Working papers in Applied Linguistics and Linguistics at York 4, 1-11

Is ChatGPT taking over the language classroom? How language ideologies of large language models impact teaching and learning

Mandy Lau[1]
York University, Toronto, Canada

Abstract: ChatGPT generated much dialogue on the implications of large language models (LLMs) for language teaching and learning. Since language teachers are uniquely positioned to teach metalinguistic awareness, they can support their learners' understanding of how LLMs are shaped by language ideologies and how their outputs are indexical of social power. This awareness would help learners be more conscientious in using LLMs, deciding how to interact with them and adapt their outputs for their purposes. This article introduces LLMs as statistical systems that predict linguistic forms. It surfaces two language ideologies that have shaped their development: the belief in the separability of language from its social contexts and the belief in the value of larger text corpora. It also highlights some ideological effects including uneven language performance, text outputs that reflect biases, privacy violations, circulation of copyrighted materials, misinformation, and hallucinations. Some suggestions for mitigating these effects are offered.

Keywords: language ideologies; ChatGPT; large language models; artificial intelligence; natural language processing

1 Introduction

When ChatGPT was released in November 2022, it dominated media coverage and sparked speculations on the future of AI. Some notable education-specific headlines included future predictions such as the death of the college essay (Marche, 2022) or teachers being replaced (Jha, 2023), and real-time concerns about academic integrity and plagiarism (Barnett, 2023). Chomsky famously described ChatGPT as "high-tech plagiarism" with "absolutely no value with regard to understanding anything about language or cognition" and as "just a way of avoiding learning" (EduKitchen, 2023). However, perhaps one of the most significant values of ChatGPT is in provoking interest and dialogue on the implications of AI tools in teaching and learning.

ChatGPT is a chatbot built on a large language model (LLM) developed by OpenAI[2] (OpenAI, 2022). It can generate coherent and grammatically accurate texts, summarize a large volume of texts, rewrite, edit, or translate existing works, and even generate research sources with just a few prompts. In contrast to some other LLMs, ChatGPT is web-based, available for free,

[1] *Corresponding Author: laumandy@yorku.ca*

[2] GPT stands for Generative Pre-trained Transformers, a neural network machine learning model trained on large data sets of texts used to process and generate human languages. GPT is a family of large language models developed by OpenAI (GPT, GPT-2, GPT-3, GPT-3.5, and the most current version, GPT-4). ChatGPT was built on the GPT-3.5 model while the newest subscription-based ChatGPT Plus can access both GPT-3.5 and GPT-4 (see OpenAI, 2022 and OpenAI, 2023a). Microsoft's Big Chat was built on GPT-4 (Lardiniois, 2023).

 ISSN 2564-2855

and designed with a user-friendly interface. These characteristics undoubtedly encourage wide adoption in the language-learning classroom and beyond. Further, ChatGPT is not the only chatbot available to the public for free; Google launched its chatbot Bard in March 2023 (Hern, 2023) and Microsoft released Bing Chat in February 2023 (Lardinois, 2023). Other lesser-known AI writers, often built on OpenAI's LLMs, can also be readily accessed on the Internet (see, for example Writesonic, 2023, or Jasper, 2023).

As the use of LLMs becomes widespread, language teachers are engaging in more decisions and policymaking on their appropriate use. An informed approach is crucial in these processes, especially since speculations often fall into the fallacy that technology is always better than humans and always the solution (see Broussard, 2018 on technochauvinism). Thus, this article aims to debunk the hype for language teachers. The first section is an overview of how LLMs work and what they are trained to do. The other goal is to surface the language ideologies that undergird LLMs, their effects, and their implications for language teaching and learning. Specifically, the two ideologies discussed are the belief in the separability of language from its social contexts and the belief in the value of larger text corpora. Even as LLMs rapidly evolve, the fundamentals of how they work and their inherent language ideologies are less prone to change. Thus, reflecting on the language ideologies of LLMs during this early adoption phase will better inform teachers and learners as they respond to, resist, and appropriate new forms of language technologies.

