Theories and Methodologies

Generative Theories, Pretrained Responses: Large AI Models and the Humanities

SETH PERLOW

SETH PERLOW is associate professor of English at Georgetown University. He is the author of *The Poem Electric: Technology and the American Lyric* (U of Minnesota P, 2018) and editor of Gertrude Stein's Tender Buttons*: The Corrected Centennial Edition* (City Lights, 2014). He is completing a book about handwriting and electronics, tentatively titled "The Digital Hand: Electronics and Literary Manuscripts."

My poetry friends on social media now concede that large language models (LLMs) can write poems on demand, but they insist that LLM verse lacks certain qualities essential to true poetry—its style, its emotion, its creativity, or something. Many writers and humanities scholars have taken such defensive stances in their early reactions to generative AI, but we should question this reflex. To clarify the stakes and possibilities of our responses to large AI models, I first suggest that they change the relations among the intellectual disciplines that readers of this journal sustain. Turning to the related but broader political situations of these technologies, I argue that many early responses to them have been either redundant with existing social thought or manifestly reactionary.

The advent of coherent text and image generation by large AI models may trouble distinctions between the sciences and the humanities. As my opening anecdote suggests, we celebrate the arts and humanities for their qualitative, subjective, unpredictable characteristics, which contrast with the quantitative, objective, predictive values of technoscience. Large AI models, however, can write poems and close readings of poems; they render humanities discourse computable. This fact might seem threatening to the scholars and artists who value their work for its apparent irreducibility to scientific calculation. If AI will force writers and graphic artists to find other ways to get money, that is bad news indeed, but these displacements of labor reflect deeper changes in the relations among intellectual disciplines.

Even as large AI models computationally mimic what had seemed like quintessentially human creativity, these devices themselves elude scientific understanding. Engineers designing and using AI systems now confront ambiguities, idiosyncrasies, and opacities more familiar

PMLA 139.3 (2024), doi:10.1632/S0030812924000518

to rhetoricians and literary critics than to computer scientists. Scholars in code studies have long recognized the qualitative, rhetorical dimensions of computer code, but their readings normally complement a technical understanding of it. By contrast, large AI models resist technical analysis in multiple ways. Most fundamentally, the core of an LLM consists of a vast matrix of numerical parameters that reflect statistical facts about the texts used to train it, but the size of such matrices and the high dimensionality of the parameters make it infeasible to interpret them directly. Scholars can analyze the relatively few lines of ancillary code that make a model run, as well as the filters, settings, and fine-tuning that influence its outputs, but such efforts leave the central facts of the model obscure. Researchers directly interpreting model parameters use such blunt techniques as ablation, a surgical term for cutting away tissue, which in this context means removing a section of the parameter matrix and then checking for loss of function. This method recalls the story of Phineas Gage, a railroad worker who experienced cognitive and behavioral changes after an accident in 1848 that drove an iron rod through his skull, enabling early neurologists to debate which parts of the brain do what. In sum, as the computer scientist Matt Welsh puts it in a recent lecture, “Nobody understands how large AI models work.” Anyone who knows HTML can edit a website arbitrarily, but engineers lack such thorough control over large AI models. For now at least, their parameters remain terra incognita. If my Latin seems overwrought, then “Here be dragons” might do as well: the image generators have spawned such monstrous visions as Loab and Crungus, grotesque creatures called digital cryptids whose origins in the parameter matrices remain obscure.

As a result of these interpretive limitations, software designers often interact with large AI models in qualitative, rhetorically nuanced ways. The new field of prompt engineering can be understood as the rhetorical study of how AI models respond to inputs. Some early findings are so intuitive as to imply (wrongly) that LLMs have mental states. For instance, appending “Let’s think step by step” to a prompt improves quantitative and logical accuracy (Kojima et al.). Other findings are bizarre: certain strings of gibberish can jailbreak an LLM, enabling illicit outputs (Zou et al.). The qualitative, expressive dynamics of model prompting may feel more familiar to literary critics than to computer scientists. Welsh describes an LLM function he calls KidSafe, which instructs the model, “Take whatever you’re given and rewrite it so that it’s OK for kids,” perhaps by removing violence and sexuality. Facing an audience of programmers, he continues, “I challenge anyone to write down the algorithm for that. . . . But the language models have no problem with this.” Scholars of human-computer interfaces often use qualitative methods, but even expert programmers now find their interactions with computers grounded not in the mechanics of code but in the rhetoric of natural language. This rhetoric might get formalized and regularized through automated prompt transformation, which optimizes users’ requests before passing them to the model, but such techniques derive pragmatically from model behavior, not from parameter interpretation.

