Where We Are: AI and Writing

Truth-Telling: Critical Inquiries on LLMs and the Corpus Texts That Train Them

Antonio Byrd

The commonplace concept 'literacy crisis' has framed ChatGPT's popularity, its rapid evolution (GPT-4's release five months after ChatGPT 3.5), and its seemingly sophisticated language and knowledge performance. The concept helps scholars and teachers easily enter conversations about AI text generation technologies and how they transform our notions of authorship, research, labor, copyright, and writing. I've used the literacy crisis as a starting point for a 2023 CCCC Annual Convention panel called "ChatGPT, Magical Thinking, and the Discourse of Crisis" (Byrd et al.) and again at my home institution for a virtual workshop on LLMs that I co-facilitated with colleagues from computer science and the university writing center. On both occasions, I shared my thoughts on how resisting AI text generation technologies suggested that writing instructors want to protect ideologies and power that may restrict the flow of literacy practices. We have an opportunity to uncover conflicts or tensions over power and ideology among higher education, the companies that develop LLMs, workplaces, and marginalized communities as AI text generation technologies reshape our literacy practices.

I again use the literacy crisis as a starting point in this essay but in the opposite direction. By responding to literacy crises with a back-to-basics pedagogy, composition studies has participated in histories of linguistic punishment that have shaped how writers produce, edit, and publish texts; these punishments create the presence of power and ideology in the corpus texts LLMs use for training. While we learn new literacy practices and teaching strategies with LLMs, we also bear the responsibility for participating in the creation of the next iteration of public data that contain our contemporary ideologies on language and culture. AI and writing involve critical inquiry on our relationships with these corpus texts. Integrating this critical inquiry may maximize students' writing practices with AI and extend their rhetorical awareness of what's at stake when they go public with their writing to participate in cultural and political conversations.

ChatGPT, along with other LLMs, train on public content extracted from the Internet (called datasets or corpus texts) to learn the nuances and patterns of human languages (some can also train on datasets that contain millions of images to generate multimodal content and analyze multimodal inputs to generate text as output in response). As chatbots, LLMs seem to be an effective

knowledge system and natural language processor to human users; however, they have really created mathematical formulas to predict the next token in a string of words from analyzing patterns of human language. They learn a form of language, but do not understand the implicit meaning behind it. Meaning comes from a "shared knowledge that the interactants bring to the scene . . . That right there is the complexity of rhetoric: It is an art that attempts to deal with the inherent messiness of the richness and variability of communication context" (*Professional Communication and Network* 147). Without accessing context, they are "stochastic parrot[s]" making stereotypical word associations and negative sentiments about marginalized people without understanding why those words are harmful (Bender et al. 616). AI safety experts can intervene in the training to limit these outputs to the user and mitigate hallucinations. (We can still work around those guardrails, however. Visit jailbreakchat.com and you'll find prompts that have generated harmful outputs. OpenAI uses this crowdsourced information to make patches in its LLMs.)

But scholars and teachers in writing studies recognize the rhetorical context and histories of linguistic ideologies and dominant power that have oppressed, mitigated, and erased marginalized communities that LLMs have encoded. Linguistic punishment includes violence against bodies and land: colonialism, imperialism, genocide, and slavery paved the way for English dominance in contemporary global economies. The downstream impact can be felt in corpus text creation. White supremacy has led to a majority of the white workforce designing AI technologies; GPT-3 trained on just seven percent of non-English languages (Bender et al. 611). Dire consequences can await multilingual users. For example, Facebook Translate mistranslated a Palestinian's Arabic for "good morning" into "hurt them" in English. Without asking further questions first, the Israeli police arrested the user (Gebru 264). The event leans on histories of violent conflict between Israel and Palestine, histories bound up in AI technologies. Corpus texts lack Reddit and Twitter posts with different varieties of English because "moderation practices . . . make them [Reddit and Twitter] less welcoming to marginalized populations" (Bender et al. 613). Harassment, doxing, and death threats push marginalized users out, while abusive culprits and their perspectives stay online. Even digital spaces made for marginalized people may not be included in LLMs' training data because they lack enough incoming and outgoing links to show up in corpus texts. Some LLMs try to filter out abusive language, although some of that same language has been reclaimed by marginalized people (Bender et al. 613–14). LLMs train on whatever is left over from linguistic and multimodal conflicts among human users.

