AI Large Language Models Inquiry

Consultation by [Communications and Digital Committee, House of Lords](https://committees.parliament.uk/work/7827/large-language-models/)

**Written evidence submitted by the Trustworthy Autonomous Systems Hub.**

**Response Authors:**

**Dr. Joshua Krook:** Bridging Scholar in the Regulation of AI at the University of Antwerp, Visiting Academic at the University of Southampton. Krook has a PhD in Law from the University of Adelaide and previously worked in technology policy for the Australian government.

**Assistant Professor Jennifer Williams:** Assistant Professor in Electronics and Computer Science at the University of Southampton.

**Dr. Tina Seabrooke:** Assistant Professor in Psychology at the University of Southampton. Seabrooke’s research focuses on improving educational outcomes and reducing the impact of misinformation using principles from cognitive psychology.

**Dr. Eike Schneiders:** Postdoctoral researcher at the University of Nottingham. Schenider received his PhD in 2022 focusing on Human-Computer and Human-Robot Interaction from Aalborg University, Denmark.

**Assistant Professor Jan Blockx:** Assistant Professor, Faculty of Law, University of Antwerp. Blockx has research expertise in European economic law, covering areas such as competition law, state aid, the internal market, EU trade policy and technology regulation.

**Associate Professor Stuart E. Middleton:** Associate Professor at the University of Southampton. Middleton is a Turing Fellow and has over 25 years experience researching Natural Language Processing including Large Language Models.

**Professor Sarvapali D. Ramchurn:** Professor of AI and Director of the UKRI Trustworthy Autonomous Systems (TAS) Hub. Sarvapali has over 19 years of experience in developing AI solutions.

**Citation**

Joshua Krook, Stuart E. Middleton, Tina Seabrooke, Eike Schneiders, Jan Blockx, Jennifer Williams and Sarvapali D. Ramchurn, Trustworthy Autonomous Systems Hub. (2023) Response to: AI Large Language Models Inquiry, Communications and Digital Committee, House of Lords.

https://doi.org/10.5258/SOTON/P1126

**Contact details:**

[J.A.Krook@soton.ac.uk](mailto:J.A.Krook@soton.ac.uk) and [opsteam@tas.ac.uk](mailto:opsteam@tas.ac.uk)

**About the TAS Hub:**

The UKRI TAS Hub assembles a team from the Universities of Southampton, Nottingham and King’s College London. The Hub sits at the centre of the £33M [Trustworthy Autonomous Systems Programme](https://www.ukri.org/news/new-trustworthy-autonomous-systems-projects-launched/), funded by the UKRI Strategic Priorities Fund. The role of the TAS Hub is to coordinate and work with six research nodes to establish a collaborative platform for the UK to enable the development of socially beneficial autonomous systems that are both trustworthy in principle and trusted in practice by individuals, society and government. Read more about the TAS Hub [here](https://tas.ac.uk/).

**Question 1: How will large language models develop over the next three years?**

**a) Given the inherent uncertainty of forecasts in this area, what can be done to improve understanding of and confidence in future trajectories?**

Research and development around LLMs is now moving at pace as a direct result of open source releases of some commercial LLMs in 2023, such as Meta’s [LLaMa 2023], and this pace will likely accelerate in the short term as a result of quick turnaround innovation by smaller research and commercial organizations working with relatively small but still very powerful LLMs. Major LLM developments in the international research community are currently appearing monthly and it is very likely developments in the commercial space will happen at a similar pace in the near future as small agile commercial companies creatively offer LLM based services to UK society.

These rapid LLM developments are happening globally, across the UK, US, EU, China and many other countries. Open source proliferation of LLMs means the cost of entry for LLM research is low. Any country considering pausing LLM research will not stop the pace of innovation in other countries and will simply fall behind. The LLM genie is out of the bottle and it will not be possible to put it back in.

This rapid pace of LLM development means traditional mechanisms such as UK Parliament calls for evidence will likely be out of date before they are even published. A more dynamic approach is needed to track LLM developments and provide regular and useful guidance for UK Government policy makers and UK regulators.

