The Royal Academy of Engineering—written evidence (LLM0063)

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

About the Royal Academy of Engineering

The Royal Academy of Engineering is harnessing the power of engineering to build a sustainable society and an inclusive economy that works for everyone. In collaboration with our Fellows and partners, we're growing talent and developing skills for the future, driving innovation and building global partnerships, and influencing policy and engaging the public. Together we're working to tackle the greatest challenges of our age.

Introduction

Since 2019, the National Engineering Policy Centre (NEPC) has been exploring the safety and ethics of autonomous systems to understand risks and benefits across different sectors (such as healthcare and transport). The project seeks to understand how autonomous systems can be ethically designed, developed and deployed to ensure benefits are widely distributed and no one is disadvantaged.

The Royal Academy of Engineering has also recently undertaken a series of interviews with Academy Fellows and Awardees who are experts in the field, providing crucial insights on topical themes surrounding the development and deployment of generative AI.

The comments below are drawn from these two streams of work, as well as input provided by a selection of Academy Fellows, and has important implications for the Communications and Digital Committee's inquiry on the opportunities and risks related to large language models (LLMs). More broadly, this response reflects on the future of artificial intelligence (AI) opportunities and threats, and the role generative AI and LLMs will play.

The Academy is grateful for the opportunity to contribute to this important inquiry and would be delighted to assist the Committee in any practicable way as it considers the issues we have outlined.

Summary

The following document is a comprehensive Academy response that addresses the '*capabilities and trends'* and '*domestic regulation'* questions posed in the consultation's call for evidence. This summary provides the key themes which are explored fully in the response which follows.

Capabilities and Trends

The Academy's response highlights key developments of LLMs over the next three years such as increased scale, improved performance, increased diversity, improved human-machine partnerships and interfaces, and increased need for international collaboration. These development areas present opportunities to grow the UK economy, bring better quality jobs, and improve public and private services. Given the uncertainty in future trajectories, a list of suggested UK principles such as transparency, multilateral engagement, and horizon scanning are included to help improve the understanding and confidence of the direction of LLMs.

With the development of LLMs lies several opportunities and risks to consider. There are opportunities for the UK to play a major role in the development of reliable benchmarks and evaluation methodologies as well as the technological capabilities of LLMs to conduct complex analyses of large data rapidly and its increased computing power to improve efficiency and accuracy. The response also outlines the greatest risks to consider, such as the monopoly of the largest tech companies on limiting market competition, the quality of data on performance and accuracy, and the societal risk of the exploitation of LLMs with disinformation. These risks should be understood in relation to its application to reflect the needs of users and capabilities of the system.

Domestic Regulation

In this section of our response, the adequacy of the Government's AI White Paper's coverage of LLMs is addressed alongside suggestions for UK regulators to respond to LLMs. While the White Paper offers a sensible approach, more attention should be given to which stakeholders are accountable for cross-cutting issues, supporting collaboration between industry and regulators, and international engagement. Regulation of LLMs need to be agile to adapt to the consistently changing nature of LLMs while also considering governance internationally to guard against isolation. Domestic regulation will require diverse, multistakeholder expertise to ensure safe use and deployment without restricting innovation.

Further detail in response to these questions is shared throughout the response below. The Academy welcomes the opportunity to clarify any points presented and provide further contributions to the inquiry as needed.

Capabilities and trends

Q1. How will large language models develop over the next three years?

1. An LLM is a type of AI model that is trained to identify patterns within text data. While LLMs may use generative AI, models that create new content based on received input, there are applications of LLMs that are not generative. LLMs can be used for a variety of natural language processing (NLP) tasks, such as language translation, content creation and sentiment analysis. While LLMs have long been in existence, recent advances in deep learning algorithms, data architectures, computing power and data accessibility have seen their capabilities grow exponentially – making them a key element within the broader AI landscape.

