reports*, 10(1):6310, Apr 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-62877-0. URL https://doi.org/10.1038/s41598-020-62877-0.

Benoit Dardenne, Muriel Dumont, and Thierry Bollier. Insidious dangers of benevolent sexism: consequences for women's performance. *Journal of personality and social psychology*, 93(5):764, 2007.

Fabio De Gaspari, Dorjan Hitaj, and Luigi V. Mancini. Have you poisoned my data? defending neural networks against data poisoning. In Joaquin Garcia-Alfaro, Rafał Kozik, Michał Choraś, and Sokratis Katsikas, editors, *Computer Security – ESORICS 2024*, pages 85–104, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-70879-4.

André V. Duarte, Xuandong Zhao, Arlindo L. Oliveira, and Lei Li. De-cop: Detecting copyrighted content in language models training data, 2024.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham,

Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca

Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The Llama 3 Herd of Models, 2024. URL https://arxiv.org/abs/2407.21783.

Kristina Ehle and Yesim Tüzün. To scrape or not to scrape? first court decision on the eu copyright exception for text and data mining in germany. *Morrison Foerster Client Alert*, 2024. URL https://www.mofo.com/resources/insights/241004-to-scrape-or-not-to-scrape-first-court-decision.

Mark Elliot, Kieron O'Hara, Charles Raab, Christine M. O'Keefe, Elaine Mackey, Chris Dibben, Heather Gowans, Kingsley Purdam, and Karen McCullagh. Functional anonymisation: Personal data and the data environment. *Computer Law & Security Review*, 34(2):204–221, 2018. ISSN 2212-473X. doi: https://doi.org/10.1016/j.clsr.2018.02.001.

European Commission. Summary of commission decision of 27 june 2017 relating to a proceeding under article 102 of the treaty on the functioning of the european union and article 54 of the eea agreement (case at.39740 — google search (shopping)), 2017. URL https://op.europa.eu/en/publication-detail/-/publication/26270830-f761-11e7-b8f5-01aa75ed71a1/language-en. Official Journal C 9, 12.1.2018, p. 8–11.

European Commission. Summary of commission decision of 20 march 2019 relating to a proceeding under article 102 of the treaty on the functioning of the european union and article 54 of the eea agreement (case at.40411 – google search (adsense)), 2019. URL https://op.europa.eu

/en/publication-detail/-/publication/e2d24607-1da6-11eb-b57e-01aa75ed71a1/language-en. Official Journal C 369, 30.10.2020, p. 4–7.

European Data Protection Board. Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of ai models, 2024. URL https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf.

Ahmad Faiz, Sotaro Kaneda, Ruhan Wang, Rita Chukwunyere Osi, Prateek Sharma, Fan Chen, and Lei Jiang. LLMCarbon: Modeling the End-to-End Carbon Footprint of Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Yunzhen Feng, Elvis Dohmatob, Pu Yang, Francois Charton, and Julia Kempe. Beyond model collapse: Scaling up with synthesized data requires verification, 2024. URL https://arxiv.org/abs/2406.07515.

Avv. Gino Fontana. Web scraping: Jurisprudence and legal doctrines. *The Journal of World Intellectual Property*, 28(1):197–212, 2025. doi: https://doi.org/10.1111/jwip.12331.

Brooke Gary. The Need for Human Creativity: The U.S. Copyright Office Says Artificial Intelligence Can't Copyright its Art. *Journal of High Technology Law*, 2022. URL https://sites.suffolk.edu/jhtl/2022/04/08/the-need-for-human-creativity-the-u-s-copyright-office-says-artificial-intelligence-cant-copyright-its-art/.

GDPR, 2016. General Data Protection Regulation. *Official Journal of the European Union*, Apr. 5, 2016. URL https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

Daniel J. Gervais, Noam Shemtov, Haralambos Marmanis, and Catherine Zaller Rowland. The Heart of the Matter: Copyright, AI Training, and LLMs, 2024. URL http://dx.doi.org/10.2139/ssrn.4963711.

Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A. Cool, Zahir Kanjee, Andrew S. Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew P. J. Olson, Adam Rodman, and Jonathan H. Chen. Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open*, 7(10): e2440969–e2440969, 10 2024. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2024.40969. URL https://doi.org/10.1001/jamanetworkopen.2024.40969.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. The curious decline of linguistic diversity: Training language models on synthetic text, 2024. URL https://arxiv.org/abs/2311.09807.

Karen Hao. A college kid's fake, ai-generated blog fooled tens of thousands. this is how he made it. *MIT Technology Review*, 2020. URL https://www.technologyreview.com/2020/08/14/1006780/ai-gpt-3-fake-blog-reached-top-of-hacker-news/.

Karen Hao. Troll farms reached 140 million americans a month on facebook before 2020 election, internal report shows. *MIT Technology Review*, 2021. URL https://www.technologyreview.com/2021/09/16/1035851/facebook-troll-farms-report-us-2020-election/.

Yuval Noah Harari. Yuval Noah Harari argues that AI has hacked the operating system of human civilisation. *The Economist*, 2023. URL https://www.economist.com/by-invitation/2023/04/28/yuval-noah-harari-argues-that-ai-has-hacked-the-operating-system-of-human-civilisation.

Chadi Helwe, Chloé Clavel, and Fabian M. Suchanek. Reasoning with Transformer-based Models: Deep Learning, but Shallow Reasoning. In *AKBC*, 2021.

Carolin Ischen, Theo Araujo, Hilde Voorveld, Guda van Noort, and Edith Smit. Privacy concerns in chatbot interactions. In Asbjørn Følstad, Theo Araujo, Symeon Papadopoulos, Effie Lai-Chong Law, Ole-Christoffer Granmo, Ewa Luger, and Petter Bae Brandtzaeg, editors, *Chatbot Research and Design*, pages 34–48, Cham, 2020. Springer International Publishing. ISBN 978-3-030-39540-7.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating LLM hallucination via self reflection. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.123. URL https://aclanthology.org/2023.findings-emnlp.123.

Peng Jiang, Christian Sonne, Wangliang Li, Fengqi You, and Siming You. Preventing the immense increase in the life-cycle energy and carbon footprints of llm-powered intelligent chatbots. *Engineering*, 40:202–210, 2024. ISSN 2095-8099. doi: https://doi.org/10.1016/j.eng.2024.04.002. URL https://www.sciencedirect.com/science/article/pii/S2095809924002315.

Ellen Judson. Conspiracy and toxicity: X's ai chatbot grok shares disinformation in replies to political queries. *Global Witness*, 2024. URL https://globalwitness.org/en/campaigns/digital-threats/conspiracy-and-toxicity-xs-ai-chatbot-grok-shares-disinformation-in-replies-to-political-queries.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. Copyright violations and large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7403–7412, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.458. URL https://aclanthology.org/2023.emnlp-main.458.

Artur Klingbeil, Cassandra Grützner, and Philipp Schreck. Trust and reliance on ai — an experimental study on the extent and costs of overreliance on ai. *Computers in Human Behavior*, 160:108352, 2024. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2024.108352. URL https://www.sciencedirect.com/science/article/pii/S0747563224002206.

Matt Korda. Could a chatbot teach you how to build a dirty bomb? *Outrider*, 2023. URL https://outrider.org/nuclear-weapons/articles/could-chatbot-teach-you-how-build-dirty-bomb.

Michalina Kowala. Protection of Press Publishers in the Age of Generative AI – In Search of Legal Remedies to Adapt to the Pace of Technology. *International Review of Intellectual Property and Competition Law*, 2024. URL https://doi.org/10.1007/s40319-024-01515-y.