**Recommendation:** We recommend that a dedicated UK working group on LLMs be setup for a period of 3 years, to advise every 6 months on (a) horizon scanning of LLM latest developments (b) recommendations for UK strategic direction priorities around LLMs and (c) maintain a strategic risk register around impact of LLMs on society. This working group should follow the structure of existing strategy working groups, such as the new [UKRI International Strategic Advisory Group 2023]. This new working group should be small, have a balanced representation from both UK academia and commercial LLM companies, and be chaired by an expert who does not have any commercial conflicts of interest around LLMs. The new UK government announced £100M Foundation Model Taskforce, Turing or UKRI would be good choices for organizations to host such a working group as they have relevant prior experience doing this, and membership should be recruited via an open call for participation with sufficient reach to recruit top expertise to the group. If hosted by Foundation Model Taskforce the working group should be set up as an arms length independent group as there might be a conflict of interest with Foundation Model Taskforce also funding LLM developments.

This dedicated UK working group would provide advice to multiple regulators, and would fall under the central function 'Monitoring, assessment and feedback' and 'Horizon scanning' defined in Box 3.1 of the AI white paper.

**Question 2:** **What are the greatest opportunities and risks over the next three years?**

**a) How should we think about risk in this context?**

Increasing attention is being placed on AI tools and services, largely due to the distribution and availability of large language models (LLMs) such as ChatGPT. These models pose numerous risks that have no simple answer. With many generative AI tools, the question remains as to what data the models have been trained on and, equally important, how to deal with intellectual property and copyright of the human generated material used to train the models. Furthermore, the data used in training the models comes with the inherent risk of producing discriminatory systems, as demonstrated with the Microsoft Tay chatbot [Vincent, 2016]. We, as a global society, have very limited understanding of the implications of a, for instance, sexist, racist, politically biased, or in other ways discriminatory tool might have on its users.

Additional risks, which already have been picked up by the media, are over-reliance on these tools by, for example, lawyers [Weiser and Schweber, 2023], medical professionals [Nastasi et al., 2023], and students [Cotton et al., 2023]. While LLMs have the potential to be a powerful tool when used in an appropriate manner, their misuse can contribute to de-skilling of the workforce, lowering of standards in education, as well as causing physical and mental harm due to reliance on e.g., financial, legal, or medical advice provided.

To address these issues, regulations need to be put into place to control the direction in which LLMs and their uses advance. However, the current speed at which these models are developing, and being made available to the world, makes it nearly impossible for regulations to keep pace. Nonetheless, while LLMs (and other generative AI) certainly come with risks, it would be folly not to acknowledge the opportunities they can provide. Implemented and deployed in a responsible and controlled manner, LLMs have the potential to improve education through, for instance, personalizing teaching material towards the needs of individuals. In a medical context, for example, LLMs could be used to provide medical students with case-studies in a low-stakes environment ([Abd-alrazaq et al., 2023](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10273039/)). In sum, LLMs have the potential to improve learning outcomes for individuals in many educational settings, all the while reducing teaching workloads. This could help alleviate some of the problems that schools and universities are facing due to the ever growing student body and in their recovery from the COVID-19 pandemic.

**Question 3: How adequately does the AI White Paper (alongside other Government policy) deal with large language models? Is a tailored regulatory approach needed?**

LLMs are not defined explicitly, but do fall within the AI white paper's approach to define characteristics which if met would represent AI.

The potential impact of LLMs on the creative industries has not been sufficiently represented in the AI white paper. The use of LLMs to create new ‘original’ versions of copyrighted work will challenge the nature of authorship and copyright. This might undermine the financial incentives for the UK’s creative industry workers in areas such as visual arts, literature and gaming, areas which contribute significant value to the UK economy.

**Recommendation:** We recommend in Box 1.2 an explicit example for LLM impact to the creative industries be added around LLM-based plagiarism of copyrighted material, to be added to the ‘Fairness’ principle rationale.

Examples in ‘Case-study 3.9: Life cycle accountability for large language models’ are already outdated by recent innovations around low compute LLMs. The idea of ‘monitoring the amount of compute used to train’ LLMs as a proxy for potential risks around LLM impact is deeply flawed. Small low compute LLMs are already achieving performance within 80% of larger high compute LLMs for many tasks [Vicuna Team 2023]. LLM compute power is an easy to measure but too crude a metric to insufficiently characterize the impact a LLM will have and its associated risks.

**Recommendation:** We recommend UK regulators mandate LLM service providers provide legally binding risk assessments, following a similar regulation approach to the UK's Online Safety Bill. Regulators can then audit LLM service providers with regards to LLM risks and hold them accountable as needed.