While computer scientists grapple with the new salience of qualitative methods in their field, large AI models have also shifted the border between humanities and sciences in the opposite direction. In their prompt for this *PMLA* special feature, Matthew Kirschenbaum and Rita Raley call our attention to tokenizers, the software modules that parse strings of text into numbers for a model to do math with. As they note, emerging methods of tokenization produce counterintuitive ways to mathematize natural language. Casual explanations of LLMs as capturing statistical relations among words subtly misrepresent the technology. Advanced tokenizers assign numbers not to individual words, as a human would, but to clusters of letters, spaces, and punctuation—not only familiar digraphs like *th* but also weird strings like *dv* or *ug*. Tokenizers seem not to cut language at the joints, or not where we think the joints should be. By finding computational efficiencies, however, they reveal novel quantitative facts about language. The tokenizer and LLM yield a speculative but empirically grounded insight: although natural language is predictable, the means of prediction might remain obscure to our intuition.

Humanities scholars, especially those studying interpretive theories and methodologies, are uniquely equipped to address the challenges that large AI models pose. We embrace literature for its resistance to rigid logic, its nuance and ambiguity, and modern criticism offers a robust tool kit for rigorous thinking at the limits of knowledge. When today's media theorists consider what it means to use software that fabricates evidence and behaves unpredictably, they join a long tradition of analyzing technologies that deceive us and evade critical scrutiny, a tradition that includes the likes of Martin Heidegger and Plato. When a computer scientist declares that "the hottest new programming language is English," or when a journalist puzzles over how an algorithm seduces him into treating it as sentient, they underscore what key roles the rhetorician, the critical theorist, and the philosopher of mind can play in the reception of these technologies (Karpathy, Roose). These circumstances may indeed signal job opportunities for humanities majors, but we should not miss their broader implications for the scope of our disciplines.

Existing theories and methods also equip us to address some consequences of large AI models for language, literature, and pedagogy. As Kirschenbaum and Raley note, the prevalence of *ChatGPT* as popular shorthand obscures the complexity of algorithmic language infrastructures that have shaped human discourses for decades and more. The difficulty of saying when and where this informatics of language began, or what discourse now escapes its grasp, underscores the deep systematicity of language and all it sustains. Ted Underwood argues that LLMs give empirical proof of a thesis that poststructuralists gave us theoretical reasons to believe decades ago: "the thesis that language is not an inert mechanism used by individuals to express their thoughts but a system that actively determines the contours of the thinkable." When *ChatGPT* seems uncannily sentient, Underwood suggests, we should recall the links that Barbara Johnson drew between "the machine-like grammar of textuality" and "the subject's function in language" (Johnson 79). Citing Michel Foucault and Roland Barthes, Underwood concludes that "the social theory of meaning we need to understand this technology took shape long before the technology itself." At its most transformative, theory seeks not merely to demystify the textual conjuration of subjective presence, which might satisfy AI skeptics today, but to critique subjectivity itself as the conjuration of a certain mechanism, "all the way down," as Kirschenbaum and Raley put it. The greater mistake is not to see machine-written texts as expressive of consciousness but to see our own subjectivity as uninvolved with the mechanistic underpinnings of language.

For literary criticism, the rise of LLMs should further motivate the already decades-old return to aesthetics, including judgments of value. It has become easy to generate coherent texts of any length in any familiar genre. With the issue of quantity rendered moot, that of quality remains. Kirschenbaum and Raley ask, "Are we sure we know what 'good' output actually is?" This question has preoccupied aesthetic criticism for centuries, and it should not remain so marginal to literary studies today. Having earlier dismissed my poetry friends for saying that machine-written poems lack a certain something, I concede that we should join them in trying to say what makes good art good, but we should seek answers that avoid politically backward ideas about the human, the creative, the authentic. However we approach questions of aesthetic value, we should abandon the belief that aesthetics inherently provides cover for elitism, quietism, or conservatism and recognize it as a politically enabling field.