These forms of linguistic punishment evoke how composition studies responds to literacy crises. The discipline has tracked how "middle class anxieties

about the loss of status and downward mobility have repeatedly been displaced and refigured in the realm of language and literacy education" (Trimbur 280). Rhetoric's beginnings as a discipline in the United States was based, in part, on "determining whose writing was better, and whose writing was better was, in part, determined from linguistic determinations by and from certain racial and economic privileged groups. Solving the disciplinary crisis means resolving the tensions of who is seen as a viable and important part of the country's identity" (Burrows 150). Some white people believed that the Civil Rights Movement and Vietnam War protests challenged the hegemonic power of the United States. As Black and Brown students gained access to white-majority colleges and universities, writing instructors designed race-conscious pedagogies that responded to the variety of Englishes those students brought with them. The worry about Black people challenging white supremacy trickled down to these pedagogies. By the mid 1970s, they had become the source of the literacy crisis discourse: white politicians, scholars, and journalists blamed the so-called falling standards of literacy on these pedagogies; they demanded, and received, a back-to-basics approach to literacy education (Lamos 126–27). The emphasis on "standard American English" appears in GPT-4's responses to prompts. Ask it to give a list of English sentences with common errors, and GPT-4 will offer some examples that suggest African American Vernacular English is wrong(@pengowray). LLMs' linguistic practices and judgments on the language we ask it to analyze come from how humans interact with each other and act on texts (broadly defined). In this example, LLMS have learned that "teachin one correct way lend a hand to choppin off folks tongues" (Young 110). The 'literacy crisis' is coming from inside our classrooms.

As I reflect on these histories of linguistic ideology, I think about how writing instructors teach in a contemporary political moment when violent language against marginalized communities circulates widely. Without data governance, the next corpus text will contain the associations and negative sentiments of contemporary bigotries that have accelerated in the United States since 2021 (the final year OpenAI's GPT-4 trained on). A number of examples come to mind: hate speech against Asian and Asian Americans during the COVID-19 pandemic; hate speech and conspiracy theories from white supremacist, especially during the 2020 elections and after the January 6th attack on the US Capitol; hate speech against trans people as anti-trans and anti-drag bills (which implicitly targets trans people) flood Republican controlled state legislatures; the word associations those same legislatures use to poorly define and attack diversity, equity, and inclusion and critical race theory. Our political and social ideologies communicated through language turn into actions, and vice versa. We created histories that produce new texts for the next iteration of the LLMs to train. Our production of texts, however, now occur alongside

synthetic text (AI-generated content), which poses new challenges to public data; not only human and machine-generated texts running on bigotries and misinformation feed into corpus texts of the future (Kirschenbaum) but also more homogenous low-quality synthetic text disincentivizes humans from contributing their labor to public data for "fears of labor replacement or lack of attribution" (Huang and Siddarth 2).

However, in recent years, composition studies has sought more justice-informed pedagogy. More scholarship on linguistic racism, more workshops on anti-racism and assessment, and more awareness and teaching of how dominant groups use violent language to shape the political landscape against marginalized communities. Keeping in mind the need for to interrogate how these intentions that may signal "benevolent gaslighting" rather taking responsibility for composition's culpability in oppression (Prasad and Maraj 324 - 327), writing classrooms can be a counterweight to the historical moments that turn into corpus texts, with marginalized social identities now more prominent on the mainstream stage than ever before, and our pedagogies shifting (although slowly) toward linguistic justice and care. Writing instructors *across disciplines* have an opportunity to partner with their students on creating digital content that the next iteration of LLMs will one day scrape. To this end, we are positioned to launch critical inquiries on corpus texts: how they are made, what they contain, how they shape our own literacy practices when they filter through the literacy practices of LLMs, and how we participate and write in the histories that LLMs consume.

Writing instructors can conduct critical inquiry of LLMs and corpus texts with their students from multiple angles. Critical inquiry begins with choosing ethical LLMs for AI-assisted writing. As writing instructors, we would want access to two areas where we can recognize the role of human values in literacy: the front end and the back end (*Professional Communication and Network* 140). ChatGPT gives us the front-end design with its chatbot interface and confident responses to prompts. While we are drawn to the front end's capabilities for research, pedagogy, experimentation, we have no access to the corpus text which sourced "both publicly available data (such as internet data) and data licensed from third-party providers" (OpenAI 2). OpenAI refuses to share "details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar" for safety and market competition (OpenAI 2). On this basis alone, ChatGPT may not be an ethical tool for our purposes as writers and researchers.