2. While it is difficult to be confident in future projections, recent developments in LLMs provide some indication of what developments over the next few years could look like. The main developments are likely to include increased scale, improved performance, and improved human-machine partnerships and interfaces. These developments present significant opportunities to grow the UK economy, bring better quality jobs and improve public and private services due to the expansive capabilities of LLMs.

Increased Scale

3. As LLMs have increased in scale so has their accuracy. Newer versions of ChatGPT, for example, can now accurately match grammatical gender of words in languages such as German and French 100% of the time. In the earlier versions, ChatGPT was only able to achieve an accuracy of about 50%. This improvement in accuracy is the result of using larger sets of training data which allow more parameters to be set and weighted. The scale of LLMs is anticipated to increase further in response to increased demand for more accurate models and different applications. This is likely to result in more emergent behaviours, such as the ability to perform tasks that an LLM was not trained on.

Tailored for Applications

4. The launch of ChatGPT created significant interest in LLMs, prompting commercial organisations, researchers, and others to develop new approaches to using them. As this attention continues, developers are likely to address known problems with LLMs, particularly 'hallucinations' (when the model states supposed facts which are not reflective of reality) and data bias, in order to improve the effectiveness of LLMs and enable better results for more applications.

Improved Human-Computer Interfaces and Partnerships

5. Over the next three years, human-computer interfaces are also likely to improve alongside generative AI, as currently, there is a lack of emphasis on human-centric development. Advances in sensing, tracking, analysing,

and animating human nonverbal communicative signals, could see significant improvements in computers being more affective partners in human-computer interaction.

Increased Diversity

6. It is very unlikely that general LLMs such as Open AI's GPT-4, Meta's LLaMA or Google's PaLM 2 can be developed outside of the big tech companies. However, a new trend is beginning to emerge: the development of application specific LLMs which use proprietary data rather than the internet. For example, LLMs trained on data from Electronic Patient Records (EPRs), medical textbooks and research papers will have a very significant impact on healthcare. By virtue of being trained on and fed data sourced from curated and manageable databases, such LLMs are likely to have superior accuracy, consistency, usability and accountability when compared with general-purpose ChatGPT like LLMs. These databases can be shared widely and used for multiple applications.

Increased Need for International Collaboration

7. Greater international collaboration on standards and data access is already needed, and likely to increase further in the next three years. To ensure that LLMs (and AIs more broadly) are transparent and trustworthy within an increasingly international marketplace, it will be necessary for standards, for example regarding explainability, and certifications to be universally recognised. Similarly, there is a clear need for improved international data access to ensure that the datasets and reference libraries needed for the training and monitoring of the largest LLMs are not the exclusive purview of certain companies or jurisdictions.

Q1a) Given the inherent uncertainty of forecasts in this area, what can be done to improve understanding of and confidence in future trajectories?

8. The nature and speed of the development of LLMs and generative AI is difficult to accurately predict, but there are several methods that can be used to better understand uncertainties and future trajectories of LLM development. One method for uncertainties is to conduct an opportunity analysis to map the range of LLM capabilities. Another method for future trajectories is horizon scanning. Both methods uncover potential opportunities, risks, vulnerabilities, and biases that can enhance preparedness for future scenarios. These measures should be informed by insights from cutting edge researchers in academia, start-ups, and large tech companies.

Q2. What are the greatest opportunities and risks over the next three years?

9. The deployment of LLMs offers opportunities to increase productivity and improve the delivery of services within the public and private sectors, creating benefits for the UK's economy, and the health of its citizens. To

realise those benefits will require a rapid and unified response from the public and private sector to identify what best practice for adoption looks like and what risks need to be guarded against.

Opportunities

10. Since LLMs can automate cognitive work, as opposed to physical tasks, a boost to economic productivity could happen more quickly than in past technological revolutions. LLMs may raise overall productivity by helping less skilled workers the most, decreasing the performance gap between employees. Increasing use of LLMs could also provide rapid access to a breadth of high-level knowledge. This is likely to support improved productivity where risk appetites allow, and depending on progress, reduce occurrence of errors or 'hallucinations'.

