OpenUK - political heft is required to sort AI data conundrum

By George Lawton July 9, 2024

Dyslexia mode

SUMMARY: Generative AI models rely on a large pool of training data. Political heft is required to sort out questions about provenance, privacy, licensing, derivative works, and AI sweatshops.

The recent Open Source Initiative definition of open source AI deftly avoided issues around data transparency, leaving many issues to ongoing discussion. At a launch event for OpenUK's Phase Two of the Open Manifesto Report, experts engaged in a lively discussion about the future of AI training data.

OpenUK CEO Amanda Brock argues that a lot hinges on building a proper data foundation to bolster AI innovation:

--

> *ecosystem to encourage competition and innovation. We need better, more appropriate ways to use data across the ecosystem. We need different licenses, and we need a change in how developers and AI data publishers understand what licenses are and aren't.*

Lee Fulmer, Senior Advisor at McKinsey and Chairman of the Finance Advisory Board at OpenUK, surmises that people hugely underestimate the importance of data in driving AI, which has important implications for the future of competition, acceptance and adoption:

> *I think the biggest challenge that we have as a society at the moment is that most of the large language models out there are owned in commercial hands. I think they're trained on public data, and let's be honest, they're not peer-reviewed. I don't like to talk about quality because I think it's subjective, but the correctness of the data for the application is flawed. think the idea that what we need is within the open source community, a movement towards open data is hugely important if we want to drive innovation, if we want to drive acceptance and adoption, we need to have data sets that people can use and that they can trust and that, you know, maybe we do need a licensing model, to make sure that there's a validation there, rather than a cost associated. But we need some sort of mechanism to really drive things.*

Uncertain scraping landscape

Sonia Cooper, AGC IP, Microsoft, observes that the legal clarity in the UK is problematic regarding text and data mining. There was an effort to develop a code of practice that did not conclude. The UK also has not implemented the EU copyright directive. As a result, the UK has been left with a text and data mining copyright exception for noncommercial purposes but not for commercial purposes.

The UK is a signatory of the Trade-Related Aspects of Intellectual Property Rights (TRIPS) agreement, which protects ideas but not the underlying concepts. However, TRIPs has no provisions for AI training. Cooper explains:

> *know, do we have legal certainty there? Certainly not in the UK.*

Should commercial companies require a license to train models on copyrighted data? Cooper argues they should not:

> *I don't think licenses are needed for analyzing publicly available data. I don't think a license is needed for that. You might need a license agreement to access data, or you might need a license agreement for something that is a copyright infringement if you want to republish it if you want to do something, but to analyze it, there shouldn't be a requirement for it.*

But Fulmer argues this raises concerns about the potential of AI models trained on copyrighted data to be a kind of derivative work:

> *Is there not a potential issue there where the data that you do the training with creates a derivative work? Because doesn't that create a legal sort of quagmire? Because if you use a copyrighted piece of information that somebody else has, and you use that as a foundation to build something else, then you're actually building off their work without a license.*

Cooper believes it is only a copyright infringement when a model is used to generate a derivative work but not when a model outputs something based on the concepts from copyrighted work. Model developers are starting to introduce controls to prevent the recreation of original works. Beyond this, she believes that other protection mechanisms could be implemented that don't require a change in law through a voluntary code of practice. For example, by enabling people to make information available on the internet but don't want it used for AI training.

Better culture required

problems. She explains:

> *We need a better culture as AI scientists and practitioners around the data that we create or reuse that goes into these systems to train them. I mean, at the university, we can't afford to train systems anymore, but maybe this will come back in some way, or to fine-tune so to tailor some of these systems, and that's not incentivized at the moment at all. There's a focus on improving against a number of benchmarks that have very little to do with actual, real problems.*

The UK Government has invested in AI hubs for real-world data such as healthcare, chemistry, mathematics, electronics, and collective intelligence, which could help. It is important to consider practical issues to put these to use around access rights and quality.

Supply chain transparency

Simperl also believes we need to consider the labor practices used to train some of the larger AI models:

> *There's all sorts of labor rights in the data supply chain that we don't really talk about. We would not eat a sandwich that would be made of the ingredients or would not buy a t-shirt that would be made in the supply chain that we have at the moment in AI data. So that needs to improve as well.*

The process of training large models often requires feedback from humans. In the early days, this meant getting five people to agree on whether a cat was in a video. However, this data labeling process has gotten more complex and potentially traumatic for those tasked with determining whether an output is racist or toxic. It also costs more to manage this process across hundreds or thousands of people. Simperl says:

> *You need to recruit these people in a certain way that takes a data set to $250,000 at least. Now, who can afford to create that? And are we okay with a small part of*

Regulatory uncertainty in the UK may be slowing the development of better models in the UK. Getting regulations, policies, and processes wrong may be more costly in the long run. However, there are considerable questions about how open and transparent data policies will shape society for better or worse.

Neil Lawrence, DeepMind Professor of Machine Learning at Cambridge, explains:

> *This is still a technology that is reliant on data. There are still rights around the data, and there is still an issue around how large companies exploit digital markets to gain power. So those sort of regulatory agendas were somewhat disrupted here and replaced by these sort of arguments that have been made before that were being made, like ten years ago, about what might happen, but were processed through, 'What's the policy action?' And then the policy action either turns out to be impractical or misguided for reasons for the purpose and so you don't take that action. You double down on what you're doing. And we've just gone through that cycle relatively quickly across the last 18 months. Again, a series of policy actions proposed, no action taken.*

My take

Generative AI is the shiny thing that attracts the attention. However, the cool new stuff cannot be built without considerably more attention to the underlying data. This will be incredibly important as we look beyond better chatbots to applying AI to real-world problems in business, medicine, engineering, finance, and scientific discovery.

Activating this data will require making it more accessible via appropriate licenses, protecting privacy and security, and guarding against market abuse and the rights of creators. Navigating these issues through a minefield of contrary opinions will require persistence and openness.

Referring to the current state of the UK House of Lords inquiry into LLMs, Baroness Stowell of Beeston observes that the UK has allocated £400 million on AI safety but far less on

> *The Government's record on copyright is inadequate and deteriorating. We appreciate the technical and political complexities of the challenge. But we are not persuaded the Government is investing enough creativity, resources and senior political heft to address the problem.*

Hopefully, the newly-elected UK Government will muster the senior political heft the AI data conundrum deserves.

Image credit - Pixabay

Read more on: Machine intelligence and AI | Regulation | Open source | Government