2 Overview of large language models

Language models refer to systems that use contextual data to predict the likelihood of a character, word, or sentence, known as tokens (Bender & Gebru et al., 2021). Broadly, they are developed in three stages: First, a dataset or corpus of texts is assembled for training. Next, the model draws from the dataset to calculate the statistical relationships between tokens. Finally, developers fine-tune the model to improve its text generation outcomes for its intended purpose. After development, a user can feed the model a prompt (i.e., a query in the form of a sentence or paragraph) and the model will respond by generating a sequence of characters, words, and sentences (Bender & Koller, 2020; Drenik, 2023; Okerlund et al., 2022).

The current capabilities of LLMs result from advancements in natural language processing (NLP), a field of AI concerned with enabling machines to communicate with humans using human languages in ways that are natural to humans (Khurana et al., 2023). These advancements allow LLMs to be trained on much larger datasets than prior models (i.e., the texts from a large web corpus like Wikipedia), better process the complex patterns and relationships between billions of tokens and generate texts in response to a wide range of prompts—even to novel prompts that it has never encountered in training (Kasneci et al., 2023; Okerlund, 2022). In fact, state-of-the-art LLMs mimic human languaging so impressively that humans have difficulty differentiating between a human-generated text and an LLM-generated one (Okerlund, 2022). Differentiation will only become more difficult as the technology advances. The success of machine mimicry combined with the human tendency to locate meaning contributes to misleading claims that the models can "understand," "comprehend," or "communicate" in ways similar to humans (Bender & Koller, 2020; Bender & Gebru et al., 2021). However, Bender & Koller (2020) point out that "the language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning" (p.5185). In other words, LLMs do not understand languages or communicate with intent; they produce results based on the statistical predictions of the linguistic forms found in the training dataset (Bender & Koller, 2020).

While LLMs are statistical language machines (and described as stochastic parrots because of it (Bender & Gebru et al., 2021)), it does not mean they are neutral, autonomous, or rational systems. Rather, LLMs can be understood as "a type of interactional culture whose human participants are distributed in space and time" (Fester-Seeger & Schneider, 2023, p. 3). In this sense, a ChatGPT user interacts not merely with a machine but with other humans across temporal and spatial contexts, processed out of sight "in the cloud." Thus, the language activity co-produced by the user and ChatGPT cannot truly be removed from social contexts, despite its invisibility. This orientation is in line with Crawford's (2021) framing of AI systems as entirely dependent on political and social structures optimized to serve existing dominant interests. From this view, we can begin to theorize how power relations are baked into code and algorithmically distributed at scale (see Broussard, 2018; Eubanks: 2018; and O'Neil, 2016 for a discussion of how inequities become reinforced and automated in AI tools). Within applied linguistics, we can deconstruct how power via language ideologies becomes embedded in LLMs. Once deployed, LLMs circulate and reinforce language ideologies. The following section will discuss how LLMs intertwine with language ideologies and with which effects.

3 Language ideologies and their effects

I understand language ideologies to refer to the normative beliefs and conceptions about languages, language speakers, and their language practices, structured within particular social groups' moral and political interests (Gal & Irvine, 2019; Milroy & Milroy, 2012). From this view, all linguistic activities are laden with ideological perspectives (Milroy & Milroy, 2012). Language technologies are no exception (see, for example Schneider, 2020, on linguistic normativities in digital publics and Castelle, 2018, on language ideologies in abusive language classification systems). This section will highlight two interrelated ideologies: The belief in the separability of language from its social contexts and the belief in the value of larger text corpora.