**a) What are the implications of open-source models proliferating?**

Open-source models of LLMs are currently in a state of flux due to a number of confounding factors, including the entrant of major tech companies into the space, and the regulatory push for AI “licenses” for LLMs. The biggest entry into the field is Meta, which has released its own open source model, LLaMA LLM. (Meta AI, 2023). CEO Mark Zuckerberg stated upon release that "Open source drives innovation because it enables many more developers to build with new technology" (Zuckerberg, 2023). OpenAI, creator of ChatGPT, is also purportedly preparing to release an open source LLM, however, reports state that it is unlikely to be competitive or superior to their commercial model, GPT4 (Reuters, 2023). Google has previously released more primitive LLMs with Flamingo, FLAN and Palm, there is no indication that they will release an open source model competitive with their commercial product BARD either. Major companies entering the open source space may concentrate the open source community, moving it away from smaller competitor communities, such as Hugging Face [Hugging Face, 2023].

The existing proliferation of open source model LLMs has, at times, created a sense of urgency in the tech sector. A leaked document from an engineer inside Google in 2023,for example, framed open source models as a direct threat to their own LLM development of BARD. The Google engineer wrote, in part, “We have no moat… while our models still hold a slight edge in terms of quality, the gap is closing astonishingly quickly. Open-source models are faster, more customisable, more private, and pound-for-pound more capable" (Patel and Ahmad, 2023) . At the same time, OpenAI CEO Sam Altman has been openly pushing for greater regulation of LLMs, and specifically, a licensing scheme for models more powerful than GPT4. If governments were to act on this licensing rule, then this may place a damper on open source LLM development for smaller companies and more informal online communities that would struggle to meet reporting or public disclosure requirements (due to cost, organization, and/or data processing). As summarized in the recent and highly cited open-access 176-billion parameter LLM called *BLOOM* designed for multilingual natural language processing tasks like machine translation [2] “Public narratives about the technology that are driven by industry actors can lead to inflated expectations about its suitability for use (Brennen, 2018; Brennen et al., 2022), leading to misaligned research and policy priorities (Raji et al., 2022) and potentially dire consequences in e.g. medical applications.”

**Recommendation:** We recommend that UK regulators take a cautious approach to open source LLMs. However, we do not support an outright ban on open source models, nor a licensing scheme that is designed specifically to make open source models, and innovative technology UK startups, incapable of competing with big tech firms.

**Question 4: Do the UK’s regulators have sufficient expertise and resources to respond to large language models? If not, what should be done to address this?**

On the 22nd May 2023, the Trustworthy Autonomous Systems Hub held a Regulators Workshop, inviting regulators from a number of different industries to discuss the challenge of regulating AI. A consistent theme of the workshop was that regulators were under-staffed when it came to technical experts on the topic and that they had difficulties recruiting technical experts with a background in AI and computer science.

**Recommendation:** We recommend that the government encourage regulators to work closely with the Foundation Models Task Force (FMTF) to help regulators test their rules on real-world use cases. We likewise encourage the recruitment of technical experts into regulatory bodies, upskilling regulatory staff and/or encourage further dialogues between regulators and academic experts in the fields of AI and computer science. This upskilling is necessary to address the need for a tailored regulatory approach for the many nuances of LLMs in different sectors.

