# AGENCY—written evidence (LLM0028)

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

## A. INTRODUCTION

AGENCY is a multidisciplinary research team of academics with expertise in computer science (natural language processing, cybersecurity, artificial intelligence, human-computer interaction), law, business, economics, social sciences and media studies. Members of AGENCY are academics at prestigious institutions such as Newcastle University, Durham University, University of Birmingham, King's College London, Royal Holloway University of London, and University of Surrey. UK Research and Innovation supports our research through the Strategic Priority Fund as part of the Protecting Citizens Online programme. Grant title: AGENCY: Assuring Citizen Agency in a World with Complex Online Harms. Grant reference: EP/W032481/2

This call for evidence of the future of large language models (LLM) and regulation coincides with the work, expertise, and concerns of the AGENCY project, which focuses on assuring citizen agency in a world with complex online harms. We refer to citizen agency as the ability for people and society to be empowered through technology and tools that provide them with a sense of control and security in that space. Thus, we propose that people and society should be at the forefront of regulation and that regulation should aim to premeditate, mitigate, and respond to complex online harms in a way that empowers people and balances that empowerment with societal concerns (such as public health, safety and security) while ensuring respect for principles such as freedom of expression. Our team possesses specialised expertise in LLM, law, emerging technologies, and their non-regulatory solutions. Accordingly, it is our responsibility to submit our response to this call for evidence, as we are well-positioned to make a useful contribution in this area.

This is a submission from AGENCY. Specifically, the following researchers contributed to the formulation of this response: Dr Shrikant Malviya, Rebecca Owens, Dr Jehana Copilah-Ali, Prof Karen Elliott, Prof Ben Farrand, Dr Cristina Neesham, Dr Lei Shi, Dr Vasilis Vlachokyriakos, Dr Stamos Katsigiannis and Prof Aad van Moorsel.

## B. CAPABILITIES AND TRENDS

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

In the future, there is the potential for rapid and unexpected change in LLM, and we expect them to develop in the following ways:

1. **Increased proliferation of smaller, domain-specific models:** The proliferation of smaller, domain-specific models represents a notable trend in generative AI. These models are designed to cater to specific tasks, industries, or domains, and they contrast with large, general-purpose models like GPT-3. They have several benefits, i.e tailored domain-specific expertise, faster training and deployment, reduced bias and ethical concerns, improved privacy and security, customisation and personalisation.

2. **Convergence of LLMs and Multimodal Interactions:** The capabilities of LLMs will be expanded by enabling them to interact with users through various modalities beyond just text. These modalities include images, videos, audio, augmented reality (AR), virtual reality (VR), and even robotics. This expansion will enhance the depth and richness of user interactions with LLMs, making them more versatile and capable of understanding and generating content across different formats.

3. **Advancing LLM Performance through Reinforcement Learning and Human Feedback Enhancements:** A key strategy to improve the capabilities and effectiveness of Large Language Models (LLMs) is incorporating reinforcement learning (RL) and human feedback (HF) enhancements. This is critical for enhancing users agency by addressing bias mitigation, contextual understanding, ethical responsiveness, and user-centricity issues.

4. **Multilingual Capabilities:** LLM will become more proficient in handling multiple languages and understanding context and nuances across different languages.

5. **Self-improving LLMs:** Drawing inspiration from the mechanisms of human learning, next-generation artificial intelligence systems may possess the ability to self-train, opening up new uses for LLMs.

6. **Fact-checking themselves:** The current LLMs suffer from factual unreliability and static knowledge limitations of large language models. 'Hallucinations' are one of their critical issues, for example, they recommend books that do not exist and confidently forecast the weather for a given fictional city. The following requirements could be crucial before LLMs are used for widespread real-world deployment such as the ability to provide valid citations and references for the answers they provide. LLMs require plenty of improvement and innovation in this area to overcome the shortcomings of their unreliability and their stubborn tendency to provide inaccurate information confidently.

**Future trends** relating to complex online harms,

- **We foresee an increase in publicly available foundation models which allow users to fine-tune models for specific uses.** Whilst this offers the advantage of low-cost creation of robust models for many applications, it also enables malicious actors to create models that cause harm, such as a model fine-tuned on hate speech or with intentional bias against specific groups of people.

## 2. How should we think about risk in this context?

In our opinion, risk should be considered through five key issues:

1. **Access**: We should limit who has access to robust AI systems, structuring the proper protocols, duties, oversight, and incentives for them to act safely.