At the onset, LLMs assume language is a system that can be separated from human users and their social structures. This assumption is not a surprise given that NLP is informed by approaches from structuralist linguistics based on Saussurean models (Schneider (2020) discusses the relationship between structuralist linguistics and their effects on digital publics). Saussure, often acknowledged as the founder of modern linguistics, viewed language as an internal system, relegating the social, geographical, ideological, political, and other cultural dimensions of language as external (Milroy, 2001). One underlying ideology in this approach is the belief that languages can be separated from other non-linguistic phenomena (Mühlhäusler 1996 as cited in Milroy, 2001). In NLP, this ideology is clearly expressed in the literal removal of humans in human-authored content to be used as language data for machine learning. For example, GPT-3 (the language model that ChatGPT is based on) derives its determination of statistical relationships from two collections of webpages (a filtered version of Common Crawl and WebText datasets), two collections of digital books, and English Wikipedia (Brown et al., 2020). Here, the text is valued as data and not as knowledge or meaning making tied to human authors and their social contexts. This is apparent in the fact that there is no attempt to seek the relationship between the text (data) and its authors in the machine learning framework; in this way, humans are erased from their texts.

A language-as-data approach can sufficiently enable LLMs to perform formal linguistic competence in their outputs, generating texts that follow the rules and patterns of a language in coherent and grammatical ways (Mahowald et al., 2023). However, it also devalues language as knowledge or meaning making. Put another way, LLMs recycle linguistic forms found in their

training data with high human likeness while erasing the data's symbolic representation. This treatment of language as data and disregard for the sociopolitical contexts of its authors contribute to some of LLMs' problematic practices, such as copyright infringement (see, for example Eliot, 2023) or the recirculation of misinformation (see Hsu & Thompson, 2023; Okerlund et al., 2022).

Severing texts from their sociopolitical dimensions does not truly make a "neutral" language, no matter its intention. As Schneider (2020) asserts, "language is always a part of, and embedded, in social contexts" (p. 2). Consider that GPT-3 is trained primarily from two large datasets: a filtered version of Common Crawl and WebText (accounting for 62% and 22% of the weight in the training mix, respectively (Brown et al., 2020)). Common Crawl, an Amazon Web Services dataset, contains petabytes of texts scraped over 12 years (Common Crawl, 2023). WebText, an OpenAI dataset, scraped 45 million of Reddit's most popular outbound links (WebText, 2023). These large web-based datasets are highly valued because, in principle, more texts in the corpus will provide more diverse patterns for machine learning, which will lead to LLM text generation that more closely resembles diverse humans (Dickson, 2022). However, their large size and their sourcing from the "open" Internet does not guarantee a representative or equitable sample of human language practices (Bender & Gebru et al., 2021), never mind a "neutral" one.

In terms of languages, English continues to be the predominant content language of the Internet. As of March 2023, English accounts for 56.1% of all websites (see Figure 1; W3Techs, 2023). Other languages do not even come close; the next most frequently used content languages are Russian at 5.1%, Spanish, French and German at over 4%, and Japanese at 3.5% (W3Techs, 2023). Thus, using datasets such as the Common Crawl can lead to the over-representation of English and the perspectives of English-speaking content creators, thereby under-resourcing languages other than English and the perspectives of their speakers. This is true for all NLP tools: the more resourced the language is for training, the better the tool works for speakers of that language. This means that LLMs that use Common Crawl for training, such as ChatGPT, may perform with expected effectiveness for English-speaking users but less so for speakers of languages other than English.

Despite the overall large size of training data and dominance in language representation, English language datasets still contain other biases that become further encoded. For example, datasets that crawl from the most popular, user-generated content sites (i.e., Twitter or Reddit) tend to overrepresent prevalent problematic views such as white supremacy, misogyny, homophobia, ableism, and ageism (Bender and Gebru et al., 2021). This tendency may be a result of uneven participation on the Internet. For example, WebText's use of Reddit data will overrepresent the views of males—as of January 2022, 63% of Reddit users worldwide are male (Dixon, 2022)—and those residing in the United States—as of May 2022, 47% of Reddit's desktop traffic was based in the United States (Bianchi, 2023). Scraping web news or digital books offers an alternative but is still the product of social and political biases that determine whose texts and viewpoints are published. For example, Google's Word2vec is known to pick up social and political biases from Google News (Bender and Gebru et al., 2021; Bolukbasi et al., 2016; Wachter-Boettcher, 2017). Even volunteer, open-collaboration datasets such as Wikipedia contain disparities, overrepresenting the languages and knowledge of those with easy Internet access and leisure time to edit entries (Graham, 2020, as cited in Godwin-Jones, 2021). In terms of gender disparity, women are less likely to be editors of Wikipedia articles, which contributes to the underrepresentation of women in all fields and the underdevelopment of articles about women's