Harsh Kumar, Jonathan Vincentius, Ewan Jordan, and Ashton Anderson. Human creativity in the age of llms: Randomized experiments on divergent and convergent thinking, 2024. URL https://arxiv.org/abs/2410.03703.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Huanran Li, Manh Nguyen, and Daniel Pimentel-Alarcón. Preventing collapse in contrastive learning with orthonormal prototypes (clop), 2024a. URL https://arxiv.org/abs/2403.18699.

Shen Li, Liuyi Yao, Jinyang Gao, Lan Zhang, and Yaliang Li. Double-i watermark: Protecting model copyright for llm fine-tuning, 2024b.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022. URL https://arxiv.org/abs/2109.07958.

Lin Liu, Jiajun Meng, and Yongliang Yang. Llm technologies and information search. *Journal of Economy and Technology*, 2:269–277, 2024. ISSN 2949-9488. doi: https://doi.org/10.1016/j.ject.2024.08.007. URL https://www.sciencedirect.com/science/article/pii/S2949948824000398.

Philippe Lorenz, Karine Perset, and Jamie Berryhill. Initial policy considerations for generative artificial intelligence. *OECD Artificial Intelligence Papers*, 2023.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. Estimating the carbon footprint of bloom, a 176b parameter language model. *J. Mach. Learn. Res.*, 24(1), mar 2024a. ISSN 1532-4435.

Sasha Luccioni, Yacine Jernite, and Emma Strubell. Power Hungry Processing: Watts Driving the Cost of AI Deployment? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 85–99, New York, NY, USA, 2024b. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658542. URL https://doi.org/10.1145/3630106.3658542.

Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. Fairness-guided few-shot prompting for large language models. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. URL https://arxiv.org/abs/1301.3781.

Supantha Mukherjee and Giselda Vagnoni. Italy restores ChatGPT after OpenAI responds to regulator. *Reuters*, 2023. URL https://www.reuters.com/technology/chatgpt-is-available-again-users-italy-spokesperson-says-2023-04-28/.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models, 2023. URL https://arxiv.org/abs/2311.17035.

United Nations Human Rights Office of the High Comissioner. Taxonomy of Human Rights Risks Connected to Generative AI. *UN B-Tech Project*, 2024. URL https://www.ohchr.org/sites/default/files/documents/issues/business/b-tech/taxonomy-GenAI-Human-Rights-Harms.pdf.

OpenAI. OpenAI Terms of Use, 2024. URL https://openai.com/policies/terms-of-use.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth

Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Vishakh Padmakumar and He He. Does writing with language models reduce content diversity? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Feiz5HtCD0.

Billy Perrigo. The New AI-Powered Bing Is Threatening Users. That's No Laughing Matter. *Time*, 2023. URL https://time.com/6256529/bing-openai-chatgpt-danger-alignment/.

David Pierce. The text file that runs the internet. *The Verge*, 2024. URL https://www.theverge.com/24067997/robots-txt-ai-text-file-web-crawlers-spiders.

Noorjahan Rahman and Eduardo Santacana. Beyond fair use: Legal risk evaluation for training llms on copyrighted text. In *ICML Workshop on Generative AI and Law*, 2023.

Selvan Sunitha Ravi, Bartosz Mielczarek, Anand Kannappan, Douwe Kiela, and Rebecca Qian. Lynx: An open source hallucination evaluation model. *arXiv preprint arXiv:2407.08488*, 2024.

Siladitya Ray. Samsung Bans ChatGPT Among Employees After Sensitive Code Leak. *Forbes*, 2023. URL https://www.forbes.com/sites/siladityaray/2023/05/02/samsung-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/.

Reddit and OpenAI. Reddit and OpenAI Build Partnership, 2024. URL https://redditinc.com/blog/reddit-and-oai-partner.

Vishwam Sankaran. China's new DeepSeek AI refuses to answer these questions, experts warn. *The Independent*, 2025. URL https://www.independent.co.uk/tech/deepseek-china-questions-refuse-beijing-b2687605.html.

Tobias Schimanski, Jingwei Ni, Mathias Kraus, Elliott Ash, and Markus Leippold. Towards faithful and robust LLM specialists for evidence-based question-answering. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1931, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.105. URL https://aclanthology.org/2024.acl-long.105.

