Dr Xuechen Chen, Dr Xinchuchu Gao and Dr Lingpeng Kong—written evidence (LLM0031)

House of Lords Communications and Digital Select Committee inquiry: Large language models

This written evidence is a collaborative submission by an interdisciplinary group of researchers with expertise in Natural Language Processing and Machine Learning, International Relations and International Political Economy: Dr Xuechen Chen (Northeastern University London), Dr Xinchuchu Gao (University of Lincoln), and Dr Lingpeng Kong (University of Hong Kong). This response is given in a personal capacity, reflecting our personal opinions as researchers.

Dr Xuechen Chen is an Assistant Professor in Politics and International Relations and Head of Digital Governance Research Cluster at Northeastern University London. Her current research focuses on emerging powers in regional and global cyberspace governance, and norm and policy diffusion in digital governance.

Dr Xinchuchu Gao is a Lecturer in International Relations at the University of Lincoln. Her research interests lie at the intersections between international relations, international political economy and European studies. She is specifically interested in the twin green & digital transitions of the EU and global cyber governance.

Dr Lingpeng Kong is an Assistant Professor in the Department of Computer Science at the University of Hong Kong. His research interests lie at the intersection of natural language processing and machine learning, with a focus on representation learning, structured prediction, and generative models.

Summary:

- We can expect significant improvements and breakthroughs in large language models (LLMs) over the next three years, which will enable LLMs to process longer sequences, maintain context in extended conversations, and offer greater degree of personalization. Breakthroughs in statefulness and strong reasoning models for expert domain knowledge can also be expected. These advancements will allow LLMs to address the existing problem of hallucination and will significantly enhance LLMs' capability of handling complex tasks across a wider range of expert domains like medicine, law, and finance.

- LLMs will generate growing opportunities across various domains. It is likely that LLMs will increasingly serve as central controllers for various tools, with the potential to act as general problem solvers. LLM-powered autonomous agent systems are likely to emerge. The development of Internet of Things (IoT) ecosystems will be greatly facilitated by LLMs. LLM-driven hyper-personalization will transform customer interactions. LLMs will also find applications in scientific fields, making breakthroughs in mathematics, physics, chemistry, and biology.

- In terms of risks, the increasing difficulty in distinguishing between machine-generated and human-generated content poses a significant threat, leading to a surge in electronic fraudulent activities using intelligent strategies. LLMs can inadvertently introduce bias into responses, perpetuating societal biases and causing unfairness or discrimination. The lack of transparency complicates efforts to address bias, and the reinforcement learning process tends to produce homogeneous content, lacking diversity. LLMs also raise concerns about the leakage of personal or confidential information and potential intellectual property rights infringements. Additionally, it's challenging to prevent bad actors from misusing LLMs, especially in multi-round dialogues and role-playing, as smarter LLMs are more susceptible to misuse.

- The UK's AI White Paper has not sufficiently and adequately dealt with the rapid developments of LLMs due to a lack of clear definitions of key terms concerning AI systems as well as insufficient explanation regarding what roles and responsibilities that sector-specific regulators and the government should take. This may lead to the possibility of inconsistent enforcement of the guiding principles outlined in the AI White Paper.

- In comparison to other major international actors such as the EU and China, the UK has already lagged behind in developing effective regulatory measures to deal with LLMs and generative AI models. The UK should reflect on how to develop innovative approaches that bring together regulatory elements from both "horizontal" and "vertical" approaches in AI governance, drawing lessons from other jurisdictions.

- Regulatory alignment in AI governance at a global level is unlikely to happen in the near future. Regulatory divergence may continue to deepen because different international actors, such as the EU, US, China, and the UK, are developing contrasting approaches to regulating AI technologies. While different countries may increasingly share a conceptual alignment in their efforts to develop trustworthy AI, the competitive approaches they embrace, their ambitions to assume leadership roles in AI governance, and the escalating geopolitical rivalry in technology domains make regulatory convergence highly unlikely.