interests (Tripodi, 2021). These biases and disparities in the training dataset can become entrenched in the LLMs outputs, further reproducing and normalizing these views.

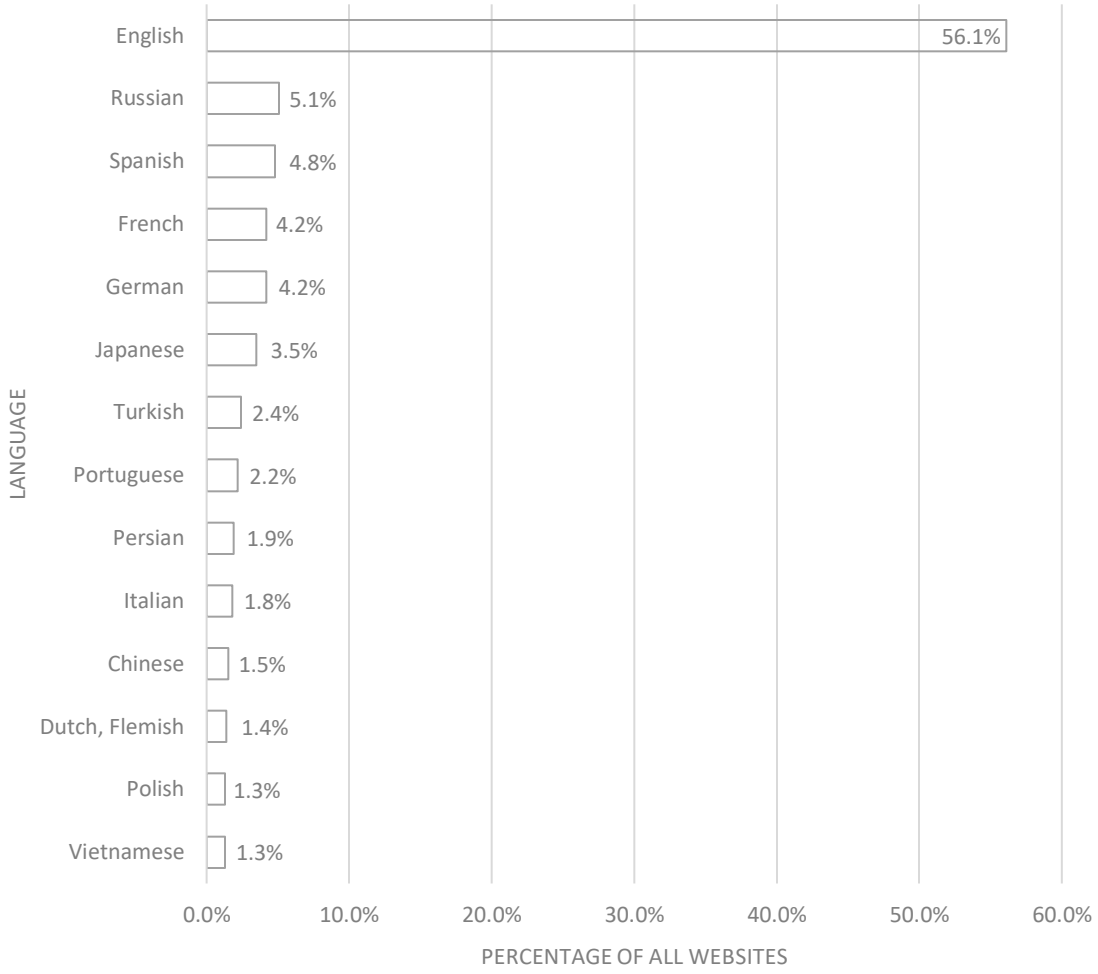

Figure 1. Content languages of all websites[3] (as of March 2023; W3Techs, 2023)