An alternative for writing instructors may be open access LLMs such as BLOOM, "the first LLM of its scale designed with commensurate efforts in responsible licensing and data governance" (Piktus et al. n.p.). Unlike GPT-4, BLOOM trained on ROOTS, a 1.6TB multilingual text corpus. Although

perhaps not as powerful as GPT-4, BLOOM was nevertheless designed with transparency in mind, and produces text from 46 human languages and 13 computer programming languages (incidentally English texts in ROOTs ranks at number three). With an open access LLM, we can continue critical inquiry of LLMs by examining corpus texts alongside their outputs. Piktus et al. share a range of use cases for critiquing BLOOM's outputs using The ROOTS Search Tool, a search engine that allows the user to do fuzzy searches of the corpus text BLOOM trained on (4). The results only show snippets of texts in the dataset, but users can request access to the full corpus for research. Writing instructors may draw on these suggested use cases to critique LLMs themselves and develop other cases. As students use BLOOM for research, their assessment not only includes checking for hallucinations, bias, and bigotry, but also looking through the text for what language and textual representations exist that may influence those responses. Gleaning the corpus text in relation to prompt engineering and output illuminates how LLMs are tools of political and cultural ideology, not neutral technologies.

I'm thinking about how this critical inquiry of LLMs, corpus texts, and writing opens possibilities for truth-telling, a rhetorical practice of speaking to the truth about human knowledge and action embedded in our technologies. The framework counters utopian thought about AI advancing civilization—an idea that echoes the literacy myth, the persistent belief that acquisition of literacy is necessary for economic development, democratic practice, upward social mobility, and other social markers of advancing toward progress and potential (Graff xvi). Building on the activities above, truth-telling involves having two conversations between writing instructors and students.

The first conversation recognizes that human progress happens in tandem with our failures to recognize our own worth and dignity. Specifically, asking questions about how corpus texts, filtered through the practices of LLMs, influence the integrity of our knowledge, writing, and perspectives on social identities and how we interact with those social identities. What completed and copyrighted materials violently stolen for LLMs says about us sets a foundation for interrogating what's missing and how to enter marginalized voices into the record using sound ethical frameworks. Work in digital writing research, and elsewhere, have found that institutional review boards lack the decision-trees needed to address the complexity of gathering internet data without consent ("The Ethics of Digital Writing Research," 716–18); talking directly to marginalized people about their expectations of internet researchers (Klassen and Fiesler 7–9), inventing critical fictions on using public data to speculate on the consequences of gathering public data (Pater et al.), and using ROOTS search tools to glean the kinds of texts present in the corpus help assert our agency over writing technologies while considering our responsibilities as writ-

ers. When we say that public data is human creativity that belongs to us, what we really mean to say is that human creativity belongs to some of us but not others. It is ironic, and perhaps no accident, that social movements like Black Lives Matter are not well-documented or properly represented in media for LLMs to process (Bender et al. 614).

What human activities and knowledge are missing and how they are gathered (or not) through web scraping leads to a second conversation with students: Data privacy matters, yes, but deliberate introduction of ourselves into the record has equal weight for consideration. Composition classrooms afford the opportunity to add a new layer to audience analysis. We've gone from writing for humans to ethically writing for algorithms that organize online content (Gallagher, "Writing for Algorithmic Audiences," 28–31; Gallagher, "The Ethics of Writing for Algorithmic," 4–5). Now we must go public with our writing knowing LLMs scrape our writing. Exploring how we participate in language creation and how we physically act on our interpretations of social reality matters for holding ourselves accountable. These activities help answer the question: what truths do we write into public spheres for LLM audiences? I'm not suggesting writing about love and empathy but rather telling the truth about bigotry; we find love and empathy with the force of activism. Marginalized people, and their allies, take the lead in truth-telling, and entering the truth into the record will be the closest we get to participating in the design of AI text generation technologies.

AI and writing require more thoughtful, careful rhetorical and ethical frameworks that shape our literacy practices. The histories that surround us and how those histories fuel language practices and vice versa only point toward how "stakes is high" for human users. Engaging with LLMs and its corpus texts may deepen our understanding how AI text generations work while informing the literacy practices we adapt as we inevitably write with and for AI text generation technologies. I hope these rough sketches of where we are today can be an easy entrance into thinking critically about LLMs for diverse writing instructors—adjuncts, lecturer, non-tenure track professors, especially. Through teaching practice, sharing resources, and learning in community we can carry the burden of LLMs together.

