Large Language Models: Trust and Regulation

David Banks, Dept. of Statistical Science, Duke University
Costanza Bosone, University School of Advanced Studies (IUSS) Pavia
Bob Carpenter, Flatiron Institute
Tarak Shah, HRDAG
Claudia Shi, Dept. of Computer Science, Columbia University

Abstract: Large language models have burst onto the scene, and may do much to change the way the world operates. They can write, illustrate, and are rapidly adding new capabilities. This panel discussion brings together a group of experts to discuss ways in these tools might evolve, particularly in the context of how trustworthy they are and the issues surrounding their regulation.

Keywords: Disinformation, GPT, Human Rights, Job Loss, OpenAI, Trust

Introduction: This is a panel discussion among people who have been thinking about and exploring data science issues that arise from large language models (LLMs). David Banks is a professor in the Department of Statistical Science at Duke University, and an occasional editor of Statistics and Public Policy. Costanza Bosone is a graduate student at the University of Pavia who has been doing research on gender biases in GPT. Bob Carpenter (identified below as RC) is a researcher at the Flatiron Institute, and a major developer of STAN. Tarak Shah is a lead data scientist at HRDAG (the Human Rights Data Analysis Group). And Claudia Shi is a graduate student at Columbia University who has been studying the moral sense of about 45 different LLMs.

Discussion:

DB: What do you see as the major benefits of large language models (LLMs)?

CS: LLMs could level the playing field in education, and make it more accessible. They could provide quality education to every child, anywhere in the world. Also, LLMs could make the production of goods and services more efficient, and thus less expensive.

CB: Education is certainly an important area. In Italy, there is sometimes a pressing concern regarding teacher understaffing, leading to overcrowded classrooms, often compounded by the inclusion of students with special needs -for instance, those who struggle to maintain high levels of attention. The
implementation of a Learning Language Model (LLM) could prove beneficial. By providing individualized tutoring through the LLM, teachers could focus on guiding and ensuring the overall engagement of students, thereby fostering a better learning environment.

TS: I’m a data scientist who works on human rights, so I focus on the benefits in that arena. Two come to mind. We have been using an LLM for information retrieval and extraction, creating structured data from unstructured sources. We had such tools before, but with an LLM we can get up and running more quickly, without having to collect a lot of training data first. This is super important in human rights work, because a lot of the information related to accountability is contained in vast amounts of paperwork generated by bureaucracies.

The other thing is that as someone who writes code all day, we’ve always had IDEs that provide such help as autocomplete and suggest code that one should use, but Copilot raises this to a whole new level. It means one can code at a higher level of abstraction, and it is a faster on-ramp for people who are just starting to code in a specific language.

RC: We didn’t always have autocomplete and other help. I started with punch cards. What this makes me think of is back in the 1980s, when I used email to communicate and exchange papers, we had no Web yet. Then in the 90s, things started to change, and lots of people asked “What am I ever going to use email for?” and “What am I going to do with the Internet?” I have the feeling that we are right on the cusp of a similar phenomenon. We are asking what can we use LLMs for, and I think that in five to ten years, we are going to be asking “How could we have lived without this?”

For me, the main use I have for GPT now is research. Yesterday I was trying to learn about semiparametric Cox proportional hazards survival models. I went to books, I went to the Wikipedia, and what I found was rather opaque terminology. So I went to GPT and asked it to explain it to me. And the nice thing about GPT is that you can dialogue with it, asking it to drill down on the things you don’t understand. It starts to explain things in terms of measure theory, and you can ask it to recast the explanation in terms of delta functions and Lebesgue integrals---things I understand.

GPT writes all my data manipulation and graphics code now. I say “Here’s my data frame” and it is off and working. It does all the manipulation, and I can look at the output for reasonability.

I also use it all the time for brainstorming. I enjoy role-playing games, and it is just a genius. I say “I have a crew of scoundrels who are planning a heist in Victorian-era Venice---give me some ideas.” The results are amazing.

All the non-native English speakers here use it to clean up their writing. GPT is an incredible editorial assistant. I hope I never have to read another paper just to copyedit it. I think we are still figuring out all the ways in which LLMs can help us.

DB: I think you are spot on, Bob. Some of the best minds in our field don’t speak English as a first language, and LLMs go a long way towards removing that impediment. I am sometimes an editor, and a few months ago I received a paper. The math looked correct, but the writing was so poor that I couldn’t send it out for review. I emailed the author, asking him to work with a native English speaker, and the next day the paper came back with perfect English. He had run it through GPT-4.

But let’s turn to the kinds of risks that LLMs pose.
CS: There are near-term and long-term risks. Immediate risks are things like bias fake news.

In the long term, we may worry about LLMs-based agents that can interact with the real world. An uncontrolled agent with knowledge like existing LLMs can be very dangerous. There are some recent works that evaluate what LLM agents can do right now. They can’t spin up an Amazon Web Services instance yet, nor can they create a bitcoin wallet yet, but these harmful tasks seem plausible for future systems. A lot of researchers and policymakers are working towards evaluating these models. I think a lot of the long-term risks come from we don’t understand the capabilities of these models, and these models are being released without a full assessment (Kinnement et al., 2024).

