Dr Elena Abrusci, Dr Hayleigh Bosher and Dr Alina Miron—written evidence (LLM0061)

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

1. This evidence is submitted by Dr Elena Abrusci (Senior Lecturer, Brunel Law School), Dr Hayleigh Bosher (Reader in Intellectual Property Law, Brunel Law School) and Dr Alina Miron (Lecturer, Department of Computer Science) at Brunel University London. The authors are members of the Brunel Centre for AI: Social and Digital Innovation.

2. This submission addresses questions 1,2 and provides regulatory and non-regulatory analysis and recommendations in response to questions 3-6.

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

3. Technical research and analysis suggest that LLMs will likely evolve along the following paths:

3.1. **'Dedicated models'**: Generalised Large Language Models have shown impressive capabilities in generating text in various topics, but they seem to lack the nuance for specific domains. LLMs have already been fine-tuned for specific applications, like models trained for medical (Med-Palm2), finance (BloombergGPT), retail (KAI-GPT) etc.[1] We will likely continue to see models fine-tuned for different tasks and the market for domain-specific models will grow to enhance user experiences and improve model performance.[2]

3.2. '**Bigger and better models'**: LLMs might continue to scale up in terms of model size, potentially leading to even more impressive capabilities. There are still many areas of research concerning LLMs, from solving the hallucination problem (they can produce factually incorrect information presented with conviction) to identifying and correcting potential model biases.[3]

---

1 *A large language model from Google Research, designed for the medical domain, Med-Palm*, https://sites.research.google/med-palm/; Introducing BloombergGPT, Bloomberg.com, https://www.bloomberg.com/company/press/bloomberggpt-50-billion-parameter-llm-tuned-finance/; KAI-GPT: The First Large Language Model Purpose-Built for banking, https://kasisto.com/blog/kai-gpt-the-first-large-language-model-purpose-built-for-banking/

2 T C Chen and others, 'Chat GPT as a Neuro-score Calculator: Analysis of a large language model's performance on various neurological exam grading scales' (2003) *World Neurosurgery*.

3 A J Thirunavukarasu and others, Large language models in medicine (2023) *Nature Medicine*, 1-11; E Kasneci and others, ChatGPT for good? On opportunities and challenges of large language models for education. (2023) Learning and individual differences, 103, 102274.

3.3. **'Multimodal models'**: LLMs trained on text-only data might produce inconsistent output on some tasks requiring basic world knowledge. We might see advancements in multimodal LLMs, where these models are trained on different data types, from text to sound and images/videos. For example, PaLM-E proposes an embodied language model where a robot is trained to perform tasks based on different real-world sensors in combination with textual data.[4]

3.4. **'Private models':** As concerns regarding privacy and security of LLMs increase, it is likely that in the next few years more 'private models' will be developed. These models prioritise data protection and could be preferred especially in highly sensitive application domains.[5]

3.5. **'Real-time information'**: It can take a long time to train LLMs, which result in their information not being up-to date. It is very likely that we will see research and development in this area to address the issue. For example, OpenAI launched plugins for ChatGPT, which extend the bot's functionality by granting it access to third-party knowledge sources and databases, including the web, meaning that it might be capable of processing more up-to-date information.[6]

What can be done to improve understanding of and confidence in future trajectories?

4. **Transparency** should be at the core of any effort to increase trust and confidence in new technologies. LLMs require even more transparency than general AI, both in terms of sharing code for the technical community and in terms of producing accessible information to the general public. The machine learning research community already share data and code and this level of transparency is extremely helpful to identify issues and limitations of the developed models. **Encouraging open-source research** could balance the close-sourced implementations that companies are employing. Moreover, an increased level of transparency could support a **model evaluation** for improving compliance with domestic and international standards and legislation. This would ultimately **improve trust and confidence from the general public** and potential users of large language models.

5. Transparency should be accompanied by **sustained technical research into risk and impact assessment methods and tools**, specifically tailored for LLMs. Transparency alone does not suffice to improve the understanding of the potential harms large language models should create. Encouraging investments into research domains like risk management and mitigation strategies is paramount in order to identify future trajectories in LLMs. Identifying challenges of LLMs like threats to privacy, increase misinformation and amplification of bias could help narrow down areas for improvement.

