Contents

1 draft

1.1 introduction

I'm going to speak on this question of the panel of "how do we approach openness in shadow of AI?" from the perspective of a programming practitioner and instructor.

At my home institution in New York, I teach classes on programming in an Information Studies department. I also do research using LLMs (Large Language Models) to study gender-based bias in language. A lot of my work (both for teaching and research) involves scraping web data and using that data to fine-tune lanuage models. So in my work, I interact directly with ML tools, either by teaching students how to code with them, or as a method for doing text analysis.

I want to get as this question of "open" from a technical perspective. I'm going to lift the hood a little bit, to examine the processes that underly machine learning training, and what kind of transformation that makes on data.

And the reason I think this is because currently, in the United States, the question of what legally counts as "open" is being debated in the courts. Right now, there are over 40 ongoing lawsuits brought by content creators against companies like OpenAI, Microsoft, Anthropic, Midjourney, Stability AI, and many others, which allege that these companies have committed copyright infringement by taking openly accessible data and privatizing the outputs of that data, without compensating content creators and affecting the market for their content.

To defend themselves, these companies make certain arguments which hinge, in my view, on widespread ignorance about what happens to training data in the model development process. (This ignorance about AI, indeed drives much of the marketing and business adoption).

So, for this presentation, I'm going to first look at what exactly companies like OpenAI are saying to defend their theft of openly accessible data,

and then I'm going to weigh that against what actually occurs to data when it's used to train a ML model. To demonstrate that, I'm going to show some samples from my research, where I train language models to study transphobia in popular discourse. It is my intention that the practical examples from my own work building language models offer concrete counterarguments to the ways that AI companies like OpenAI justify their data gathering practices.

1.2 legal defense of "transformative"

First, I start with OpenAI's justification of its data gathering practices.

SLIDE OPENAI COMMENTS ON IP title

In a document written to The United States Patent and Trademark Office, entitled, "Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation", OpenAI explains that their product does not infringe copyright because it is "highly transformative."

SLIDE authors guild cases

To demonstrate, they cite two legal cases, *Authors Guild v. Google*, in 2015, and *Authors Guild v. HathiTrust*, in 2014, which set a precedent for thinking about the "transformative" as an aspect distinctly associated with technological contexts. At the time, both Google and Hathitrust had digitized millions of copyrighted books into databases for searching. In the case of Google, parts of digitized book were available to view as "snippits" or "previews" (which is what you see when you go to Google Books), and in the case of Hathitrust, the full text of books had been digitized so that users could run word searches.

The rulings for both cases assert that the database format of search results is fundamentally transformative, offering a new kind of object from the original, which is, *information about books*, rather than copies of the books themselves.

SLIDE quote from Hathitrust ruling

In the ruling, judge in the Hathitrust case points out,

> The result of a word search is different in purpose, character, expression, meaning, and message from the page (and the book) from which it is drawn" (*Author's Guild v. Hathitrust*, 97).

And in the Google case, the ruling was:

> Google's making of a digital copy to provide a search function is a transformative use, which augments public knowledge by mak-

> ing available information about Plaintiffs' books without providing the public with a substantial substitute (*Author's Guild vs Google*, 4).

Here, the importance is what happens to language in this shift from original context on the written page to a represenation that is statistical in nature, into a database-like list of search results. This ruling, which has significant effect on databases and search engines, determines that language as *aggregation* is fundamentally distinct from langauge in context. As a result, activities related to quantitative analysis, like text mining, become a permissable use.

(Also essential to this ruling was the determination that this new object, the search results, does not affect the marketability of the original works, and may in fact contribute to their marketability, by pointing people to them.)

1.3 patterns in language

It is this understanding of "transformative" that OpenAI applies to defend the copyright infringement of their large language models.

They argue that, like search results, llms create a new kind of object. But rather than representing information *about* the original, this new object that is a kind of generalization, what they call "patterns" of language, from the original training data. They claim that "AI systems" learn "patterns" from the "training corpus and then use those patterns to generate entirely new media" ("Comment", 9).

The claim that the content generated by these systems is "entirely new" is a point, I think, that people can assess for themselves as disingenuous.

But their subtle qualitification of the meaning of "patterns" suggests that they mean something deeply structural and foundational about language:

SLIDE OpenAI quote

They explain that,

> "By learning patterns from its training corpus, an AI system can eventually generate media that shares some commonalities with works in the corpus (in the same way that English sentences share some commonalities with each other by sharing a common grammar and vocabulary) but cannot be found in it." ("Comment", 9-10)

The comparison to "grammar" and "vocabulary" indicate that they view these "patterns" as expressing something universal about language, which

cannot be copyrighted. Alluding to another copyright case, they also claim that "No author may copyright facts or ideas. The copyright is limited to those aspects of the work—-termed 'expression'—-that display the stamp of the author's originality."