11. LLMs can rapidly conduct very complex analyses of large amounts of data, making them useful for many traditionally time- and resource-intensive applications. For example, LLMs can significantly hasten the discovery and development of drugs by quickly analysing large numbers of complex chemical and biological interactions, allowing for the rapid identification of drug candidates.

12. LLMs can also quickly extract information from large amounts of unstructured data, for example posts on social media. Applied to a business context, this can allow for an organisation to aggregate data from a variety of different sources and structure it relatively rapidly. This has applications in creating greater visibility of supply chains, or closer understanding of consumer interests.

13. If organisations and states are supported to develop domain specific LLMs based on shared access to curated data sources there will be market diversification. This will result in a global LLM marketplace that better represents a diverse range of cultural, societal, and economic values. There is, however, also a risk that the development of sovereign LLMs, without appropriate international cooperation agreements, could exacerbate divisions between countries and organisations by creating new barriers to accessing data and collaborating internationally.

14. There is an opportunity for the UK to play a major role in the development of comprehensive and reliable benchmarks and evaluation methodologies to assess the capabilities of LLMs across a wide range of contexts and markets. This will require greater transparency of commercial models, greater assessment of the reliability of such models, their training and evaluation regimes and their performance data. Such transparency has the potential to improve the safety and trustworthiness of LLMs, increasing public and market confidence, leading to significant societal and economic benefits.

15. Increased computing power will enable larger data sets to be used with greater efficiency, creating opportunities to further improve the accuracy of LLMs. The value of this opportunity could be further increased through

developing agreements to improve international access to computing power.

Risks

16. Presently, only the largest technology companies in the world have the capacity to duplicate what OpenAI has done with GPT-3 and GPT-4. While smaller companies and governments can produce smaller-scale models independently, there are certain tasks that can only be accurately completed by models of a certain size. As governments actively work with these companies to define the direction of AI regulation, there is a risk of excluding innovators and limiting market competition.

17. LLMs and generative AI models can be exploited by nefarious actors. For example, tools such as ChatGPT can be used to generate effective phishing emails that could be used to initiate a cyberattack. The exploitation of LLMs also poses a societal risk. LLMs can be used to produce convincing disinformation in the form of 'deepfakes' and falsified news reports. In a political context, this constitutes a significant risk and forms a potential obstacle to democracy. This could be mitigated by mandating the use of watermarking technology to be built into image and video generators to identify AI-generated content. Researchers are developing tools which alter photos in ways that are invisible to the human eye but prevent them from being manipulated.

18. A multistakeholder approach is needed to define and determine appropriate uses for LLMs that are both technically feasible and socially desirable. Appropriate uses will differ depending on sector and the specific tasks according to risk appetites for different applications, emphasising the need for various stakeholders like researchers, funders, and industry to provide their perspectives alongside government and regulators. Together these stakeholders will be in a better position to consider what determines appropriate use.

19. Even when LLMs and generative AI are used appropriately there will be risks of incorrect answers or biased outcomes. The effectiveness of existing frameworks for risk management and reporting accountability must be considered by appropriate stakeholders in this context. Clarity on who is accountable for the way that LLMs are developed, trained, and deployed will be crucial to minimising the risks associated with large-scale deployment. These deployment risks should also be managed through the implementation of safeguards such as runtime monitoring or fault-tolerant architectures.

20. The quality of data is crucial to the functionality and performance of LLMs. Poor quality or polluted data affects performance and accuracy. LLM performance could be drastically improved by using closed-ended LLM methods that are based around the use of well-curated and monitored datasets, for example enterprise databases, rather than internet-scale public datasets. As companies seek to increase the scale of their training databases for LLMs, it is crucial that stakeholders collaborate to access and incorporate sources of high-quality, representative data.