So, is the solution to increase the training data with even more texts? Some educators have attempted to address these gaps creatively, such as through hack-a-thons or edit-a-thons. Tripodi (2021) studied an edit-a-thon organized to encourage more women editors and increase the representation of women in science on Wikipedia. It is possible that this strategy can restore agency for users while improving the digital knowledge landscape and the performance of LLMs trained on Wikipedia. However, the strategy is only short-term and highlights a few concerns: first, the onus of representation falls back on the shoulders of those marginalized, adding to their workload. Second, the labour arrangements of an edit-a-thon (or Wikis in general) provide no compensation. This normalizing of a gift economy (Kelly-Holmes, 2019) only serves to further the equity gap. Third, while the women in this group were tasked with creating new biographies, other women participants may hypothetically be tasked to edit out misogynist or sexist entries. This practice can expose women to more misogyny and sexism with harmful effects, much like the psychological

[3] Only the languages that make up over 1% of all websites are included in this figure.

harms experienced by content moderation workers in Kenya who "clean" OpenAI's language data from hate speech, sexual violence, or extreme violence (Perrigo, 2023).

Another way learners can contribute to increasing more representative training data is through the repeated use of LLMs. All text inputs by the user are stored by the developer and can be used in future iterations (C&J, 2023). In other words, all LLM users are contributing language data that can be used to improve training (among other commercial or functional purposes).[4] However, this raises privacy issues whereby personal data may be reproduced publicly, hacked, or accidentally leaked (C&J, 2023).[5]

Finally, more training data may not matter much when LLMs are designed to provide one answer without corresponding contextual information. This design does not enable what Deepak P (2023) calls "user exploration capability". In contrast to search engines, LLMs do not yield multiple links for users to explore the larger context in which the information is situated (P, 2023). Algorithmic logics also tend to reward what is most popular, frequent, or profitable in its outputs (see, for example, Noble (2018) on Google's prioritizing of profit in Search rankings; Schneider (2020) on the conflation of frequency with correctness in machine translation; WebText (2023) for a direct rationale on why only the most popular Reddit links are used for training; and West et al. (2019) on how voice assistants prioritize results to generate simple, single answers). Further, LLMs will provide any plausible-looking answer, even if it lacks data relevant to the prompt. In such cases, LLMs may generate fictitious or nonsensical responses known as hallucinations (P, 2023). These output tendencies have implications for learning, requiring further critical thinking and fact-checking practices.

4 Implications for language teaching and learning

Language teachers are uniquely positioned to teach metalinguistic awareness, particularly the relationship between language, cultural norms, and sociopolitical power. As Milroy (2001) asserts, language needs to be defined "in the very terms that Saussure excluded from the remit of linguistics as a subject"—namely, locating languages within their social, geographical, ideological, political, and cultural dimensions (p. 541). Language teachers can support learners in identifying how language ideologies are a part of LLMs and how their linguistic outputs can be indexical of social power. This awareness can help learners be more conscientious in using LLMs, deciding how to interact with them and adapt their outputs for their purposes. This is particularly crucial for language learners who are incentivized to use and trust an LLM's formal linguistic competence more than their own capabilities in their additional language. In addition, per typical terms and conditions of use, learners are entirely responsible for what they do with the LLMs' outputs (see the terms for OpenAI, 2023b). Thus, developing a more informed stance will increase learner agency as they make intentional choices about using the tool.

