ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

RSS EVIDENCE FOR COMMUNICATIONS AND DIGITAL LORDS SELECT COMMITTEE INQUIRY INTO LARGE LANGUAGE MODELS

5 September 2023

1. Introduction

1.1.1 This is the Royal Statistical Society's response to the Communications and Digital Lords Select Committee's inquiry into Large Language Models (LLMs). The RSS is a professional society for statisticians and data scientists, with over 10,000 members.

1.1.2 There are two areas where we think there is a need for urgent government investment. First, we call for the establishment of aa Centre for AI Evaluation (CAIE). Evaluation in the context of generative AIs (including those that use LLMs) is vital – doing evaluation well will allow us to have a much better understanding of the future trajectory of developments in LLMs. But it is also challenging – technology is moving rapidly and some of the concepts that need to be evaluated, such as fairness, are not clear-cut. Investing a relatively modest £10m in CAIE could have a significant impact and is an area where the UK is well-placed to lead the world.

1.1.3 Second – as the RSS first argued in 2020 – scaling LLMs is an important and pressing challenge facing the UK's AI industry. If the UK is to become a leader in efficient, explainable, scalable LLMs this will investing in a unit of open source software developers and providing funding for increasing computing power and building new data sets. If the UK is to act before large American companies form an oligopoly and capture most of the value from AI we suggest that initial funding of £250m over three years will be required.

1.1.4 We also comment on the recent government white paper on AI. Our four key recommendations are that the government should:

- Establish a new Centre for AI Evaluation (CAIE) with an initial investment of £10m.
- Invest £250m over three years in a unit of open source software developers and increased computing power.
- Strengthen leadership on AI – involving organisations with deep expertise to bring in practical experience, knowledge and capacity.
- Build data science capabilities among regulators.

2 Capabilities and trends

2.1.1 As the Committee will be aware, AI research and development is rapidly evolving and any predictions of how LLMs will develop will be somewhat uncertain. The types of trends that we might expect to see continue are increasing model sizes (which would likely improve AI's performance in language related tasks); more efficient architectures (so that larger models can be handled without needing a correspondingly massive increase in computational power); improved capability across a wider range of modes of output (increasing focus on capabilities for improving generation of audio, images and video). There will also, increasingly, be more specialist models for specific industries such as medicine and finance.

2.1.2 These trends are trends in how the models may develop, but there we also expect to see certain trends and advances in how teams design LLMs into broader systems (products and services in the context of industry). These are likely to be just as impactful, and will be key to the real innovation of putting AI to good use in a wide range of situations. A good example is the development of learned services – services which use AI to allow individuals to draw together different services from a range of organisations.[1]

2.1.3 As the public is becoming more aware of some of the challenges associated with AI, we would also hope to see a focus from developers on understanding and mitigating biases in LLMs to ensure that – as AI is used for a wider range of tasks – it does not amplify existing inequalities. Relatedly, it will be important that developers focus on enhancing the explainability and interpretability of LLMs – for example, by including links to the source information where any key facts used are stated – so that users can better understand the decision-making process.

2.1.4 The focus of our submission, though, is on making the case for investment in two areas where we think there is a pressing need. In this section, we set out a case for the creation of the CAIE: this would have an important role in helping to shape the trajectory of future development of LLMs.

[1] More detail on learned services is available in From chatbots to connections: How AI will change services

2.1.5 At present, we see an insufficient focus on the question of evaluation methodology in the context of LLMs. Evaluation plays a critical role in assessing the performance and capabilities of LLMs, and without improving evaluation methods, we will not gain adequate insight into the strengths, weaknesses, and behaviours of these models. While it is welcome that the AI Foundation Model Task Force will get some priority access to models to help improve evaluation – we believe that this alone does not meet the scale of the challenge and that a devoted centre is warranted. This is important because an ever increasing number of businesses will be looking to use AI systems and it is vital that there is an agreed methodology for evaluating their impact. The problem is especially complex because evaluation has to be proportionate and there are so many different use cases – different types of evaluation are required based on whether the product is intended to advise, recommend, decide, create, entertain etc.