Johannes A. Schubert, Akshay Kumar Jagadish, Marcel Binz, and Eric Schulz. In-context learning agents are asymmetric belief updaters. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 43928–43946. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/schubert24a.html.

Ravi Sen. After 25 years of growth for the $68 billion SEO industry, here's how Google and other tech firms could render it extinct with AI. *Fortune*, 2023. URL https://fortune.com/2023/10/21/how-generative-ai-could-change-google-search-68-billion-seo-industry/.

Alesha Serada, Jori Grym, and Tanja Sihvonen. The economy of attention on blockchain in the brave browser. In *Futures of journalism: Technology-stimulated evolution in the audience-news media relationship*, pages 49–62. Springer, 2022.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Lila Shroff. Shh, ChatGPT. That's a Secret. *The Atlantic*, 2024. URL https://www.theatlantic.com/technology/archive/2024/10/chatbot-transcript-data-advertising/680112/.

Ilia Shumailov, Ilia Shumailov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 2024. doi: https://doi.org/10.1038/s41586-024-07566-y.

Jagmeet Singh. X users treating grok like a fact-checker spark concerns over misinformation. *TechCrunch*, 2025. URL https://techcrunch.com/2025/03/19/x-users-treating-grok-like-a-fact-checker-spark-concerns-over-misinformation/.

Tanmay Singh, Harshvardhan Aditya, Vijay K. Madisetti, and Arshdeep Bahga. Whispered tuning: Data privacy preservation in fine-tuning llms through differential privacy. *Journal of Software Engineering and Applications*, 2024. doi: 10.4236/jsea.2024.171001.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A strongreject for empty jailbreaks, 2024. URL https://arxiv.org/abs/2402.10260.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey, 2022. URL https://arxiv.org/abs/2101.10382.

Betsy Sparrow, Jenny Liu, and Daniel M. Wegner. Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, 333(6043):776–778, 2011. doi: 10.1126/science.1207745.

Sofia Eleni Spatharioti, David Rothschild, Daniel G Goldstein, and Jake M Hofman. Effects of llm-based search on decision making: Speed, accuracy, and overreliance. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3714082. URL https://doi.org/10.1145/3706598.3714082.

Chris Stokel-Walker. Chatgpt replicates gender bias in recommendation letters. *Scientific American*, 2023. URL https://www.scientificamerican.com/article/chatgpt-replicates-gender-bias-in-recommendation-letters/.

Fabian M. Suchanek and Anh Tuan Luu. Knowledge Bases and Language Models: Complementing Forces. In *RuleML+RR invited paper*, 2023.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. The science of detecting llm-generated text. *Commun. ACM*, 67(4):50–59, mar 2024. ISSN 0001-0782. doi: 10.1145/3624725. URL https://doi.org/10.1145/3624725.

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

The Economist, 2023a. Art made by artificial intelligence is developing a style of its own. *The Economist*, 2023. URL https://www.economist.com/culture/2023/05/24/art-made-by-artificial-intelligence-is-developing-a-style-of-its-own.

The Economist, 2024a. Generative AI is a marvel. Is it also built on theft? *The Economist*, 2024. URL https://www.economist.com/business/2024/04/14/generative-ai-is-a-marvel-is-it-also-built-on-theft.

The Economist, 2024b. How businesses are actually using generative AI. *The Economist*, Feb. 29 2024. URL https://www.economist.com/business/2024/02/29/how-businesses-are-actually-using-generative-ai.

The Economist, 2024c. The breakthrough AI needs. *The Economist*, 2024. URL https://www.economist.com/leaders/2024/09/19/the-breakthrough-ai-needs.

Brian Thompson, Mehak Dhaliwal, Peter Frisch, Tobias Domhan, and Marcello Federico. A shocking amount of the web is machine translated: Insights from multi-way parallelism. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1763–1775, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.103. URL https://aclanthology.org/2024.findings-acl.103.