I worry about the possible loss of plurality in different viewpoints. Bob made a good point about how creative LLMs can be, but we don’t know the ways in which their imaginations might be limited.

TS: These are all good points, but I am going to stick to my sphere. At HRDAG we often use multiple systems estimation to infer, say, the number of unreported civilian casualties in a conflict area based on the amount of overlap among several lists of victims. This work entails named-entity resolution, since there are often small differences in the details for people on two lists. For example, if one police report has a man named Alberto Sanchez aged 41 who was killed on April 14 in Caicedo by the FARC, is this the same person as Roberto Sanchez, aged 40, who is reported by the Catholic Church records as having been killed in Caicedo on April 13?

We already had tools for named-entity resolution, and often there is uncertainty about the results. With LLMs, they make things so easy that it is tempting to accept uncritically whatever comes out. LLMs can make it very easy to do bad analyses.

Copilot makes it easier to write code, but we don’t have corresponding tools for testing the code to ensure it does exactly what we need it to do. This was a problem before LLMs, but now LLMs may do it well enough that we get lazy about checking.

HRDAG works around the world, and some of these tools have been incredibly valuable when moving between English and Spanish. But for other languages, we have had situations in which the translation was imperfect and since I don’t speak those languages, it is hard to tell when that happens. LLM quality varies a lot across languages.

We need to learn more about how and when LLMs break down.

CB: There are always risks with innovation. I can imagine that soon people will be using ChatGPT as a personal assistant: it can correct text, collect data, list pros and cons of a decision, and so forth. But we don’t know how much we can trust it. If we are cautious and careful, I think most risks can be minimized. After all, a human personal assistant can make mistakes or mislead us too.

LLM bias is another problem. We know it exists, and many are beginning to study this.

The way we learn may have to change. LLMs are getting better at writing proofs and essays. If we think those are skills we want people to have, we shall need to ensure that educational systems have ways to teach those skills in a world with LLM assistants.
RC: I think the risks are the flip side of the benefits. These tools are giving people a lot more leverage. They give people more reach, and in some sense that is a lot like the Internet or Meta. We have democratized the ability to get content out in the public domain.

I think one of the risks that we’ve not talked about yet is job loss. LLMs will make people more productive, and in the short term at least, that will mean that someone will lose a job. This happened to the secretary pool when computers came in. I think LLMs will take over a lot of low-level writing and coding tasks. Right now, they won’t displace really good writers or really good programmers, but LLMs are getting better and that could happen.

LLMs are limited in their ability to manage large projects because they have essentially no memory. When we start adding that feature, they will be able do a lot more. When we give them access to a terminal or a wallet, things will change in ways that I cannot foresee.

DB: The President’s Council of Advisors on Science and Technology put out a call for input on ways to mitigate the risk of disinformation and deep fakes, because LLMs can easily produce these. The American Statistical Association responded with a number of suggestions (cite). We shall probably need to learn how to curate our information in a more formal way. If we all learn that we need to have fact-checking as a regular part of our lives, that could be a good thing for democracy.

DB: Can LLMs be regulated?

CB: I think there are two dimensions in your question.

We don’t know how a LLM is generating text, which may lead to biases or hallucination—a when the LLM creates nonsensical or wrong content. One kind of regulation might be to require companies to put the LLM code in the public domain, which is of course unlikely. Developing such codes requires time, energy and long-term investment, and different stakeholders are involved, so with different players in the game, whether it’s fair to share or not needs careful discussion.

The other aspect is how and where data are stored. In other words, who can access AI-stored data and where are such data stored? Unpacking this scenario involves investigating into the complexity of data storage and retrieval. In this regard, regulations aiming at clarifying such aspects could be promoted.

RC: I honestly don’t know how one would regulate LLMs. I’ve thought about it. Regarding Costanza’s point, even if one shared all the code and data, that still wouldn’t change behavior. Other strategies would be to restrict access to LLMs, or, in cases, ensuring access, to stop tech companies from hoarding this capability.

I think people want to regulate LLMs to prevent harm, which is as tricky as trying to regulate the Internet. I don’t see much success in that arena. If you use an AI to do something illegal, then you are doing something illegal. Do we need to separately regulate the AIs? I’m not a lawyer—I don’t know.

I am curious to know what kinds of regulations people are discussing. And it is unclear how general regulations could be enforced.

CS: I defer to governance experts. One governance expert that I admire is Gillian Hadfield at the University of Toronto. She wrote about using regulatory markets for AI governance --- "where governments require the targets of regulation to purchase regulatory services from a private regulator"
(Hadfield and Clark, 2023). She wrote about it in a recent paper called: Regulatory Markets: The Future of AI Governance. I recommend it!