---

4 D Driess and others, Palm-e: An embodied multimodal language model (2023) *arXiv preprint arXiv:2303.03378*.

5 N Carlini and others, The secret sharer: Evaluating and testing unintended memorization in neural networks (2019) *28th USENIX Security Symposium (USENIX Security 19)*, 267-284.

6 ChatGPT Plugins [Beta], https://platform.openai.com/docs/plugins/introduction

6. Any work on large language models should be based on **multi-stakeholder engagement**. LLMs research and future trajectories are not developed in a bubble. There are several stakeholders involved, from policymakers, to general public and companies. Collaborative efforts can lead to a more comprehensive understanding of potential trajectories, and having a platform where stakeholders could provide input and feedback is essential.

Risks and opportunities of large language models (LLMs)

7. Large language models bring opportunities, including improved productivity and likely rapid advancements in different research fields. However, **the praise of these opportunities should not overshadow the several risks they may bring**.

8. LLMs have significant **risks of hallucination**, which could lead to harmful effects on individuals and society. 'Hallucinations' is a phenomenon where a large language model generates non-existent and false content. This **could mislead the users and, depending on the context where it is deployed, produce significant damages.** Galactica, a LLM by Meta used for summarizing academic papers, solve math problems and write scientific code, was shut down after just three days because it was not able to distinguish truth from falsehood.[7] There are other examples of LLM incorrect information already causing concerns. For example, two US lawyers were fined after ChatGPT cited non-existent case law.[8] In Australia, a regional Mayor has threatened to sue OpenAI for defamation if it does not correct ChatGPT's false claims that he had served time in prison for bribery.[9] This is a general issue with most LLMs and developers should closely monitor the behaviour of the LLM and act swiftly to limit the harm they may cause.

9. When using LLMs, there is a **high risk of overreliance on the model output**. While these models have demonstrated remarkable capabilities in generating human-like text and assisting with a wide range of tasks, excessive **dependence** on their output can lead to various problems, from increased impact of misinformation, **distortion** and amplification of biases.[10] To counter this, public and private efforts should be made to increase the digital literacy and understanding of the general public. Digital literacy empowers users to comprehend the fundamentals of LLMs, such as how they work, their limitations, and their potential for generating accurate and inaccurate information.

---

7 W D Heaven, Why Meta's latest large language model survived only three days online, (MIT Technology Review, 18 November 2022) https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/

8 Two US lawyers fined for submitting fake court citations from ChatGPT (The Guardian, 23 June 2023), https://www.theguardian.com/technology/2023/jun/23/two-us-lawyers-fined-submitting-fake-court-citations-chatgpt

9 Australian mayor readies world's first defamation lawsuit over ChatGPT content (Reuters, 5 April 2023) https://www.reuters.com/technology/australian-mayor-readies-worlds-first-defamation-lawsuit-over-chatgpt-content-2023-04-05/#:~:text=SYDNEY%2C%20April%205%20(Reuters),against%20the%20automated%20text%20service.

10 A Birhane and others, Science in the age of large language models (2023) *Nature Reviews Physics*, 1-4.

10. **Large language models increase the risks of proliferation of content manipulation and misinformation.** The possibility to produce a high amount of content, in short time and with ease could be used by malicious actors to fuel disinformation campaigns. Research confirms that as the use of LLMs for malicious campaigns drives down the costs and increases accessibility, new actors will be likely be tempted to spread disinformation.[11] Moreover, it is also extremely difficult for existing human and even AI powered tools to confidently distinguish content produced by generative AI. This would further challenge the existing measures in place to counteract misinformation, disinformation and other harmful content.

11. Even more than other systems using AI, LLMs have a **structural risk of unintended privacy leaks.** These leaks occur when LLMs inadvertently generate or disclose sensitive information about individuals or organisations, often due to the vast amount of data they have been trained on. These leaks usually happen due to two situations:

11.1. **Training process**: LLMs learn from a diverse range of sources which can contain personal information, confidential documents, or proprietary data. Several studies have shown that **unintended memorization is a persistent, hard-to-avoid issue that can have serious consequences**.[12]

11.2. **Generation stage**: LLMs excel at rephrasing and paraphrasing, which can inadvertently reveal private details when given seemingly harmless prompts. Both situations require that the technical development of LLMs and its regulatory framework are strongly rooted in data protection and that the ICO is equipped with the right expertise to monitor compliance.