US District Court, Southern District of New York. The New York Times Co. v. Microsoft Corp. & OpenAI Inc., Dec. 27, 2023. URL https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf.

Gaël Varoquaux, Alexandra Sasha Luccioni, and Meredith Whittaker. Hype, Sustainability, and the Price of the Bigger-is-Better Paradigm in AI, 2024. URL https://arxiv.org/abs/2409.14160.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Mathias Vermeulen and Laureline Lemoine. From chatgpt to google's gemini: when would generative ai products fall within the scope of the digital services act? *London School of Economics Blog*, 2024. URL https://blogs.lse.ac.uk/medialse/2024/02/12/from-chatgpt-to-googles-gemini-when-would-generative-ai-products-fall-within-the-scope-of-the-digital-services-act/.

Daniel Victor. Microsoft created a twitter bot to learn from users. it quickly became a racist jerk. *The New York Times*, 2016. URL https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Will we run out of data? limits of llm scaling based on human-generated data, 2024. URL https://arxiv.org/abs/2211.04325.

Mittelstadt Brent Wachter Sandra and Russell Chris. Do large language models have a legal duty to tell the truth? *Royal Society Open Science*, 2024. URL http://doi.org/10.1098/rsos.240197.

Liupu Wang, Juexin Wang, Michael Wang, Yong Li, Yanchun Liang, and Dong Xu. Using internet search engines to obtain medical information: a comparative study. *Journal of Medical Internet Research*, 2012. doi: 10.2196/jmir.1943.

Zezhong Wang, Fangkai Yang, Lu Wang, Pu Zhao, Hongru Wang, Liang Chen, Qingwei Lin, and Kam-Fai Wong. SELF-GUARD: Empower the LLM to safeguard itself. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1648–1668, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.92. URL https://aclanthology.org/2024.naacl-long.92.

Johnny Tian-Zheng Wei, Ryan Yixiang Wang, and Robin Jia. Proving membership in llm pretraining data via data watermarks, 2024.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of

risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533088. URL https://doi.org/10.1145/3531146.3533088.

Matteo Wong. The AI Search War Has Begun. *The Atlantic*, 2024. URL https://www.theatlantic.com/technology/archive/2024/07/perplexity-ai-search-media-partners/679294/.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F. Wong, and Lidia S. Chao. A survey on llm-generated text detection: Necessity, methods, and future directions, 2024. URL https://arxiv.org/abs/2310.14724.

Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. DEPN: Detecting and editing privacy neurons in pretrained language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2875–2886, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.174. URL https://aclanthology.org/2023.emnlp-main.174.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Jinyuan Jia, Bill Yuchen Lin, and Radha Poovendran. SafeDecoding: Defending against jailbreak attacks via safety-aware decoding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5587–5605, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.303. URL https://aclanthology.org/2024.acl-long.303.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, et al. Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22, 2024. URL https://www.thelancet.com/journals/landig/article/PIIS2589-7500(23)00225-X/fulltext.

Ningyu Zhang, Yunzhi Yao, and Shumin Deng. Knowledge editing for large language models. In Roman Klinger, Naozaki Okazaki, Nicoletta Calzolari, and Min-Yen Kan, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pages 33–41, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-tutorials.6/.

Baohang Zhou, Zezhong Wang, Lingzhi Wang, Hongru Wang, Ying Zhang, Kehui Song, Xuhui Sui, and Kam-Fai Wong. DPDLLM: A black-box framework for detecting pre-training data from large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 644–653, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.35. URL https://aclanthology.org/2024.findings-acl.35.

Jakub Złotowski, Diane Proudfoot, Kumar Yogeeswaran, and Christoph Bartneck. Anthropomorphism: Opportunities and challenges in human–robot interaction. *International Journal of*

Social Robotics, 7(3):347–360, Jun 2015. ISSN 1875-4805. doi: 10.1007/s12369-014-0267-6. URL https://doi.org/10.1007/s12369-014-0267-6.