In the classroom as well, LLMs might heighten attention to the quality of student writing over its quantity. As a student, I viewed the length of essay assignments as a major indicator of difficulty. To some degree, I still do. Perhaps we should ask more advanced students not for longer essays but for better ideas. We ostensibly ask them for both, but in practice, it often seems easier to count pages than to evaluate arguments. As Kirschenbaum and Raley suggest, to define "better" in such contexts is a genuine intellectual task.

As to the connected but broader political consequences of AI, some recent efforts to address these seem redundant with existing lines of social thought. For example, while I recognize environmental and labor politics as urgent concerns, large AI models do not transform either. In the most famous critique of LLMs, Emily M. Bender and coauthors note that training an LLM has been "estimated to require as much energy as a trans-American flight" (612). Every ton of carbon matters, but do they know how many tens of thousands of flights crisscross the North American continent every day? The environmental economist Akhil Rao estimated the even more trivial carbon and water costs of using AI models. He equates the energy cost of making seventy-two AI-generated images with that of playing video games in HD for an hour. To eat a quarter-pounder, Rao reports, uses about as much water as seventeen thousand queries to *ChatGPT*. We should certainly worry about the environment, and the continuing growth of model size and usage will increase their effects. Large AI models join other resource-intensive computational infrastructure, such as cryptocurrency and other blockchain technologies, in hastening our planetary ruin. But while AI companies have motives to pursue efficiency, the most popular blockchains use a proof-of-work system that rewards whoever computes the most, directly incentivizing consumption. A focus on the climate impact of AI both distorts the scale of this impact and overlooks the more substantive consequences of this new technology. Similarly, Bender and others raise valid concerns about the invisible human labor of fine-tuning and filtering AI platforms. This work pays little, often involves objectionable material, and exacerbates geopolitical inequalities. But it differs neither in quality nor in scale from the microtask and moderation labor economies that have supported social media and other digital infrastructure for years. The critique of labor politics in the tech industry—like the discussion of environmental footprint—remains urgent but bears no special relation to large AI models per se.

These new technologies do intensify and transform other familiar problems. If AI produces racist, sexist, and otherwise hateful language, we can thank the human texts it mimics, but AI models do worse than duplicate and amplify poisonous human discourses. They enable new kinds of harm, such as bespoke propaganda automatically tailored to individual readers; deepfake pornography, which differentially affects women and gender-nonconforming people; or the expanding use of proprietary, opaque algorithms in decisions about such matters as mortgage and insurance rates, policing strategy, and criminal sentencing, which entrenches racial and other inequities. These problems will get worse. Their legibility as problems, however, suggests that politically engaged humanities scholars are well equipped to respond.

I close by addressing one area where responses to large AI models have been surprisingly reactionary—that of intellectual property. Objections to large AI models on copyright grounds amount to a dramatic reversal of attitudes among the American left, what Kirschenbaum and Raley call a "new copyright fundamentalism." Two decades ago, many people now urging copyright infringement claims against AI firms were deriding Metallica and the RIAA for their lawsuits against Napster and its successors. Back then, piracy was cool; we called it "sharing." To defend intellectual property rights against new technology was a sellout move. The recent shift in favor of policing intellectual property has all the pathos of defending carriage makers against the rise of the automobile, but its consequences are worse. As Kirschenbaum and Raley note, the new copyright conservatism "necessitates a concession to a market economy for culture," which reduces artistic production to fungible labor time and expresses the value of artworks in dollars. Art making is far older than capitalism, and we should not degrade the former by reducing it to the latter. Copyright conservatism abdicates the commitment, once familiar in left political circles, to the open sharing of knowledge, of techniques, and indeed of art itself, a commitment for which the hacktivist Aaron Swartz died and many others have suffered. The utopian rhetoric of "free culture" was not just an excuse to stop buying CDs. It was an effort to celebrate the collaborative, appropriative, transformative processes through which arts and cultures live (Lessig). Must every remix or collage be approved, licensed, and paid for? If lawsuits from the likes of George R. R. Martin and Jodi Picoult suggest they cannot pay the bills otherwise, then we should respond with better ways to value creative work, not old weapons of cultural constraint. Of course, we owe Silicon Valley nothing: we should not only support public datasets and oppose what Kirschenbaum and Raley call "enclosure of the language commons," whether by tech firms or the people suing them, but also hack, pirate, and otherwise mess with the proprietary AI models however we can.