The previous section highlighted the effects of two interrelated language ideologies embedded in LLMs: The belief in the separability of language from its social contexts (language as data) and the belief in the value of larger text corpora (more language input means better

[4] As of this writing, ChatGPT's terms and conditions stipulates that all text inputs, file uploads, or feedback provided by individual users are collected by default and used to improve services, including training machine learning models (OpenAI, 2023b). Users can choose to opt out manually (OpenAI, 2023b). Users enrolled in ChatGPT's business offerings are also exempted from content collection (OpenAI, 2023b). There are also other LLMs that are more privacy and security oriented and developed to run entirely offline.

language output). The following table summarizes the possible consequences of these ideologies and offers suggestions on how learners can mitigate their effects.

Table 1. Suggestions to mitigate the consequences of language ideologies embedded in LLMs

Consequence	Suggestions
Uneven language performance	Users should be aware that different languages will operate at different performance standards and possibly draw from different sources. The effectiveness of LLMs will vary according to the language the learner inputs in the prompts. Users can try using different languages for the same prompt to explore how the outputs may differ. The use of LLMs for translation purposes may also have different degrees of accuracy and will need human verification. This could be a teaching and learning opportunity to reflect on the appropriateness and accuracy of the LLMs' translation. Learning how LLMs work for different languages is also an opportunity to explore how languages are conceptualized in digital culture. For example, users may infer which languages have the most prestige, why some languages and varieties do not work as effectively or are missing, or how they may adapt their language practices to make LLMs work for them.
Text outputs that are tilted towards the dominant, the popular, the frequent, or the profitable.	Users should be aware of how an LLM's outputs may reflect dominant voices on the Internet or corporate interests. They can work around this by asking prompts within specific and varying viewpoints to elicit less dominant perspectives. (Example prompt: write a paragraph about [topic] from the perspective of [a theoretical approach or a person]). They may also compare the outputs with other perspectives external to the LLM (i.e., a search engine from a different corporation or an academic or journalistic source). Some of the outputs may also contain stereotypes. Teachers can address this pre-emptively by deconstructing how these stereotypes circulate digitally and preparing classroom supports to counter the harmful effects of encountering stereotypes (Noble, 2018).
Privacy violations	Users can take steps to protect their privacy. They should avoid sharing private or sensitive information in prompts (C&J, 2023). They may also use a private VPN when using the tool, remove personal details when opening an account, consider sharing the account with multiple users, or learn to remove data or delete the account (Véliz, 2020).

Consequence	Suggestions
Copyright infringement	Users can prompt LLMs to produce citations. They can verify the citations outside the LLM to determine how the output should be cited or incorporated. Users can verify the text output using third-party plagiarism detection software. Since users can also ask the LLM to paraphrase or summarize a text as a prompt (and repeat this prompt multiple times), it may be beneficial to review academic integrity policies and citation practices.
Circulation of misinformation or fictionalized nonsense (hallucinations)	Users can vary their prompts to include different positions, specific purposes, or narrow conditions to get different outputs. (Example prompts: What would [person with specific position] say about [topic]? Tell me about [topic] for educational purposes.) They can also prompt for citations, web links, or alternative sources and verify these externally (P, 2023).

Overall, the efficacy of the suggestions depend on widening literacy skills to include skills such as prompting for different purposes and cross-platform fact-checking. These suggestions stress the importance of being vigilant about text inputs (including both language and content) and evaluating sources during fact-checking processes. Unless otherwise specified, these strategies draw from my explorations and classroom experiences and have not been reviewed or validated. Additional studies on how learners make sense of LLMs and exploit their affordances for their purposes could be a practical next step.

5 Conclusion and next steps

Will LLMs like ChatGPT take over the language classroom? Not for those who believe that language is more than data and writing is more than generating form. The hype around LLMs necessitates a critical understanding of how LLMs work and what they are trained to do. It also requires us to sharpen our definition of what language is and what we do with it. To this end, this paper broadly introduced LLMs as statistical systems that predict linguistic forms and discussed two language ideologies that have shaped the models: the belief in the separability of language from its social contexts and the belief in the value of larger text corpora. It also highlighted some ideological effects and their implications for language teaching and learning. Future analysis of other embedded language ideologies could help expand and propagate the understanding of LLMs as sociolinguistic machines and place them within a longer sociopolitical history. Further studies on the inferential processes that learners bring to LLMs will also help teachers and learners respond to new NLP technologies.