<u>Questions addressed</u>:

1. How will large language models develop over the next three years?

1.1. There will be significant improvements in terms of large language models (LLMs)' capacity of processing long sequences (e.g., text with more than tens of thousands of words), allowing users to generate longer pieces of texts and to handle longer conversations without losing context. Such improvement may potentially derive from two sources: firstly, the progress in transformer architecture will provide better resolution for long sequences without sacrificing accuracy. Secondly, advances in training objectives and training data will encourage large language models to capture long-distance knowledge, which will significantly improve the models' ability to understand and generate long documents.

1.2. There will be significant improvements in terms of statefulness (i.e., memorization of unlimited history) as well as the level of personalization. Currently, achieving personalization in large language models remains infeasible because memorizing all history requires individually fine-tuning the model. However, breakthroughs in these technologies are likely to occur within three years, which will allow users to obtain highly personalized services and tailored customer experiences.

1.3. Strong reasoning models for expert domain knowledge are expected to emerge within the next three years. The hallucinations problem, which refers to the phenomenon that the model speaks false knowledge as if it is accurate, is expected to be significantly mitigated when developing strong reasoning models. The progress of LLMs in general language understanding makes it possible to inject the knowledge required for specific fields (e.g., medical, legal, and financial domains) to complete complex tasks (e.g., contract compliance checks, financial data queries, and data mining).

2. What are the greatest opportunities and risks over the next three years?
a) How should we think about risk in this context?

2.1. In terms of opportunities, one can expect that LLMs will become the core controller for other tools. LLM-powered autonomous agent systems are likely to emerge. It is noteworthy that several proofs-of-concept such as AutoGPT and GPT-Engineer, demonstrate that LLMs have the potential to act as general problem solvers. In fact, in many application fields, automated programs such as industrial automation have already been widely deployed. Whilst these existing applications rely primarily on a single AI system for a specific task, the advancement of LLMs and the development of LLM-powered autonomous agents will enable the models to serve as an overarching control center, providing high-level guidance for a series of heterogeneous intelligent systems to handle complex tasks.

2.2. LLMs are likely to play a significant role in developing Internet of Things (IoT) ecosystems. As our surroundings and devices increasingly connect to the IoT, LLMs will enable us to interact with and manage these devices through natural language interfaces, powered by their impressive natural language processing capabilities. This integration will foster a heightened awareness of device status and context within the broader ecosystem, leading to smarter and more efficient operations.

2.3. Advancements in LLMs will play an important role in reshaping marketing and sales in a wide range of business sectors. For example, hyper-personalization strategies are likely to surge across a wide range of business sectors, resulting in a paradigm shift in ways in which companies and business organizations interact with customers.

2.4. It is also worth noting that LLMs will be increasingly applied in a wide range of scientific fields such as mathematics, physics, chemistry, and biology. These new AI tools and strengthened capabilities of LLMs will

make significant breakthroughs in the domains of natural science (e.g., proving mathematical theorems, developing new medicines and materials). For example, AlphaFold is an artificial intelligence program developed by DeepMind, a subsidiary of Alphabet, which performs predictions of protein structure.

2.5. Natural language interfaces will be more widely used to handle more complex tasks and break down language barriers. One can expect the emergence of LLM-powered "perfect translators" that are capable of multilingual translation.

2.6. In terms of risks, it is worth noting that it will be increasingly difficult to distinguish between machine-generated content and human-generated content. As a result, a significant surge in electronic fraudulent activities is likely to happen. By adopting LLMs, these fraudulent activities will use more intelligent and sophisticated strategies (e.g., creating highly personalized and convincing messages which are tailored to a specific victim) to increase the success rate of fraud.