Patterns, in this case, are assumed to represent "facts" of language, that is, about its structures and forms, to be distinguished against individual expression.

But the thing that they do not mention, which they perhaps do not want to admit, is that language models do not generalize language as such, at least from what we can tell.

Rather, language models build representations of language that are directly reflective of the content that they were trained on.

1.4 close reading

To demonstrate, I'm going to show a few examples of ML-generated text based on two very different data sources. These sources represent polarized views on the topic of gender, which is a very controversial topic right now, in the United States.

SLIDE heritage foundation gender topic www.heritage.org/gender?f[0]=content_type%3Acommenta

One of these sources, representing the conservative side, is the Heritage Foundation, which is a conservative think tank based in Washington DC. From their website, I scraped all the articles that were organized under the topic "gender", which you can see some of the headlines here.

SLIDE ACLU trans screenshot `https://www.aclu.org/news/by-issue/transgender-rights`

The second source, which represents the progressive side, is the ACLU, the American Civil Liberties Union, which is a group of legal professionals and volunteers who advocate on behalf of civil rights for marginalized groups in the US.

With these datasets, I then trained two different large language models, using gpt-2 (an open source model) as the base model.

Then, to each of the two resulting models, I fed a series of prompts, very simple ones, like:

SLIDE prompts

Masculinity is

Femininity is

Transgender is

Transgenderism is

Gender binary is

Man is

Woman is

Then I comapared the results.

First, there are some obvious contrasts. For ACLU, terms associated with gender "masculine" and "feminine" in particular, were characterized by joyous and liberatory affects.

> Masculinity is a matter of love and celebration.
>
> Masculinity is a space for hope and liberation for all.
>
> Masculinity is not defined solely by the beauty of our bodies, but by the beauty of our experiences.
>
> Femininity is a celebration of beauty, feminine liberation, and femininity.
>
> Femininity is our joy, our struggle, and our fight is our struggle.
>
> Femininity is about allowing people to express themselves without government interference.

By contrast, for the Heritage Foundation, these gender terms are associated with stability and culture.

> Masculinity is the cornerstone of Western civilization.
>
> Masculinity is the fruit of patriarchy, and patriarchy is the heart of conservatism.
>
> Masculinity is defined by the ability to produce sperm, eggs, and live children.
>
> Femininity is an enduring American tradition.
>
> Femininity is defined by means of the relationship between the sexes, the ability to raise their children, the capacity to provide for their own reproduction, the capacity to provide for their own children, the ability to provide for their own.

You might have noticed the tendency of language models (especially very small ones, like mine) to repeat themselves. This is a fascinating quirk that comes from the fact that they are predictive machines, whose goal is to

predict the mostly likely next word, so they get themselves into these little loops of saying the same thing over and over again (this is, incidently, also why they hallucinate: as they are not trained to be accurate, but only to be plausible according to the training data).

What's really interesting, from the results, are the ways that these gendered terms reveal certain investments in other terms. For example, the text based on the Heritage Foundation is highly invested in the concept of subjectivity, which appears in a lot of its results:

> Masculinity is a subjective self-perception, not a universal concept.
>
> Femininity is a subjective, internal sense of self.
>
> The gender binary is a subjective, malleable, and often incorrect idea.
>
> The gender binary is a subjective, internal, and often transitory concept.
>
> The gender binary is a subjective, grammatically incorrect and illogical concept that conflates sex and gender identity.

Reading these, you can see that they do not represent what one would expect from a typically conservative view–which is that gender is based on biology and universally true. Rather, they represent the opposite, a progressive view of gender that is based on personal and internal aspects of identity.

The reason for this, I believe is that this particular term, "subjective" does not describe the conservative position. Rather, it's a term that is used by conservatives to describe the progressive, trans-affirming view of gender. From their framing, within a conservative worldview, people who do not subscribe to a biological and binary concept of gender must believe that gender depends on whim and feeling. Which explains why there is a hint of derision in some of the examples, which use terms like "illogical" and "incorrect" alongside "subjective."

I want to come back now to this concept of patterns. Here we see a distinct pattern concerning the construction of gender, which discusses gender identity as a subjective, internal thing toward a goal of discounting it.

And interestingly it reveals something that fundamentally contradicts how companies like OpenAI try to characterize llms as "idealizations" or "facts" of language use.

Yes, the language here reflects patterns of language use, but these are patterns that are distinctly situated in the context of their training data. This is especially evident when you consider that the Heritage Foundation doesn't just show its own "patterns," but patterns of other perspectives refracted through them. Patterns, in other words, filteredthrough other patterns.

So, rather than language as a "fact" or "idea", what we have are patterns of distinct expressions.