2. **Alignment:** Ensuring that the AI system will act as intended in agreement with socialised human ethical sensibility (i.e., values and norms)

3. **Raw intellectual power:** Grade the generative AI systems on raw intellectual/processing power, which depends on the level of sophistication of the algorithms and the scale of computing resources of datasets (source).

4. **Scope of actions:** Point out the potential for harm in AI systems based on the scope of actions that can be indirect, for example, through human actions (misinformation, data privacy or cybersecurity risks) or directly through the system itself/other AI agents ( stereotypes, unfair discrimination, exclusionary norms, toxic language etc), such as intentionally training LLM biased against specific groups of people, or training them with specific misinformation (sources).

5. **Unlicensed Text Usage:** One significant concern and potential legal problem is the unauthorised use of extensive text for training LLMs. Many websites, digitised books, and magazines prohibit such usage or allow reuse once the source is properly acknowledged or referenced. When a Large Language Model generates output containing verbatim text from sources, it typically does not meet these requirements, potentially leading to legal issues (sources).

## C. NON-REGULATORY AND REGULATORY OPTIONS

We are calling for **a human-centred, responsible innovation approach** towards developing and regulating LLMs.

- **Regulators and companies must work together** to facilitate substantial transparency in generative AI technologies, thereby ensuring accountability to society. To counteract the potential for deception, scams or other forms of misuse originating from generative AI, a comprehensive two-pronged security strategy can be implemented, comprising technical measures and policy interventions. Watermarking is an example of a technical measure to help know whether AI generates the output.

From a **regulatory standpoint**,

- **Regulatory interventions are appropriate before LLMs come to market.** This will help ensure that AI systems are designed and deployed in ethical, legal, and technically robust ways and ensure citizens are protected from complex harms. **The best way to achieve this is policy measures akin to the 'secure by design' approach** (source) where safeguards, fail-safes and other mechanisms for harm reduction are co-designed with users to ensure they are protected from harms before they occur.

- **Transparency obligations should be imposed on the creators of LLMs** in order to address risks by providing users with the agency to understand and evaluate LLM operations. In our view, the proposed EU AI Act (source) may provide some guidance as it obligates companies that develop LLM to guarantee the prevention of generating illicit content and provide summaries of copyrighted data used during training. Such an approach may provide a reference point for developing a framework that the UK government can use to ensure LLM transparency and improve user trust.

- **To complement this, the government needs to create a framework for AI liability similar to the EU's proposed AI Liability Directive.** This would make it easier for users suffering AI harm to bring civil liability claims against manufacturers and organisations using AI by creating a rebuttable presumption of causality, thereby allowing users to be protected.

**Non-regulatory options**

- We propose **the inclusion of frameworks** to address risks and capitalise on opportunities that can be applied to services engaged with LLM. One such example is the adoption of an adaptable **Corporate Digital Responsibility (CDR) framework** (source)

- The CDR framework can encourage services that utilise and deploy LLM to engage in practices and exercises that consider ethical practices from design to delivery in socially, economically, and environmentally responsible ways. CDR also allows stakeholders to focus on the responsibility of LLM to enhance positive societal impacts while minimising harmful and unintended consequences beyond legal obligations.

- Incorporating pre-existing tools can support teams working with LLM **to assess unintended consequences**; an example is the Consequence Scanning Manual available in resources from the Open Data Institute (source).

- These practices introduce non-regulatory exercises that facilitate consideration of the impact LLM can have on the future in a way that ensures **responsible innovation through adherence to ethical procedures** that do not hinder innovation (for example, the UKRI Framework for responsible AI research and innovation (source).

- A **common barrier to sustaining non-regulatory practices** in the LLM space - is the fear that implementing additional procedures such as CDR would slow processes and ultimately hinder innovation.

- We propose the following examples of ways in which the **CDR framework can work in practice**, such as,

    - Including exercises and knowledge sharing on CDR, Responsible Innovation, and Ethics by Design when onboarding employees who will work with LLM.

    - To be supported at different delivery stages by including a CDR Champion on teams who actively seek CDR innovative approaches with company-wide feedback as a simultaneous top-down and bottom-up approach to CDR compliance to permeate company culture.

    - We also propose that the responsibility of the CDR Champion should rotate throughout the LLM software development life cycle to avoid burnout and to inculcate CDR as a common practice.