2.7. Another risk is that LLMs may produce responses that are influenced by bias. Bias refers to a tendency to favor a particular perspective, belief, or opinion over others. This bias can impact how people perceive, understand, and assess information or situations, often resulting in unfairness or discrimination. Biases can originate from various sources, including biased training data, societal biases reflected in user interactions, or even biases introduced during the fine-tuning process. These biases can take the form of gender, racial, or ideological biases, leading to unequal representation or unjust treatment in the information provided. Algorithmic bias is of particular concern because it can reinforce and perpetuate existing societal biases. Users who heavily rely on LLMs and ChatGPT may unknowingly absorb these biases, distorting their understanding of various subjects. The lack of transparency and interpretability in these models further complicates efforts to identify and address bias.

2.8. In addition, given that the training basis of large language models is the maximum likelihood estimation of observed data, there is a lack of diversity in the reinforcement learning human feedback (RLHF) process. This will inevitably lead to the tendency that LLMs produce content that is increasingly homogeneous and less representative of the diversity of the society. With the increasing amount of machine-generated content in the world and the increasing interaction between people and generated models, this problem may become more serious in the near future.

2.9. Another major risk of LLMs is concerned with the potential leakage of personal or confidential information. During the training process of large models, due to the irregular collection of data, private or confidential information is easily and explicitly leaked during the generation phase, or potentially leaked. For example, confidential know-how may be processed into new inferences and leaked. It is difficult to supervise these leaks with the existing techniques. Besides, LLMs may expose works that are protected by intellectual property rights to the risk of infringement and

therefore generate uncertainty regarding the ownership of the content produced by LLMs.

2.10. Another significant risk is that from a technical perspective, it is difficult to prevent bad actors from using LLMs to undertake cybercrimes. In the current framework, bad actors and inappropriate usage of generative AI are often judged in RLHF, based on which LLMs can refuse to answer the queries. However, in reality, bad actors can use various methods to achieve the goal of deceiving and circumventing the security settings of models, which is more difficult to control in multi-round dialogue and role-playing. In fact, more intelligent LLMs do not necessarily imply higher levels of safety. Instead, smarter LLMs are more vulnerable to bad actors' misuse of technologies.

3. How adequately does the AI White Paper (alongside other Government policy) deal with large language models? Is a tailored regulatory approach needed?
a) What are the implications of open-source models proliferating?

3.1. Although the life cycle accountability for LLMs was discussed as a case study in the document, the AI White Paper did not sufficiently and adequately deal with the implications of the surge of LLMs. Firstly, there is a lack of clear definitions on key terms such as foundation models, large language models, and generative models. When key terms lack clear definitions, it becomes challenging to interpret and apply regulations consistently. Regulators may find it difficult to enforce rules effectively as different parties may interpret the regulations differently. Moreover, the uncertainty stemming from undefined terms can hurdle investment and innovations within industries subject to these regulations. Businesses might be reluctant to pursue business expansions and investments in new technologies if they are unsure about the regulatory landscape they will face.

3.2. Secondly, although the AI White Paper set out principles guiding future development of AI models and tools, these principles require interpretation and translation into actions by individual sectoral regulators. This raises significant questions about who ought to take the responsibility for implementing the overarching principles of AI governance within a particular sector, and whether they possess the capabilities and resources required to deliver these principles. For instance, in some domains, there are no domain-specific regulators. This means that it is not clear who is responsible for implementing the general principles introduced by the AI White Paper. Moreover, some regulators might have limited access to AI expertise and mechanisms for regulatory cooperation, potentially leading to the possibility of inconsistent enforcement of the guiding principles outlined in the AI White Paper.

3.3. Thirdly, according to the AI White Paper, there will be a duty for sector-specific authorities such as the ICO, CMA, FCA and Health and Safety Executive and the Human Rights Commission to comply with the AI regulations by issuing both individual and joint guidance where AI crosses

multiple sectors. Nevertheless, the AI White Paper provides very little details regarding how this will work in practice. Therefore, it is likely that businesses receive conflicting advice from multiple regulators. Different regulators need to figure out how to interact with each other to ensure regulatory consistency and enhance regulatory coordination.

