Text, not Language: Inexpert Provocations on Language Models

Chris Forster

2023-03-16

`http://cforster.com/2023/03/llms-post`

Like lots of folks, I've been watching the discourse and debate about ChatGPT in particular, and about Large Language Models more broadly, with great interest. With the release of GPT-4 a couple days ago (as I was trying to write this), the already torrential discourse about these objects has only intensified. It feels impossible to stay on top of the news stories, blog posts, and demos related to these technologies. People are discussing these things with an intensity typically reserved for defining *digital humanities*.

I've felt dissatisfied with the existing discourse in ways I've had a hard time articulating; this post tries enunciate that dissastisfication in the form of a series of provocations, none of which are sufficiently justified.

For what it's worth (and as much as a record for myself), as I write this, these are the essays, articles, and posts that I've been thinking about most in the last couple weeks:

- Writings by Emily Bender and others: Bender and Koller, "Climing Towards NLU: On Meaning, Form, and Understanding" ; Bender et al, "On the Dangers of Stochastic Parrots" ;

this profile of Bender
- Ted Chiang, "ChatGPT Is a Blurry JPEG of the Web"
- Chomsky et al., "The False Promise of ChatGPT"
- Ted Underwood, "Mapping the Latent Spaces of Culture"
- Stephen Wolfram, "What is ChatGPT Doing… and Why Does it Work?" (this is the single best introduction to the technology behind ChatGPT, etc).

Provocations

1. **What "intelligence" is here is human, not machine**: Emily Bender, among others, notes that the phrase "artificial intelligence" is hype; and while I am not among those folks who are unimpressed by these models, I think it is useful to recognize how human, and indeed, cultural these objects are. If these models are able to produce output that seems to be motivated by knowledge (or intelligence), it is because of the massive amounts of text on which they were trained, text that was itself generated by (human) intelligence. They are representing, and re-representations, of human intelligence rather than being *intelligent* themselves.

   Everyone has their preferred metaphors here; I've seen people describe large language models as essentially extractive technologies, drilling into the existing and accumulated textual record to extract a set of tokens, weights, and probabilities ("text mining," as they say). I've been trying on a slightly different metaphor, imagining them as a sort of massive act of statistical casting, pouring a gelatinous network of neuronal goo into the structured chaos of text on the internet, and allowing gradient descent to gradually harden it into a model. My point, though, in either metaphor, is that what we are marvelling at is less the technology of extraction or casting, but the extracted (or cast) material.

2. **This knowledge is social, not individual**: Moreover, the

knowledge these models represent is distributed and social; it cannot be owned by one entity, nor localized in a single space. Indeed, since these models are trained on internet text, they are trained, at least potentially, on all of us. As Saffron Huang writes, "We are each, in a small way, an author—perhaps more aptly, a ghostwriter—of ChatGPT" (Huang). Of course, *exactly* who is an author of ChatGPT is impossible to say, without a reckoning with the training data and its sources.

Their essentially social basis is, I would conjecture, a source of their power. As direct models of collective "text use", and so as indirect models of collective knowledge, the model can represent "more" than any one person knows. Yet, rather that seeing them as Borg-like, they seem equally describable as embodiments collective/historical intelligence, of the sort that we might also call *culture* or *ideology*.[1]

3. **Chat as an interface is *not great***: To judge from results, if you're trying to productize an LLM, chat *is* the killer app for marketing the large matrices of numbers that at some level *are* an LLM.A blog post by Peter Levine making an apposite point comes to my attention via Mastodon as I'm writing this. By adopting the conventions of human conversation, chat bots from ELIZA forward have proven to be shockingly compelling.

   Yet, the representation that motivates these technologies is not *essentially* conversational. ChatGPT takes a large-scale representation of "meaning" and then uses reinforcement

[1] If we understand these technologies as essentially extracting, or representing, some common, social, object, then we might follow Huang in asking, "is it right that private companies which draw on the thoughts and words of the commons, then sell ads and distribute flattened opinions back to the very same public?" . In so much as we understand language, and the text of the internet, as "public things," it is hard not to see these technologies as acts of enclosure.

learning from human feedback (RLHF) to make interacting with it feel conversational (and, I take it, to sand off the rough edges). This is, though, only one particular *interface* for these models. The model itself, as Benjamin Schmidt notes, "doesn't really participate in a conversation—it doesn't even know which participant in the conversation it is!" (Schmidt). As an interface, though, "chat" predisposes a user to treat the model as a sort of superhumanly knowledgable person. It inspires an immediate anxiety—is it smarter than *me*? Could it do *my job*? (Does it want me to leave my wife? (Roose)) All the specular logic of the doppleganger—jealous rivalry!—is right there, as a consequence of this interface choice. But these models are far more like a talking library (a weird, peculiarly curated library) than they are a person.