21. Presently, a lack of appropriate training data constitutes a bottleneck for the development of LLMs in some sectors. If managed carefully, there is an opportunity for the UK to leverage its unique data assets, such as NHS digital health records. Under privacy preserving data sharing agreements, the NHS could be a potential source of high-quality data, especially if LLMs continue to improve their ability to interpret free text in medical records.

22. There is a risk that the UK fails to leverage its significant strength in ML and AI to realise the economic and societal benefits from LLMs and generative AI. Should the UK fail to develop rapidly as a hub for the development and implementation of LLMs, and other forms of AI, it is likely to lose influence in international conversations on standards and regulatory practices. UK domestic AI regulatory frameworks must therefore be outcome focused, globally relevant and stimulate innovation.

Q2a) How should we think about risk in this context?

23. The risks associated with LLMs should be thought of in relation to where they will be applied. Use cases for LLMs should appropriately reflect the capabilities of the system, and the needs of users. Where there is a mismatch between capability and need, policymakers will need to develop safeguards.

24. Government departments and regulators with oversight of early adopter sectors should add new risks associated with AI to their existing risk registers to ensure preparedness for any reasonable worst-case scenarios associated with those risks. These risks should be allocated to accountable officials with responsibility for assessing risks and opportunities as future technology trajectories changes. Processes established for the reporting of risks should also conform to international accounting guidelines.

25. Policymakers should also be attentive to the quality of the data being used by LLMs as a way to minimise the uncertainty of outcomes. The quality of the data used to train models has a direct impact on their performance. However, many stakeholders still do not have a clear understanding of how data is collected, or how it is being used – making unintended consequences of deployment a significant source of risk. As policymakers are in a unique position to oversee how LLMs access our data, and how LLMs trained on public data are governed, this should be seen as a key interface for the reduction of risk.

26. There is a need to actively try to anticipate the risks and unintended consequences of deployment in different environments. It may be helpful to create a reporting system for harms and close calls. Policymakers could begin developing specific environments for applications to be tested. Having a diversity of voices included in the development of such environments is crucial to ensuring these environments reflect the contexts in which models would be deployed and the diversity of impacts on different demographics.

Q3. How adequately does the AI White Paper (alongside other Government policy) deal with large language models? Is a tailored regulatory approach needed?

27. While the principles in the AI White Paper offer a foundation to consider and deal with some of the key risks associated with LLMs, it does not address how LLMs will be updated, monitored, or how accidents, injuries and other forms of harm that could result from deployment will be investigated. It is important to note that this same point should also apply to non-LLM systems.

28. It could agree principles with industry to maximise confidence and opportunities while managing risk similar to the US AI Bill of Rights, for example, which has five principles meant to guide the use of AI—systems should be safe and effective, non-discriminatory, protective of people's privacy and transparent; people should be notified when a system makes a decision for or about them, be told how the system operates and be able to opt out or have a human intervene.[1] The following are suggested principles gathered from selected Academy fellows to include (with the caveat of the existing implementation challenges from a societal or business perspective):

    a. Open to Scrutiny – enabling independent assessment of the properties and qualities of LLMs and sharing of best practice and lessons learned from LLM training and deployment.

    b. Open Competition – creating a rich ecosystem of service providers by regulators exercising their power to monitor market concentration and prevent the formation of monopolies within the LLM provider market.

    c. Secure Data Infrastructure – establishing and maintaining a robust data infrastructure that maintains high levels of data quality whilst respecting private, sensitive and personal data, building on existing GDPR legislation.

    d. Transparency – embedding expectations for transparency in the development of models and their reasoning to ensure content produced is explainable. (This is likely to be technically challenging, given the back-box nature of many machine learning systems.)

    e. Multilateral Engagement – working with international governments (US, Canada, China and the EU), regulators, Learned Societies and Professional Bodies to develop international standards and frameworks for LLM development and deployment.