3.4. Foruthly, it remains unclear what roles the UK government will play in guiding collaboration among sector-specific regulatory institutions. We observe the government's intention to encourage the regulators themselves to produce guidance and advice. Businesses therefore will seek assurance that adhering to the guidance and advice issued by regulators will effectively minimize the risk of AI regulation violations, particularly for small to medium sized enterprises (SMEs) lacking in-house legal expertise. To bolster businesses' confidence, providing clarification on respective roles of governments and regulators in issuing guidance and advice would be beneficial.

3.5. Fifthly, although the regulatory framework set out in the AI White Paper is deliberately designed to be flexible, the pace of AI technological developments still raises the question about whether a faster response is needed. As AI technologies develop rapidly, the government and regulators need to continually reassess and update the regulatory framework to address emerging challenges.

3.6. Sixthly, there is a significant concern regarding the potential for conflicts between the principles outlined in the AI White Paper and those present in other existing regulatory frameworks. For instance, given that a substantial portion of data processed by AI systems is personal data, there is a natural overlap between the AI regulation principles and the data protection principles. One example is that the AI White Paper's fairness principle is much like data protection's fairness principle. Therefore, it is important for the AI White Paper to frame AI principles in a way that is compatible with their meanings under data protection law. Maintaining compatibility between the principles will minimize unnecessary burden for businesses and advance technological innovation.

3.7. Regarding the implications of open-source models proliferating, one can expect that open-source models can contribute to boosting the effort in understanding and developing LLMs technologies, including but not limited to more efficient training and inference, stronger reasoning ability, interpretability, and safety. It can also help dismantle the monopoly of big tech companies in this sector. Furthermore, it is expected that opportunities for new applications of open-source models would emerge. Of particular interest is the llama2 model introduced by Meta, which utilizes an open-source license that is friendly for commercial use. This makes it incredibly convenient to directly apply open-source models and to develop them further for commercial purposes.

4. *How does the UK's approach compare with that of other jurisdictions, notably the EU, US and China?*
   a) *To what extent does wider strategic international competition affect the way large language models should be regulated?*
   b) *What is the likelihood of regulatory divergence? What would be its consequences?*

4.1. Firstly, in comparison to other major international actors such as the EU and China, the UK lags behind in developing effective regulatory measures to deal with LLMs and generative AI models. It is noteworthy that on 14 June 2023, the European Parliament adopted a compromise text for the EU's AI act, paving the way for the final step in the legislative process[1]. In this compromise text, the European Parliament has proposed a series of amendments which take into consideration the recent development in LLMs and generative AI in the hope of mitigating the risks. Specifically, the compromise text provides, for the first time, clear definitions for both "foundation" and "generative" AI models, which helps clarify the boundaries and scope of the obligations to which different AI systems are subject. The amended text also details new requirements on providers of foundation AI models (which include generative AI models as a sub-category), including obligations to show "compliance-by-design" throughout the development of the AI systems to achieve "adequate performance, predictability, interpretability, corrigibility, safety and cybersecurity" as well as obligations to register the foundation model in an EU-wide database. Generative AI systems will be subject to additional requirements concerning user transparency, more extensive testing, as well as the documentation and publication of summaries of the training data.

In a similar vein, the Interim Administrative Measures of Generative Artificial Intelligence Services (Generative AI Measures)[2] were issued by the Cyberspace Administration of China, along with six other authorities, on 13 July 2023 and entered into effect on 15 August 2023. Similar to the EU's text, China's Generative AI Measures seek to provide definitions on the key terms such as "generative AI technologies", "generative AI services", and "generative AI service provider" in an explicit manner. The measures also set out a diverse array of obligations on generative AI services providers, which include aspects such as training data requirements, tagging and labelling standards, data protection requirements, and safeguarding user rights.

Compared to the EU's and China's documents, the UK's AI White Paper did not provide a rigid definition on AI systems. The absence of clear definitions of the key terms may generate a higher degree of legal uncertainty, as well as confusion about the scope, objects, and subjects of regulation both for regulators and industrial stakeholders. Although the UK's AI White Paper emphasized that "LLMs will be a core focus of our monitoring and risk assessment functions" and that life cycle accountability of LLMs will be a priority area for ongoing policy

1 https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf
2 http://www.cac.gov.cn/2023-04/11/c_1682854275475410.htm

development, a more comprehensive roadmap with concrete steps and methodologies to achieve these goals need to be further developed and specified.