12. **LLMs are likely to amplify existing bias**. This phenomenon occurs because LLMs learn language patterns and associations from the vast amount of text data they are trained on, and this data often reflects the biases present in society. This can lead to **discriminatory or exclusionary messaging.**[13] A rigorous risk and impact assessment informed by a strong enforcement of the Equality Act are needed to limit the harm this may cause.

13. There are **risks relating to intellectual property law for creators, rightsholders and AI developers**. This is because there is uncertainty about how copyright law and related rights apply in the context of AI. The law and process must be clarified in relation to AI training (input) and AI generated (output).

---

11 Forecasting potential misuse of language models for disinformation campaigns- and how to reduce risks (Stanford Internet Observatory Cyber Policy Center, 11 January 2023), https://cyber.fsi.stanford.edu/io/news/forecasting-potential-misuses-language-models-disinformation-campaigns-and-how-reduce-risk

12 N Carlini and others, The secret sharer: Evaluating and testing unintended memorization in neural networks (2019) *28th USENIX Security Symposium (USENIX Security 19)*, 267-284.

13 S Moni, Overcoming Algorithmic Gender Bias In AI-Generated Marketing Content, Forbes, 25 July 2023, https://www.forbes.com/sites/forbescommunicationscouncil/2023/07/25/overcoming-algorithmic-gender-bias-in-ai-generated-marketing-content/?sh=778e78ac1639; R Navigli et al. 'Biases in Large LanguageModels: Origins, Inventory and Discussion' (2023) *ACM Journal of Data and Information Quality*.

Regulatory and non-regulatory recommendations (in response to questions 3-6)

General recommendations

14. **LLMs are based on the same functioning of other AI and ADM systems, and they should therefore be regulated as such.** As argued elsewhere, the current UK legal framework on data protection, non-discrimination and human rights is appropriate to regulate the core components of any AI and ADM system and any specific regulation should build on them and align with the existing legislation.[14] The Data Protection Act 2018, the Equality Act 2010, and the Human Rights Act 1998 should inform any regulation of AI and LLMs. However, the AI White Paper fails to provide a **framework for an adequate enforcement and application of the existing regulation** to the specific challenges posed by AI and LLMs.

15. The **AI White Paper falls short of protecting individuals against the risks that AI**, and large language models, can pose to their human rights. As confirmed by the Equality and Human Rights Commission, the White Paper does not provide enough guarantees of how it will protect individual human rights.[15] As large language models continuously develop and change, **an adaptive regulatory framework is essential**. This should cover all stages of the AI lifecycle. However, this regulatory framework to be effective needs to be accompanied by thorough human rights impact and risk assessments, robust oversight mechanisms and accessible remedies and grievance mechanisms.

16. The White Paper briefly foresees the need to conduct risk assessment of AI technologies, including LLMs, but it puts the burden on the regulators and not on the providers or deployers of the technology. In the context of LLM, it is **crucial to undertake continuous and rigorous risk and impact assessment**, to be able to identify the likely impact on the individuals and society. However, as these assessments should be conducted frequently and demand advanced resources and access to information, **they can be effectively conducted only by the provider and deployer of the technology and not by the regulator**.

17. Any regulation of large language models needs to be supported by **a strong oversight body**. The oversight mechanism will be able **to identify the emerging risks posed by LLMs and adjust the enforcement of the legislation as needed.** This function could be carried out by the Digital Regulation Cooperation Forum (DRCF) or separately by the existing regulators but will require appropriate investment in terms of resources, skills and powers. Alternatively, a dedicated body could be established, following prior proposals for a Digital Authority or following the example of the European Artificial Intelligence Board provided by the EU AI Act. A strong and well-equipped oversight body will

---

14 E Abrusci, R Mackenzie Gray-Scott, 'The questionable necessity of a new human rights against being subject to automated decision making' (2023) 31(2) International Journal of Law and Information Technology.

15 Equality and Human Rights Commission, Consultation response to the AI White Paper, 2023.

deal with all systems using AI, without the need to establish a dedicated body for LLMs only.

18. In addition to this, any regulation should ensure that individuals who may be impacted by the harmful effects of large language models have **easy access to grievance mechanisms and remedies**. These should be ultimately provided by the providers and deployers of the LLMs but should be supervised by relevant regulators.