- A CDR framework is also a valuable addition to support standards in this area, such as the ISO 27001 certification in Information Security Management and the IEEE Standards Association, P7010- Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems.

- We propose that **post-market human oversight must be involved** to identify errors in the LLM output and mitigate against complex online harms.

## D. DOMESTIC REGULATION

### 1. How adequately does the AI White Paper and other Government policies deal with LLM, is a tailored regulatory approach needed?

We believe that the **sector-specific regulation** proposed by the UK Government White Paper AI Regulation: A Pro Regulation Approach **is ill-equipped** to protect users from the complex harms stemming from LLM for several reasons.

- The multifaceted nature of foundational LLM, such as GPT4, means that they **influence every area of society and can facilitate complex online harms** such as misinformation and the production of malicious content.

- It needs to be clarified **how the central functions will coordinate key stakeholders,** such as those representing vulnerable groups who may be affected by LLM and aid the regulatory interpretation of principles that cannot be easily quantified.

- From the government's proposals, it is **uncertain how the framework will apply to LLM in many sectors** and how it will apply to domains without a clear regulator. This lack of clarity may lead to different regulatory standards being applied to LLM, which may not protect users.

- Further information is needed on **how the government intends to uphold citizens' procedural rights** when trans-sectoral disputes occur.

Therefore, to promote consumer trust in the technology, **an integrated, cross-sectoral regulatory strategy is better suited for LLM as it would enable regulators to pool expertise and resources, promoting a more efficient approach to regulation.**

**2. 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?**

Existing **UK regulators do not have the required expertise, resources and powers** to effectively comprehend the complex technical and ethical aspects of LLM and improve public trust in using AI models.

In particular, we would like to draw attention to the fact that:

- **The recent AI: A Pro Innovation Approach White Paper proposes significant new regulatory responsibilities without providing any additional funding.** As such, there is a need for the government to allocate clear funding to regulators specifically for the regulation of LLM.

- **There is also an urgent need to improve regulators' technical capabilities.** Currently, whilst regulators such as the ICO and Ofcom have in-house access to specialists in natural language processing, many smaller regulators do not have the capabilities to assess LLM.

- **Regulators do not have the necessary power to audit LLM systems.** Much of the information used is proprietary; this means that some regulators do not have the necessary mandate to access it and understand how LLM functions.

- **Regulators should be allowed to pool resources** to address capability gaps and promote capacity building.

## E. INTERNATIONAL CONTEXT

### 1. How does the UK's approach compare with that of other jurisdictions, notably the EU, the US and China?

In our opinion, the **UK's regulatory approach differs significantly** from other jurisdictions as they opt for **a more stringent approach to regulation.** We call on the government to learn from these approaches and adopt certain regulatory features from other jurisdictions to ensure the **creation of a human-centred, responsible innovation approach to regulation.**

*EU*

- **The UK should adopt a right to know they are communicating with an LLM.** Under the proposed EU AI Act, companies must disclose when content is AI-generated, guarantee the prevention of generating illicit content, and provide summaries of copyrighted data used during training. In our view, the UK should adopt a similar approach to enhance user agency and increase trust in LLM.

- **Clear penalties are needed.** The EU AI Act (source) allows regulators to impose fines amounting to 6% of global annual turnover. However, no clear sanctions are provided in the UK's current legislative response.

- **Easy routes for redress for LLM harms are needed.** The EU proposed AI Liability Directive (source) makes it easier for users suffering from AI harms to bring civil liability claims against manufacturers and organisations using AI by creating a rebuttable presumption of causality. The UK should take a similar approach and establish a clear liability framework to protect users against the complex harms of LLM.

*US*

- **The UK should enshrine the rights of users of LLM** like the US AI Bill of Rights (source) to promote a human-centred approach and promote users' trust in LLM**.** Such an approach ensures user's rights to the creation of safe and efficient systems, protection from algorithmic discrimination protection, the promotion of user agency, and notification of AI use.

*China*

- **The Government should implement a misinformation prevention obligation on LLM companies.** In April 2023, China released a draft of its generative AI regulatory approach. It requires securing the truth, objectivity and diversity of training data, does not discriminate, and provides truthful information (source). Whilst we recognise that this is a high standard for

companies to achieve given the complex harms that can arise from the spread of disinformation through LLM, we call on the government to require companies to actively prevent the spread of misinformation through their products actively in the development stage by ensuring the products scrape truthful information and improve their tendency to hallucinate information.

*4 September 2023*