**Question 6: How does the UK’s approach compare with that of other jurisdictions, notably the EU, US and China?**

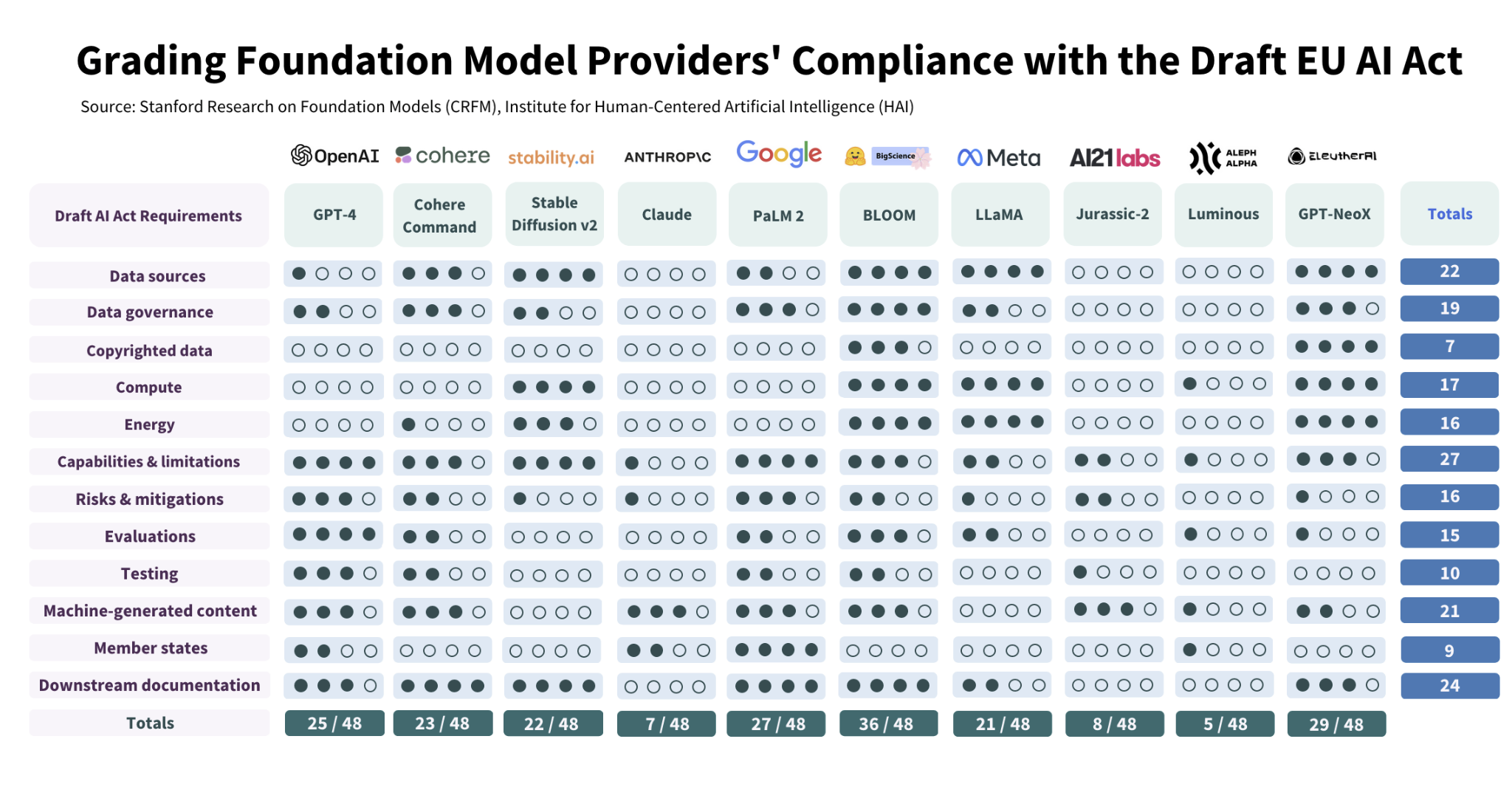
**EU:**

The European Union (EU) has adopted a centralized, top-down view of AI regulation that attempts to harmonize AI rules across the EU. A range of recent regulatory initiatives will directly impact the usage and provision of LLMs in Europe, including the Digital Services Act (DSA, 2022), Digital Markets Act (DMA, 2022) and, most importantly, the proposed AI Act (expected to pass in late 2023 or 2024). Taken together, these regulations aim to create a fair, more transparent and accountable digital landscape, regulating the digital marketplace and provision of digital services (DSA), providing for fair competition between large and small firms (DMA), and providing rules for transparency and accountability for generative AI, including LLMs (proposed AI Act).

Most important for the regulation of LLMs is the proposed AI Act. At a basic level, it requires all AI systems that interact with natural persons to be “designed and developed in such a way” so that users know that “they are interacting with an AI system” (Article 52, proposed AI Act). This ‘show or tell’ provision, will mean that users will have to either be shown in some manner that an LLMs content was generated by AI by nature of the content, or explicitly told that the content is AI generated, with the usage of a watermark or a disclosure statement.

The latest amendments proposed by the European Parliament would impose further transparency requirements on LLMs, as a subset of foundation models, regarding input data, use of copyright materials, and restrictions on the generation of illegal content (Article 28b, proposed AI Act). Prior to being put on the market, the LLM would also need to be registered in a dedicated EU database. If they are used in high-risk applications (such as certain types of biometric identification, law enforcement, employment, education, etc.), LLMs face further transparency obligations regarding stating their intended purpose, approximate accuracy, risks of use, limitations, and where appropriate, input data (Article 13, proposed AI Act).