Works Cited

@pengowray (Pengo wray). "it gets worse." *Twitter,* 23 Mar. 2023, 4:38 p.m. twitter.com/pengowray/status/1639018762019151875.

Albert, Alex. *Jailbreak Chat*. Feb. 2023. www.jailbreakchat.com/. Accessed 24 Apr. 2023.

Bender, Emily M. et al. "On the Dangers of the Stochastic Parrots: Can Language Models Be Too Big?" *FAccT '21: Proceedings of the 2021 ACM Conference on*

Fairness, Accountability, and Transparency, Virtual Event, Canada, March 3 - 10, 2021. Edited by Lilly Irani, Sampath Kannan, Meg Mitchell, and David Robinson, Association for eComputing Machinery, 2021, pp. 610–23.

Burrows, Cedric D. "A Historical and Cultural Rendering of the Rhetoric of Disciplinary Crisis." *Composition Studies* vol. 50, no. 3, 2022, pp. 149 - 152.

Byrd, Antonio, et al., panelists. "ChatGPT, Magical Thinking, and the Discourse of Crisis." The 2023 Annual Convention of the Conference on College Composition and Communication, 17 Feb. 2023, International Ballroom North, 2nd Floor, Hilton Chicago, Chicago, IL.

Gallagher, John R. "Writing for Algorithmic Audiences." *Computers and Composition* vol. 45, 2017, pp. 25 - 35. dx.doi.org/10.1016/j.compcom.2017.06.002

—. "The Ethics of Writing for Algorithmic Audiences." *Computers and Composition*, vol. 57, 2020, pp. 102583. doi.org/10.1016/j.compcom.2020.102583

Gebru, Timnit. "Race and Gender." *The Oxford Handbook of Ethics of AI*, edited by Markus D. Dubber, Frank Pasquale, and Sunit Das, Oxford U P, 2020, pp. 253–70.

Graff, Harvey J. *The Literacy Myth: Literacy and Social Structure in the Nineteenth-Century City*. Academic Press, 1979.

Huang, Saffron, and Divya Siddarth. "Generative AI and the Digital Commons." *Collective Intelligence Project*, 20 Mar 2023, doi.org/10.48550/arXiv.2303.11074.

Kirschenbaum, Matthew. "Prepare for the Textpocalypse." *The Atlantic*, 8 Mar. 2023, www.theatlantic.com/technology/archive/2023/03/ai-chatgpt-writing-language-models/673318/. Accessed 6 Apr. 2023.

Klassen, Shamika, and Casey Fiesler. "'This Isn't Your Data, Friend': Black Twitter as a Case Study on Research Ethics for Public Data." *Social Media + Society*, vol. 8, no. 4, 2022, pp. 1–11. doi.org/10.1177/20563051221144317.

Lamos, Steve. "Literacy Crisis and Color-Blindness: The Problematic Racial Dynamics of Mid-1970s Language and Literacy Instruction for "High-Risk" Minority Students." *College Composition and Communicatio*n, vol. 61, no. 2, 2009, pp. 125–48.

McKee, Heidi, and James E. Porter. *Professional Communication and Network*: *A Rhetorical and Ethical Approach*. Routledge, 2017.

---. "The Ethics of Digital Writing Research: A Rhetorical Approach." *College Composition and Communication*, vol. 59, no. 4, 2008, pp. 711–49.

OpenAI. *GPT-4 Technical Report*. 27 Mar 2023, arxiv.org/pdf/2303.08774, Accessed 8 April 2023.

Pater, Jessica, et al. "No Humans Here: Ethical Speculation on Public Data, Unintended Consequences, and the Limits of Institutional Review." *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, 2022, pp. 1–13. doi.org/10.1145/3492857.

Prasad, Pritha and Louis M. Maraj. "'I Am Not Your Teaching Moment': The Benevolent Gaslight and Epistemic Violence." *College Composition and Communication* vol. 74, no. 2, 2022, pp. 322–51.

Piktus, Aleksandra. "The ROOTS Search Tool: Data Transparency for LLMs." *Hugging Face*, 27 Feb. 2023, huggingface.co/papers/2302.14035.

Trimbur, John. "Literacy and the Discourse of Crisis." *The Politics of Writing Instruction: Postsecondary*. edited by Richard Bullock and John Trimbur, Boynton/Cook Publishers, 1991, pp. 277–95.
Young, Ashanti Vershawn. "Should Writers Use They Own English?" *Iowa Journal of Cultural Studies,* vol. 12, no. 1, 2010, pp. 110–17.