CB: LLMs are gaining and have already gained the ability to access the Internet and use various tools. But it can go the other way, too. Some people want to ensure that an LLM does not have access to certain documents or websites. Since August 2023, the New York Times and The Washington Post have prevented ChatGPT from training on their content. In an article published on the Washington Post (Tiku, 2023) it was suggested that publishers could be paid in exchange for granting access to the chatbot. This speaks to regulation as well. A university might want to protect its research or prevent premature release of research papers, or a journalist might want to protect a scoop. Like the Internet, LLMs have made it much easier and much less expensive to acquire and distill information.

TS: That engages with questions of privacy and security. HRDAG sometimes acquires personal testimony about killings and disappearances, and we encrypt such data. We have an obligation to the data and the people who provided it. So the decision process for using an LLM to translate or curate or summarize such data is not very different from the decision process we use to choose whether to use Google Maps Geocoder or a translation service.

Often we have legal documents that contain a great deal of identifying information about, say, a witness to a human rights violation. If, somewhere down the line, a next-generation LLM were to somehow access or reconstruct such documents, that would be a serious problem. It may be that governments and businesses will decide to keep their most sensitive information far away from LLMs.

DB: If an LLM could access trusted sources on the Internet, it could fact check its statements, reducing the problem of hallucination. And if an LLM is used for education, one could require that it not make a factual statement that wasn’t supported by one of the 27 textbooks approved for that state. There is a nice article on LLMs for data science education in Tu, Zou, Su and Zhang (2024). And there is a thoughtful discussion of payoffs and perils of using LLMs for data analysis in Glickman and Zhang (2024).

RC: Allowing an LLM to access the Internet only gives it more words and context. Doing what you suggest, requiring that the LLM source its statements to 27 specific textbooks, would be very hard to engineer into the current architecture of an LLM. You would need a system as smart as GPT on the outside, acting as a watchdog.

CS: One question is whether the LLM should access the Internet. A separate question is whether the LLM should react to material on the Internet. If it can generate text and then publish it on the Internet, that could be a very bad situation.

The current generation of LLMs have been trained on text produced by humans. But the next generation will be trained on a lot of text produced by LLMs. Going forward, it will be difficult to figure out what will be the effect of giving LLMs the ability to react to the Internet.

But I agree with what Bob said. To solve the problem of truthfulness, you probably need to pose it in a context. You can give the LLM a database of 27 textbooks, and check whether a claim is supported by one or more of them. I don’t know how computationally expensive that would, but it is definitely computationally expensive. If, instead, one could fine-tune the model by training on the 27 textbooks, that would be much less expensive. There is a tradeoff between accuracy and compute time.
DB: What do you all think data scientists should be thinking about regarding LLMs?

TS: I noted my experience with LLMs in terms of information retrieval, and I think that is a really important use case for data scientists. This sidesteps some of the issues with truth and ground truth. We want the answer with respect to a specific document collection, which is not the same thing as deciding whether or not the documents are true.

We've been using LLMs to surface information for human review. I am very focused on starting with unstructured data, and putting it into a structured format. As a data scientist, that seems like a very important application.

CB: For me, the big application is coding. Very often people give me a code that doesn’t work and sometimes even the purpose of the code is nebulous. But when I give the code to ChatGPT, it often finds the errors quickly, saving me a lot of time. For the moment, I think it is super fine.

But it comes back to how much should we trust a LLM. Sure, I trust it to tell me that I forgot to close a bracket, but I don’t know if I can trust it for data cleaning. Data cleaning is complicated and a LLM might clean away important signal that I want to study. At the same, LLMs are meant to train themselves. Even if LLMs don’t really do data cleaning now, in the future they may become smart enough that data scientists will want to use them for that purpose.

RC: It depends on how you define “data scientist”. I think there are two different roles with respect to this technology. One role is to analyze data, work with clients, draw conclusions. The flip side is when one is a developer, trying to fine tune these LLMs, build new applications, and so forth. I think you could call these people data scientists as well.

I think everybody is going to be a client of LLMs. We are all going to be centaurs, half human and half AI (a metaphor attributed to Kasparov by Westover (2023)). AI augmentation of human intelligence is coming, and partly here already (cf. Agrawal, Gans, and Goldfarb, 2023). If people want to keep up, they need to master these new tools. Like Costanza, I use LLMs to debug code. But as Tarak said, I think the more interesting problem is how to customize this technology.

Coming back to coding, code is just text. But it has a compiler, and so it means something specific, unlike the ambiguity of natural language. So I think there is an opportunity to leverage the interpreters, the compilers, with natural language models. Data scientists will have a special place in this domain.

CS: I don’t think I can comment on data scientists in general, but I can talk about what interests me. I care about the science of LLMs. These LLMs generate a lot of text, and I want to evaluate and measure features of that text. I am also interested in understanding the mechanisms of these models and their societal impact.

DB: Thank you all for you time and comments. This has been a wonderful and enlightening conversation.

References


©2024 David Banks, Costanza Bosone, Bob Carpenter, Tarak Shah, and Claudia Shi. This article is licensed under a Creative Commons Attribution (CC BY 4.0) International license, except where otherwise indicated with respect to particular material included in the article.