4. **These are not models of *language* but models of *text***: We call these "large language models," but they are not, in fact, *models of language*. They are models made from language, and which make language. But they have "learned" language the same way that they have "learned" the other things they are able to produce compelling output about. They generate well-formed, grammatical sentences because grammaticality is a property of the training data. That training data, though, is better understood as *text* rather than *language*. This accounts for the ease with the models seem to generate non-natural language text (say, computer programs). The *tokens* that GPT learns, for instance, are not words or any other linguistic objection, but are a product of a tokenization scheme that has proven useful for this task (see the OpenAI tokenizer).

   And the text data on which they are trained is not *language*. This is true in a few senses; its data is exclusively *written* (rather than being spoken or signed), and so it is a narrower slice of language, shaped by its medium; moreover it seems to include some things that strain the definition of *language*—

output logs from software, say (see the "weird tokens" below). I don't mean to be dogmatic, but it seems useful to at least acknowledge that this text data exists in a weird relationship to language as conventionally construed.

But my more essential point is that the same way these models learn grammar and linguistic meaning, is how they are also simultaneously learning the plot of *Hamlet*, what makes a joke funny, and the forty-seventh digit of pi. It absorbs these things through the same stochastic processes. As such it seems better to describe them as models of *culture*: a culture that carries grammars and languages and genres along with it; a culture that surely maps to no existing lived community of humans—a culture whose borders and contents remain unclear so long as the training data is undisclosed; but like a culture, it mixes languages and knowledges together in ways that are not easily separable. This, I think, is fundamentally a point Ted Underwood has made; we should understand "neural models as models of culture rather than intelligence or individual language use" (Underwood).

5. **They are made out of bias**: Understanding these models as models *not* of language or *knowledge*, but of culture, touches on the vital question of LLMs and bias. Many talk about the bias that these models can learn through their training, and then propose ways of mitigating or managing it. Yet, is there a meaningful distinction in LLMs between *knowledge* and *bias*? These models are *made of bias*; there is a fundamental continuity between the ways that model is able to output grammatical sentences, make observations about the color of sky, and the other biases that these models contain. All were learned the same way.

6. **These are objects to be interpreted**: Looking over my points, I'm surprised at how critical I sound, because I find these objects *really fascinating*, and I am convinced that people like

literary and cultural historians should be more and better engaged with them (and *not* primarily because these models may soon be authoring a portion of all terms in humanities classes).

Understood not as intelligences but as texts, produced as all texts are—out of other texts, and made of biases that inhere in the same substrate as language, these are fascinating objects. I find myself fascinated precisely because I am impressed and I want to understand how they're working—not just how the model was trained, tuned, and reinforced, but how can we understand *exactly* what is being modeled? This is something that would require us to compare a model to its training data; but of course that alone would hardly be sufficient, because the training data is simply so massive that we couldn't *read it* even if we wanted. Are there ways to read the models?

The example of "weird tokens" or "glitch tokens" (well dicussed in this computerphile video, and elaborated with interesting examples here) provides one way of approaching them that is not simply chatting with them: finding strategies for examining the strange weird castings of the training data. Such an approach should not be totally alien to anyone involved in the study of texts, their interpretation, and their history. It is the inaugurating assumption of fields like literary studies that the texts we read remain strange and alien to us, worthy of additional consideration, study, and exploration. And it seems useful to find ways of adopting such an attitude to these new textual objects.

Works Cited

Bender, Emily M., et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜." *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ACM, 2021, pp. 610–23, https://doi.org/10.1145/3442188.3445922.

Bender, Emily M., and Alexander Koller. "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*, Association for Computational Linguistics, 2020, pp. 5185–98, https://doi.org/10.18653/v1/2020.acl-main.463.

Chiang, Ted. "ChatGPT Is a Blurry JPEG of the Web." *The New Yorker*, Feb. 2023.

Chomsky, Noam, et al. "Opinion | Noam Chomsky: The False Promise of ChatGPT." *The New York Times*, Mar. 2023.

Huang, Saffron. "ChatGPT and the Death of the Author." *New Statesman*, Feb. 2023.

Roose, Kevin. "A Conversation With Bing's Chatbot Left Me Deeply Unsettled." *The New York Times*, Feb. 2023.

Schmidt, Ben. "You've Never Talked to a Language Model." *Ben Schmidt: Blog*, https://benschmidt.org/post/2023-02-19-sydney/, Feb. 2023.

Underwood, Ted. "Mapping the Latent Spaces of Culture." *The Stone and the Shell*, Oct. 2021.

Weil, Elizabeth. "You Are Not a Parrot." *Intelligencer*, https://nymag.com/int artificial-intelligence-chatbots-emily-m-bender.html, Mar. 2023.

Wolfram, Stephen. *What Is ChatGPT Doing … and Why Does It Work?* https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/, Feb. 2023.