---

[1] "FACT SHEET: Biden-☐Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI." The White House, 21 July 2023, www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/

f. Horizon-scanning – identifying potential opportunities, risks, vulnerabilities and biases through scenario development in collaboration between regulators, policymakers, industry and civil society.

g. Improved Awareness – develop communications and disseminate guidance to a range of stakeholders, from company board level to individual members of the public, to support the effective use of LLMs.

29. These principles should be specifically promoted within high potential sectors, such as telecommunications, health and care or financial services. Sector champions could be appointed to the Foundation Model Taskforce to deliver this. Beyond promoting the principles, sector champions could also serve as a connector between industry and the relevant regulators, thereby improving the flow of knowledge within the innovation ecosystem and ensuring understandings of risk and opportunity so UK sectors can be responsive to changing future trajectories.

30. While the AI White Paper does recognise the importance of international engagement on AI regulation, more convergence in AI-based regulatory approaches with Europe, the United States, and China, will be necessary, as LLM development and deployment scales-up globally. The current draft of the EU's AI Act would ban the use of software that creates an unacceptable risk, defined as covering most uses in predictive policing, emotion recognition and real-time facial recognition.[2] The UK AI Safety Summit should be used to gain buy-in from international governments and industry.

31. There are cross-cutting issues associated with the deployment of widely applicable AI systems (such as LLMs) that are not presently the responsibility of any regulator. Government needs to consider if there is a role for specific regulations that deal with these issues including, factuality, precision, possible harm, bias and explainability. Technical standards also have an important role to play as they provide practical ways to assess the system and promote consistency. They are also not prescriptive, which can enable them to remain responsive and agile.

32. The development of further principles to guide regulators in their approaches to the use of AIs (inclusive of LLMs) within their sector could be beneficial. This is particularly important in the context of ensuring understandings of risk and opportunity are responsive to changing future trajectories. Different sectors do have different needs, and effective governance of LLMs will differ drastically from case to case depending on the context of their deployment. This could be supported by the creation of sector champions within the Foundation Model Taskforce, who support a stronger dialogue between industry and regulators.

---

[2] "EU AI Act: first regulation on artificial intelligence." European Parliament, 14 June 2023, www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

Q3a) What are the implications of open-source models proliferating?

33. One of the key benefits of open-source models proliferating is that they provide easy access to LLMs for a wide community. The general public, and even researchers, do not have access to the code for LLMs like ChatGPT, which can limit confidence and understanding of future trajectories. With an open-source approach, LLMs can foster research and development at significantly larger (and more inclusive) scale than closed-source models.

34. Open-source models operate internationally and their governance is highly influenced by leading technology companies in the United States, as well as those outside dominant western actors in this field such as China and India. It is therefore necessary for the UK to align its approach to governing these models to guard against isolation. However, open-source models are extremely difficult to regulate as they are very fluid, used globally, and embraced by independent developers. If models are introduced and changed at a fast pace, then it will be difficult to exert direct control over the models themselves. While the UK government should explore opportunities to support the development of open-source project governance internationally, managing their use should be prioritised in the short-term.

35. It may also be difficult to control how these models evolve, as open-source models are often developed by disparate and heterogeneous communities of developers. As such, it may be necessary to place responsibility for open-source models with the organisations that make the decision to deploy them (similarly to the way a supplier importing products into the UK or EU assumes responsibility for their import).

36. As open-source models continue to proliferate, there is a risk that divergence in approaches to intellectual property rights (IPR) and the patenting of AI based innovation could limit the accessibility of models. While the UK remains aligned with EU practices in these areas, there is divergence in IPR approaches to software and AI globally. Governments should explore opportunities to conduct international IPR benchmarking exercises to evaluate performance and identify best practice to create a more level and innovation-friendly global ecosystem.

Q4. 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?