Researchers at Stanford (Bommasani et al, 2023) have [studied](https://crfm.stanford.edu/2023/06/15/eu-ai-act.html) the ten most popular LLMs, including GPT and Bard, and found that *none of them* currently comply with the transparency rules provided in the latest version of the proposed AI Act (See diagram below). The models that came closest to compliance were open source LLMs, referring to models built by the community using the principles of transparency at their core.



Given the above diagram and analysis, it is clear that LLM providers face a new and significant regulatory hurdle in Europe to match the requirements of AI transparency. Some of the transparency obligations (such as the provision of input data), contradict market imperatives of secrecy and confidentiality. Few companies wish to share the data that they train their models on.

While several members of the European Parliament expressed a wish for centralized enforcement of the proposed AI Act, the current text of the proposal does not provide for this, mainly due to a lack of (financial) resources at the EU level. Much of the enforcement will therefore need to take place at a national level, although coordination between national authorities is meant to take place through a central EU AI Office. The role and powers of this central body will probably be intensely debated in the coming months. (Note that the DMA and DSA do provide for centralized enforcement at an EU level, at least as far as the largest tech companies are concerned.)

Since it may well be the first comprehensive AI legislation, the EU AI Act may act as an example for other jurisdictions. However, certain factors suggest that it may not have the same exemplary effect as the General Data Protection Regulation that was adopted in 2016 (which led to copy-cat legislation in many jurisdictions and the dubbing of a “Brussels Effect” (Bradford 2020)). The fact that multiple jurisdictions, including the US, UK and China, are currently creating their own regulatory provisions, means that the “gap in regulation” that occurred before the GDPR (and the general right to privacy online) is in any event less significant now.

**China:**

China is taking a more robust, restrictive approach to the regulation of LLMs when compared to the EU, UK and US. In particular, China is placing a strong emphasis on state oversight of AI, values alignment with state principles, political ideology and ‘socialist’ ideals (loosely termed), truthfulness and accuracy of generated content, and taking a proactive approach to determining the kind of future they would like to see. A focus on the impact of technology is at the core of their response. At the same time, state controls and a focus on moving towards “accuracy” in LLM content could potentially hamper innovation.

In April 2023, the *Cyberspace Administration of China* created a draft set of regulations on generative AI (including LLMs), that aimed to solidify these state principles. The draft regulations mandated that companies abide by CCP censorship rules. This meant that all LLMs would have to abide by the country’s strict publishing requirements on content relating to China’s leaders, controversial history or topics in news and current affairs. [Che, NYT, 2023] LLMs would also need to reflect “socialist core values,” and avoid undermining state power or national unity. These topics were defined broadly, but included at a minimum, restrictions on calls for violence, civil disobedience or other disruptive activities as defined by the CCP. The draft AI regulations required that training data and model outputs be “true and accurate,” which posed a problem for LLMs that generated “hallucinations” (untrue or unverified claims). A report by the Carnegie Endowment for International Peace called this “true and accuracy” requirement an “insurmountable hurdle for AI chatbots to clear” [Sheehan, 2023]. LLMs, by nature of being predictive language tools, typically suffer from the problem of “hallucinations,” whereby predictions produce unverifiable or false information, such as fake links, book titles or case studies. The obligation on LLMs in China to provide “true and accurate” information would have placed a strong, if not impossible, regulatory burden on technology companies, to solve a currently unresolved technological problem. However, these draft regulations were not to last.

In August 2023, the *Cyberspace Administration of China* implemented an *updated* set of rules for generative AI, with significant changes from the above draft provisions. This final version significantly weakened the provisions relating to generative AI. Article 2, which previously covered all generative AI (including for R&D purposes), now only covers *public-facing* generative AI. The original draft would have limited research and development, whereas the new regulation may not. These changes re-orientate the regulation towards a more pro-innovation focus, with looser rules and standards allowing for the potential for greater research and development of LLMs in China.