For now, case law in the United States appears to favor AI companies. Although the Supreme Court's recent decision in *Warhol v. Goldsmith* narrows the scope of fair use, some legal scholars argue that the use of copyrighted texts to build computer systems remains protected, even if the texts were illicitly obtained (Henderson et al.; Samuelson et al.). Verbatim reproduction of copyrighted works, as alleged in the New York Times suit against OpenAI, does seem illegal, especially if it reduces the market for such works. (Given that LLMs are not designed or used to store individual texts, verbatim reproduction will likely remain marginal.) By contrast, stylistic imitation of human artists by AI systems, as by humans, appears permissible; the dismissal of most claims in the *Andersen v. Stability* class action suggests as much. Most troubling of all are the claims, still pending in *Andersen* and elsewhere, that AI firms break the law by using copyrighted works to train their models. Setting aside the legal merits, how do such claims differ in principle from the idea that a person must not enter a library, read a few books, and leave with the knowledge they contain?

The copyright police will rightly say that language models are not people, but both operate within networks of language technologies that affect what we can read and write. Both generate complex but predictable language based on texts we have previously read. To view the self as fundamentally, not just incidentally, entangled with textual machines is an authentic intellectual challenge, one posed in theory decades ago and in practice today. As we grapple with this challenge, corporate interests will gladly leverage reactionary defenses of intellectual property to further enclose and commodify information of all kinds. Those on the political left must not help close our remaining open libraries, whether in buildings or on servers, by adopting the wrong language of value for their contents.

Works Cited

Bender, Emily M., et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜." *FAccT 2021: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Mar. 2021, pp. 610–23.

Henderson, Peter, et al. "Foundation Models and Fair Use." *Journal of Machine Learning Research*, vol. 24, 2023, pp. 1–79, www.jmlr.org/papers/volume24/23-0569/23-0569.pdf.

Johnson, Barbara. "Rigorous Unreliability." *Yale French Studies*, no. 69, 1985, pp. 73–80.

Karpathy, Andrej [@karpathy]. "The hottest new programming language is English." *X*, 24 Jan. 2023, twitter.com/karpathy/status/1617979122625712128.

Kojima, Takeshi, et al. "Large Language Models Are Zero-Shot Reasoners." *Advances in Neural Information Processing Systems*, no. 35, 2022, pp. 22199–213, arxiv.org/abs/2205 .11916.

Lessig, Lawrence. *Free Culture: How Big Media Uses Technology and the Law to Lock Down Culture and Control Creativity.* Penguin Books, 2004.

Rao, Akhil. "AI Is Like a Very Tiny Hamburger." *The Efficient Frontier*, 15 Jan. 2024. *Substack*, akhilrao.substack.com/p/ai-is-like-a-very-tiny-hamburger.

Roose, Kevin. "A Conversation with Bing's Chatbot Left Me Deeply Unsettled." *The New York Times*, 16 Feb. 2023, www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html.

Samuelson, Pamela, et al. "Comments in Response to the Copyright Office's Notice of Inquiry on Artificial Intelligence and Copyright." US Copyright Office, 30 Oct. 2023, www.regulations.gov/comment/COLC-2023-0006-8854.

Underwood, Ted. "The Empirical Triumph of Theory." *Critical Inquiry: In the Moment*, 29 June 2023, critinq.wordpress.com/2023/06/29/the-empirical-triumph-of-theory.

Welsh, Matt. "Large Language Models and the End of Programming: CS50 Tech Talk with Dr. Matt Welsh." *YouTube*, uploaded by CS50, 29 Oct. 2023, www.youtube.com/watch?v=JhCl-GeT4jw.

Zou, Andy, et al. *Universal and Transferrable Adversarial Attacks on Aligned Language Models.* 27 July 2023, llm-attacks.org.