Acknowledgements

My research activities are supported in part by funding from the Social Sciences and Humanities Research Council. I wish to thank Dr. Eve Haque, Dr. Geoff Lawrence, Dr. Antonella Valeo, and Dr. Robert Godwin-Jones for their support and feedback in developing this paper.

References

Barnett, S. (2023, January 30). ChatGPT is making universities rethink plagiarism. *Wired*. https://www.wired.com/story/chatgpt-college-university-plagiarism/

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/V1/2020.ACL-MAIN.463

Bender, E., Gebru, T., McMillan-Major, A., Shmitchell, S., & Anonymous. (2021). On the dangers of stochastic parrots: can language models be too big? *Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3-10, 2021, Virtual Event, Canada*, 271–278. https://doi.org/10.1145/3442188.3445922

Bianchi, T. (2023). *Regional distribution of desktop traffic to Reddit.com as of May 2022 by country*. Statista. https://www.statista.com/statistics/325144/reddit-global-active-user-distribution/

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *30th Conference on Neural Information Processing Systems*, 1–9. https://dl.acm.org/doi/10.5555/3157382.3157584

Broussard, M. (2018). *Artificial unintelligence: How computers misunderstand the world*. The MIT Press.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*. https://arxiv.org/pdf/2005.14165.pdf

C, D., & J, P. (2023, March 14). ChatGPT and large language models: What's the risk? *National Cyber Security Centre UK*. https://www.ncsc.gov.uk/blog-post/chatgpt-and-large-language-models-whats-the-risk

Castelle, M. (2018). The linguistic ideologies of deep abusive language classification. *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, 160–170. https://doi.org/10.18653/v1/w18-5120

Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Crawl, C. (2023, April 7). *Want to use our data?* Common Crawl. https://commoncrawl.org/the-data/

Dickson, B. (2022, August 30). *LLMs have not learned our language - we're trying to learn theirs*. VentureBeat. https://venturebeat.com/ai/llms-have-not-learned-our-language-were-trying-to-learn-theirs

Dixon, S. (2022). *Distribution of Reddit users worldwide as of January 2022, by gender*. Statista. https://www.statista.com/statistics/1255182/distribution-of-users-on-reddit-worldwide-gender/

Drenik, G. (2023, January 11). Large language models will define artificial intelligence. *Forbes*. https://www.forbes.com/sites/garydrenik/2023/01/11/large-language-models-will-define-artificial-intelligence

EduKitchen. (2023, January 21). *Chomsky on ChatGPT, education, Russia and the unvaccinated*. [Video]. YouTube. https://youtu.be/IgxzcOugvEI

Eliot, L. (2023, February 26). Legal doomsday for generative AI ChatGPT if caught plagiarizing or infringing, warns AI ethics and AI law. *Forbes*. https://www.forbes.com/sites/lanceeliot/2023/02/26/legal-doomsday-for-generative-ai-chatgpt-if-caught-plagiarizing-or-infringing-warns-ai-ethics-and-ai-law

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. Picador.

Fester-Seeger, M.-T., & Schneider, B. (2023). *AI technology as interactional human culture: Language, data practice and social struggle*. European University Viadrina Frankfurt (Oder).

Gal, S., & Irvine, J. T. (2019). *Signs of difference: Language and ideology in social life*. Cambridge University Press.