The original draft demanded “truth and accuracy” in generated content. The updated regulation instead asks companies to “adopt effective measures to elevate the quality of training data, and to strengthen the truth, accuracy, objectivity and diversity in training data”. Provisions were also added to accommodate generative ai “by service type,” meaning that science fiction images need not be “true and accurate”. The “truth” requirement would have been almost impossible to implement, so this new focus on transparency is likely a more workable situation for tech companies to follow. At the same time, these regulations are still provisional in nature, meaning they could be subject to further changes.

China’s other rules on labeling generated content match similar proposed requirements in the EU and UK. China’s *Deep Synthesis* regulation requires the labeling of synthetically generated content, including LLMs, to make it clear that the information was artificially generated. Users will have to be informed that the content has been generated by AI, and this may typically occur in the form of a watermark or public disclosure statement. This is similar to the proposed watermarking of content in the EU’s AI Act (s. 52) and the UK’s AI Regulation White Paper.

All three of the major Chinese regulations on AI “require developers to make a filing to China’s algorithm registry, a newly built government repository that gathers information on how algorithms are trained, as well as requiring them to pass a security self-assessment" [Sheehan, 2023]. In effect, this amounts to a security screening of LLMs and will lead to very close oversight of LLM providers in China.

**References:**

Anthony J. Nastasi, Katherine R. Courtright, Scott D. Halpern, Gary E. Weissman. Does ChatGPT Provide Appropriate and Equitable Medical Advice?: A Vignette-Based, Clinical Evaluation Across Care Contexts, medRxiv 2023, doi: https://doi.org/10.1101/2023.02.25.23286451

Anu Bradford, The Brussels Effect (OUP 2020).

Benjamin Weiser and Nate Schweber, 2023, The ChatGPT Lawyer Explains Himself, https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html

Chang Che, China Says Chatbots Must Toe the Party Line (The New York Times, 24 April 2023).

Debby R. E. Cotton, Peter A. Cotton & J. Reuben Shipway (2023): Chatting and cheating: Ensuring academic integrity in the era of ChatGPT, Innovations in Education and Teaching International, DOI: 10.1080/14703297.2023.2190148

European Commission, European Centre for Algorithmic Transparency <https://algorithmic-transparency.ec.europa.eu/index\_en>.

Future of Life Institute, Open Letter: Pause Giant AI Experiments https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Hugging Face, 2023, https://huggingface.co/.

J. Scott Brennen. An industry-led debate: how UK media cover artificial intelligence, 2018.

James Vincent, 2016, Twitter taught Microsoft’s AI chatbot to be a racist asshole in less than a day https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

Josephine Wolff and Nicole Atallah, 2021, Early GDPR Penalties: Analysis of Implementation and Fines Through May 2020, Journal of Information Policy, 11: 63–103.

Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., ... & Al-Shaibani, M. S. (2022). BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.

Matt Sheehan, China’s AI Regulations and How They Get Made (Carnegie Endowment for International Peace, 2023).

Mark Zuckerberg, 2023, cited in Meta Challenges OpenAI and Google with Open-Source AI, Industry Week, https://www.industryweek.com/technology-and-iiot/emerging-technologies/article/21269825/meta-challenges-openai-and-google-with-opensource-ai

Meta AI, 2023, Introducing LLaMA: A foundational, 65-billion-parameter large language model, https://ai.meta.com/blog/large-language-model-llama-meta-ai/

Dylan Patel and Afzal Ahmad, 2023, Google "We Have No Moat, And Neither Does OpenAI", *Semi-Analysis*.

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*.

Reuters, 2023, OpenAI readies new open-source AI model, The Information reports, https://www.reuters.com/technology/openai-readies-new-open-source-ai-model-information-2023-05-15/

Rishi Bommasani, Kevin Klyman, Daniel Zhang and Percy Liang, 2023, Do Foundation Model Providers Comply with the Draft EU AI Act?

Scott Brennen, Philip N Howard, and Rasmus K Nielsen. What to expect when you’re expecting robots: Futures, expectations, and pseudo-artificial general intelligence in UK news. Journalism, 23(1):22–38, 2022.

UKRI International Strategic Advisory Group, UKRI, 2023 https://www.ukri.org/who-we-are/work-for-us/join-an-advisory-committee-panel-or-network/ukri-international-strategic-advisory-group/

LLaMa, Meta, 2023, https://ai.meta.com/blog/large-language-model-llama-meta-ai/

Vicuna Team, Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%\* ChatGPT Quality, 2023, https://lmsys.org/blog/2023-03-30-vicuna/