37. Many regulators have been supporting innovation through early engagement to try and understand the issues, agreeing on solutions with industry, and enabling joint research. However, there are capacity and capability shortages, and recruitment of those with the necessary technical skills is very difficult as regulators are unable to compete with industry salaries.

38. Many regulators require additional expertise and resources to independently assess the risks and opportunities associated with AIs,

including LLMs, and address them sufficiently. Without this understanding there is a risk that they will not be well-placed to respond to technological developments or previously unforeseen problems.

39. This expertise gap could be partially addressed through the creation of a central body of expertise for regulators to call on, similarly to the National Cyber Security Centre. Such a body could collaborate with regulators to provide specialist knowledge on LLMs and AI more broadly. This would help regulators build on their existing domain expertise thus enabling better assessment of the risks and potential mitigations unique to their sector. The Alan Turing Institute, as the UK's national institute that specialises in data science and AI, would be well-placed to fulfil such a function.

40. There is also a role for continuous professional development of those working in regulation. This will upskill regulators to better understand these technologies, the existing and emerging standards around LLMs, and how to adopt them. There is potential for a body such as the Institute of Regulation to collaborate with members of the National Engineering Policy Centre to provide the technical engineering expertise in the design and development of CPD courses.[3]

41. Greater engagement with a diversity of stakeholders is critical to understanding the extent of possible impacts across all of society. This should include licencing authorities, regulators, platforms, consumer groups and civil society groups, particularly surrounding the ethical implications, in advance of issues arising.

5. What are the non-regulatory and regulatory options to address risks and capitalise on opportunities?

Technical Standards

42. Well-functioning regulatory systems are supported by technical standards that encourage the use of good practice and define the conditions that systems must be tested under. While crafting complex technical standards effectively can be challenging, they are a crucial step towards greater transparency and trustworthiness. Standards can also facilitate international alignment on regulation, promoting international collaboration and knowledge-sharing and access to international markets.

43. There are numerous existing international standards that are not designed solely for LLMs but are, however, applicable to LLMs as a type of AI system. Standards such as IEC 62304 and 82304-1 in the software development cycle or ISO 27001 for cybersecurity are examples of this. Each of the IEC, ISO, BSI, ETSI and ITU are presently engaged in the development standards that can address cross-cutting issues associated with widely applicable AI systems (such as LLMs) will be key to ensuring

[3] National Engineering Policy Centre (2023) Autonomous systems: a workshop on cross-cutting governance

that organisations and governments are able to act with certainty and collaborate with ease.

44. Playing an active role in supporting the development of LLM standards should therefore be a priority for the UK. In doing so, the UK should devote special focus to helping craft standards committees that are diverse in both experience and perspective, and representative of academia and industry of different scales.

Codes of Conduct

45. Sector-specific codes of practice, or conduct, are not legally binding in themselves but support legislation and provide further guidance. Where AI systems have increasing levels of autonomy in areas such as well-defined maritime applications and self-driving vehicles, Codes of Practice have been developed as a flexible way to ensure safety when trialling these systems while the future regulation and legislation develops. These codes can build trust and push a culture change within the profession and should accordingly be considered a powerful means to address risks in the short term. The value of such codes is greatly diminished should they not be in use internationally.[4]

Risk-Based and Proportionate Regulation

46. While nonregulatory mechanisms play an essential role in a well-functioning regulatory system, too much reliance on industry generated codes of practice or standards may create additional risk.

47. Regulatory intervention must be risk-based and proportionate so that innovation is not stifled. Regulators in the UK should therefore seek to build on the well-established ideas of safety cases and ethical assurance cases to continue the development of their understanding of the appropriate risk/benefit trade-offs in deploying LLMs.

48. In doing so, regulators should devote special attention to the through-life value of potential regulation. Models and the environments within which they are deployed are not static, and so regulation must be responsive to their ever-changing profile of risk and opportunity.

September 2023

[4] National Engineering Policy Centre (2020) The journey to an autonomous transport system: identifying challenges across multiple modes