4.2. A comparison between the UK's approach with that of other jurisdictions (i.e. EU, US, and China) shows that the UK and the US seem to share more common ground in terms of AI regulation, whereas a more substantial regulatory divergence can be observed between the UK and the EU (and between the UK and China). Notably, the US approach to AI governance and regulation is sectorally specific, highly dispersed, and is distributed across various federal agencies[3]. Despite the existence of several guiding documents from the White House on AI, there has been a lack of a consistent and coordinated federal approach to AI governance and regulation in the US. The Blueprint for an AI Bill of Rights (AIBoR)[4] published by the White House Office of Science and Technology Policy in October 2022 set out a non-binding roadmap for the responsible use of AI based on five core principles[5], providing a comprehensive explanation of the negative impacts of AI on economic and civil rights. It outlines five key principles aimed at reducing these negative effects and includes a list of recommended actions for federal agencies to take. The AIBoR emphasizes a sector-specific approach to governing AI, where policies are tailored to specific sectors like healthcare, labor, and education. Such an approach relies heavily on the actions of these federal agencies to implement the guidelines instead of centralized and coordinated action. Alignment between the UK's and the US's approach can be identified in the sense that the UK's AI White Paper also adopts a non-binding, and principle-based approach which attaches great importance to similar aspects such as safety, explain ability, and fairness. In addition, instead of pursuing a centralized approach, the UK's AI White Paper seeks to empower existing regulators and envisages that regulators will issue their own guidance interpreting the key principles in their policy domains. Both the UK and the US are currently adopting a "light-touch" approach based on a relatively fluid regulatory framework, allowing the industry to thrive and innovate at will.

4.3. The UK's approach differs significantly from that of the EU. Specifically, the EU's AI Act is better conceptualized as a "horizontal" regulation, which signifies that it sets out rules for AI across all sectors and applications. Besides, the EU's AI Act proposes a risk-based approach, detailing four levels of risk for AI systems. Different rules and obligations will apply depending on the levels of risk an AI system poses to fundamental rights. Compared to the UK's principle-based approach, the EU AI act represents a more rigid and prescriptive legislative framework which imposes

---

3 Engler A. (2023). The EU and U.S. diverge on AI regulation: A transatlantic comparison and steps to alignment. Available at: https://www.brookings.edu/articles/the-eu-and-us-diverge-on-ai-regulation-a-transatlantic-comparison-and-steps-to-alignment/

4 The White House, Blueprint for an AI Bill of Rights (Washington, D.C., 2022) https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

5 The five principles emphasized in AIBoR include: (1) safe and effective systems; (2) Algorithmic discrimination protections; (3) data privacy; (4) notice and explanation; (5) human alternatives, consideration, and fallback.

obligations at all stages of the lifecycle of AI systems. In addition, the EU approach is built upon a much more coordinated network of both new and existing regulators which include a centralized European AI Board that would oversee the implementation of the regulation across the EU, and national authorities for AI in each EU member state. Furthermore, whereas the EU approach proposes up to 40,000.00 EUR or 7% turnover for violations, the UK has not introduced any penalties at the current stage. Whilst some argue that rigid and onerous regulations may hamper AI innovation, the EU's approach may be in a better position to ensure greater degree of legal certainty which thereby may enhance consumer trust in AI systems. It is also noteworthy that the EU AI act should not be considered as the only key legislation in AI governance. Rather, the EU's AI governance regime is multifaceted and builds upon existing legislations (e.g. GDPR) as well as new legislations (e.g. Digital Services Act and Digital Market Act).