Godwin-Jones, R. (2021). Big data and language learning: Opportunities and challenges. *Language Learning & Technology*, *25*(1), 4–19. http://hdl.handle.net/10125/44747

Hern, A. (2023, March 21). Google's Bard chatbot launches in US and UK. *The Guardian*. https://www.theguardian.com/technology/2023/mar/21/googles-bard-chatbot-launches-in-us-and-uk

Hsu, T., & Thompson, S. A. (2023). Disinformation researchers raise alarms about A.I. Chatbots. *The New York Times*. https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html

Jasper. (2023, April 7). *Meet Jasper*. Jasper. https://www.jasper.ai/

Jha, S. (2023, March 28). Will ChatGPT take your job - and millions of others? *Al Jazeera*. https://www.aljazeera.com/features/2023/3/28/will-chatgpt-take-your-job-and-millions-of-others

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*. https://doi.org/10.1016/J.LINDIF.2023.102274

Kelly-Holmes, H. (2019). Multilingualism and technology: A review of developments in digital communication from monolingualism to idiolingualism. *Annual Review of Applied Linguistics*, *39*, 24–39. https://doi.org/10.1017/S0267190519000102

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*, 3713–3744. https://doi.org/10.1007/S11042-022-13428-4

Lardinois, F. (2023, February 7). Microsoft launches the new Bing, with ChatGPT built in. *TechCrunch*. https://techcrunch.com/2023/02/07/microsoft-launches-the-new-bing-with-chatgpt-built-in/

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). *Dissociating language and thought in large language models: a cognitive perspective*. 2023. https://arxiv.org/abs/2301.06627v1

Marche, S. (2022, December 6). The college essay is dead. *The Atlantic*. https://www.theatlantic.com/technology/archive/2022/12/chatgpt-ai-writing-college-student-essays/672371/

Milroy, J. (2001). Language ideologies and the consequences of standardization. *Journal of Sociolinguistics*, *5*(4), 530–555. https://doi.org/10.1111/1467-9481.00163

Milroy, J., & Milroy, L. (2012). *Authority in language: Investigating standard English* (4th ed.). Routledge.

Noble, S. (2018). *Algorithms of Oppression*. NYU Press.

O'Neil, C. (2016). *Weapons of math destruction*. Crown Books.

Okerlund, J., Klasky, E., Middha, A., Kim, S., Rosenfeld, H., Kleinman, M., & Parthasarathy, S. (2022). What's in the chatterbox? Large language models, why they matter, and what we should do about them. In *University of Michigan Technology Assessment Project*. https://stpp.fordschool.umich.edu/research/research-report/whats-in-the-chatterbox

OpenAI. (2022, November). *Introducing ChatGPT*. OpenAI. https://openai.com/blog/chatgpt

OpenAI. (2023a, March 14). *GPT-4*. https://openai.com/research/gpt-4

OpenAI. (2023b, March 14). *Terms of use*. OpenAI. https://openai.com/policies/terms-of-use

P, D. (2023, March 22). What we can learn from how ChatGPT handles data densities and treats users. *Data & Society: Points*. https://points.datasociety.net/what-we-can-learn-from-how-chatgpt-handles-data-densities-and-treats-users-f62b2cb222c6

Perrigo, B. (2023, January 18). Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic. *Time*. https://time.com/6247678/openai-chatgpt-kenya-workers/

Schneider, B. (2020). Language and publics in a global digital world. What is linguistic citizenship in the 21st century? *Studia Universitatis Babeș-Bolyai Studia Europaea*, *65*(2), 45–70. https://doi.org/10.24193/subbeuropaea.2020.2.03

Tripodi, F. (2021). Ms. Categorized: Gender, notability, and inequality on Wikipedia. *New Media & Society*. https://doi.org/10.1177/14614448211023772

Véliz, C. (2020). *Privacy is power: Why and how you should take back control of your data*. Penguin Random House.

W3Techs. (2023, April). *Usage statistics of content languages for websites, April 2023*. https://w3techs.com/technologies/overview/content_language

Wachter-Boettcher, S. (2017). Technically wrong: Sexist apps, biased algorithms, and other threats of toxic tech. In *W.W. Norton & Company*.

WebText. (2023). *WebText*. https://paperswithcode.com/dataset/webtext

West, M., Kraut, R., & Chew, H. E. (2019). *I'd blush if I could: Closing gender divides in digital skills through education*. https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1

Writesonic. (2023, April 7). *Writesonic: Best AI writer, copywriting & paraphrasing tool*. Writesonic. https://writesonic.com/