19. All these regulatory efforts should be accompanied by **sustained investment into digital literacy**. Equipping all individuals in society to understand and consciously use large language models is necessary to ensure confidence and trust. Digital literacy programme that goes beyond the basic understanding of the web should be provided as part of a **multi-stakeholder initiative** where all the actors involved, public authorities, private companies and civil society, contribute to make end users aware of the risks and benefits associated with LLMs and the needed precautions to take when using these tools.

Intellectual property and copyright analysis and recommendations

20. The White Paper mentions risks in relation to mental health, privacy rights and human rights, but not any threats to intellectual property, creators or culture. **The White Paper mentioned intellectual property rights only once**, and then only to state that the Government intended to follow the suggestions of Sir Patrick Vallance. This suggested an extension of the current text and data mining copyright exception. There is a **lack of evidence that AI firms are facing a barrier to innovation as a result of copyright**. Only 13 out of 88 responses to a UK IPO consultation were in favour of a broad any-purpose exception.[16]

21. It appears that the **Government fails to understand that AI firms and copyright rightsholders are not mutually exclusive**. Copyright industries - particularly the creative industries – are utilising AI and AI firms could be rightsholders, for example in their database.

22. There needs to be consideration as to **whether AI generated content is protectable by copyright or not.** Copyright only protects "original" literary, artistic, dramatic and musical works. In copyright law the test for originality relates to the creative choices made by the creator, it is about drawing inspiration from what has already done and adding a personal touch. Some argue that AI generated content is merely a borrowing of what already exists, and since it does not add any personal touch, it does not quality for copyright protection.

23. **Some copyright works, such as sound recordings, do not have the same originality requirement.** This is simply because a sound recording is a fixation of a musical work and therefore cannot be original by definition. This

[16] Consultation outcome: Artificial Intelligence and IP: copyright and patents, (28 June 2022), https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents

right is shorter than the rights of original creations, lasting 70 years since publication rather than 70 years after the death of the creator. If it is decided that AI generated works should be protected it will also need to be decided to what extent that work is protected. For example, how long should it last and what type of rights apply to it. **To decide whether or not to grant copyright protection to AI generated works would require considering copyright justifications** - the reasons that these rights are granted – and if they apply in this context. Copyright, under UK principles, essentially aims to encourage and reward the creation and dissemination of knowledge and culture.

24. AI, and LLMs, use creators' work to generate more work. As the recent report from **the UK Science, Innovation and Technology Committee** noted that one of the 12 essential challenges for AI and policy is that AI models and tools make use of other people's content and policy must establish the rights of the originators of this content, and these rights must be enforced.[17] The recent report from the Culture, Media and Sport Select Committee also called for the Government to drop any plans to allow AI copyright-free use of text and data to protect the creative industries.[18] **Therefore, the Government should:**

24.1. **Confirm that there will be no extension to the text and data mining copyright exception**. If any exception is deemed necessary, it must be narrow and specific to overcome a barrier to innovation that is clearly evidenced.

24.2. **Confirm that the use of AI training data includes the use of copyright protected content,** for which rightsholders are entitled to be remunerated and creators are entitled to be acknowledged. For this licensing systems and transparency requirements need to be in place.

24.3. **Consider if AI generated works should be protected by copyright**. AI generated works are unlikely to be deemed original for the purpose of copyright protection under UK Law. The US Copyright Office has also rejected registrations of AI generated works as protectable.

24.4. If the Government did decide that AI generated works could be protectable, then to **consider the justification and extent of the rights granted**. In particular, the Government should consider the impact of AI generated works on creators and human created works. Copyright mechanisms can be used to uphold and protect human created works to a higher degree than AI generated works for example by limiting the scope of AI generated rights.

September 2023

---

17 Science, Innovation and Technology Committee, The governance of AI, Interim Report, https://committees.parliament.uk/committee/135/science-innovation-and-technology-committee/publications/

18 Select Committee on Culture, Media and Sport, 30 August 2023, https://committees.parliament.uk/committee/378/culture-media-and-sport-committee/news/197222/abandon-artificial-intelligence-copyright-exemption-to-protect-uk-creative-industries-mps-say/