4.4. Different from that of the EU, China's approach in AI regulation is both vertical and iterative.[6] Beijing's approach leans towards a vertical approach in the sense that regulators tend to take a bespoke strategy, setting out different regulations and legislations to deal with different applications or types of AI systems, as evidenced in the Generative AI Measures which is the latest addition to AI governance regulatory tools in China after the adoptions of the Algorithm Provisions[7] in 2021 and the Deep Synthesis Provisions[8] in 2022. Apart from being vertical, the process of developing new regulations is iterative: when the regulators identify a regulation as flawed or insufficient, they will introduce a new one to address any gaps or broaden its coverage, as manifested in the case of the Generative AI Measures which extended the measures for Deep Synthesis Provision. Whilst such an iterative process may generate uncertainty in terms of compliance, Chinese regulators consider it as a reasonable trade-off for effectively overseeing a rapidly evolving technology landscape. In addition, some horizontal components can also be observed in China's regulations. Some regulatory tools, such as the algorithm registry, are deployed across different vertical regulations. Regulators from the UK may draw lessons from the case of China to explore how to develop more effective approaches in AI governance that incorporate regulatory tools characterized by both vertical and horizontal elements. We also believe that a more nuanced understanding of China's approach in AI governance is much needed in the context of the UK, as the majority of the existing analyses fail to capture the complexity of China's AI governance landscape which involve multiple stakeholders at different levels in the agenda-setting and decision-making processes to regulate AI.

4.5. Regulatory alignment in AI governance at a global level is unlikely to

6 Sheehan, M. (2023). China's AI Regulations and How They Get Made. Available at: https://carnegieendowment.org/2023/07/10/china-s-ai-regulations-and-how-they-get-made-pub-90117. Zhu, S and Ma, G, The Chinese Path to Generative Ai Governance. Available at SSRN: https://ssrn.com/abstract=4551316 or http://dx.doi.org/10.2139/ssrn.4551316

7 http://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm

8 http://www.cac.gov.cn/2022-12/11/c_1672221949318230.htm

happen in the near future. Regulatory divergence may continue to deepen because different international actors, such as the EU, US, China, and the UK, are developing contrasting approaches to regulating AI technologies. The EU follows a more sector-specific and horizontal strategy, while the US opts for a decentralized approach featuring federal guidance with local adaptations. China, in contrast, tends to be vertical, as regulators often adopt a tailored approach. Also, China's approach prioritizes flexibility and adaptability, even sometimes at the cost of certainty.

Pursuing contrasting approaches, these actors all proactively compete for more normative and regulatory power in the field of AI governance. For the EU, once its AI Act enters into force in late 2023/early 2024, there will be an extraterritorial policy influence globally. This influence is often referred to as the "Brussels Effect", where EU regulations and laws are externalized to other other jurisdictions due to the EU's regulatory capacity and market size. Moreover, given the growing cooperation between the EU and US on AI governance, it is likely that the Brussels Effect for AI governance becomes a "Transatlantic effect". The establishment of the EU-US Trade and Technology Council (TTC), along with a joint AI Roadmap, signifies progress in their collaborative efforts to address AI-related risks. In May 2023, the EU and the US announced their intention to draft a voluntary code of conduct on AI, which would be open to other like-minded countries to sign up to.

However, despite their collaboration, the EU and the US are simultaneously competing for leadership roles in AI governance. For instance, it is noted that the US, with support from the UK and Canada, is working to water down a proposed binding worldwide treaty on AI put forth by the Council of Europe.

In line with China's overall strategic aspiration moving from a norm-taker to a norm-shaper, China searches for its role in global AI governance. Chinese AI experts have been actively participating in setting global AI standards. For instance, China's "Information technology–Artificial Intelligence–Reference architecture of knowledge engineering", proposed by China's Electronics Standardardization Institute, has been approved by ISO/IEC JTC 1, Information Security, sub-committee SC 42, AI. Additionally, China's AI regulations gain global impacts as default settings for Chinese technology exports.

Therefore, although different countries may share a conceptual alignment in their efforts to develop trustworthy AI, the competitive approaches they adopt and their ambitions to assume leadership roles in AI governance make regulatory convergence highly unlikely.

September 2023