2.1.6 There are similarities between the urgency of the current need for investment in evaluation methodologies in the context of AI as there was in the early 1990s to invest in methodology in the context of health. Then, there were a great many trials that were being funded but there was no clear centre of expertise on methodology for evaluating the evidence in terms of cost/benefit. As a result the NHS R&D Health Technology Assessment Programme was set up in 1993 to ensure that high-quality research information on the costs, effectiveness and broader impact of health technologies was produced in the most efficient way for those who use, manage and provide care. This was a massive success, establishing the UK as a world leader in evaluating health care. Given the UK's track record in healthcare evaluation, there is an opportunity to establish ourselves as genuinely world leading in AI evaluation methodology as well.

2.1.7 To see why this is so important it is helpful to consider some concrete examples. Consider the application of AI in healthcare for medical imaging. AI models, particularly deep learning-based algorithms, have shown significant promise in analysing medical images and assisting healthcare professionals to make accurate diagnoses. Evaluation is important here for many reasons – here we highlight three illustrative examples. First, the accuracy of the AI is crucial – because errors can have serious consequences for patients – and proper evaluation is needed to ensure that the performance is at least at the level of human doctors. Second, the AI must be able to generalise – medical images can vary greatly and evaluation is necessary to check the ability of the AI to generalise well. Third, AIs need to be evaluated to ensure that they are properly balancing sensitivity (the ability to detect true positive cases) with specificity (the ability to avoid false positives) – if the balance between sensitivity and specificity isn't right you risk either missing cases or unnecessarily worrying a large number of patients.

2.1.8 Evaluation is also important when generative AIs, like Chat-GPT, are used to help people produce content for school or university work. If this technology is to assist learning there are a number of things that it must do. First, provenance is important – the AI must be able to detail how the answer was produced and what was used as inputs. Second, it is important to understand the types of sources that were used – did it only refer to real sources (rather than, in the case of hallucinogenic AI, invented sources) and did it respect copyright when accessing information? Evaluation has a key role to play in assessing how generative AIs produce their responses and making sure that they do so accurately and responsibly.

2.1.9 It is important to emphasise that some elements of an AI model will need to be evaluated within the context of a broader system. Different levels of evaluation and explanation are required for different systems depending on the extent to which they are advisory or automated. In the context of the medical imaging example, the performance requirements would be quite different for a system that simply generated a response of "yes", "no" or "maybe" than they would be for an augmented system that provides a recommendation to a medical professional. There is also the question of the baseline that is used to evaluate the performance of an AI – eg, in the context of medical imaging, human doctors will also make mistakes.

2.1.10 Evaluation of generative AIs is difficult, especially when technology is changing at a rapid pace, and when key concepts like fairness is so difficult to tie down.[2] There are also limitations to current evaluation methodologies in the context of generative AIs using LLMs, which prevent us understanding their capabilities and limitations. There are a few particularly pressing problems:

[2] Courtland's (2018) *Bias Detectives: the researchers striving to make algorithms fair* remains a useful discussion of the difficulties in aiming for fairness in AI and algorithmic systems.

- Existing evaluation metrics can overly focus on narrow and task-specific benchmarks, overlooking the models' broader performance across various domains and tasks.
- The lack of standardised evaluation criteria makes it challenging to compare results across different studies – risking inconsistent and potentially misleading conclusions.
- Evaluation datasets frequently fail to represent the complexity and diversity of real-world language usage, resulting in models that perform well on artificial data but struggle to generalise to practical scenarios.
- Evaluation needs to be continuous – model performance changes over time, but a lot of evaluation work is designed for one-off checks pre-deployment. That model is fine for standard algorithms, but is not sufficient when dealing with AIs that change post-deployment.
- Communication of evaluations in a way that is helpful to very different audiences – eg, regulators, teams using LLMs to design products, people using or affected by those products – is challenging but vital if evaluations are to have impact.
- The assessment of biases and fairness in LLMs is an ongoing challenge, with current methods often insufficient to capture subtle biases or to understand the potential legal and ethical implications of the model's outputs.

Addressing these limitations is crucial to ensure that the evaluation of LLMs aligns more closely with real-world requirements and contributes to the responsible and meaningful deployment of these powerful AI systems. The EU plans for regulating AI – that focus on the risk impact of the system and safety based on use rather than by technology – is the right approach and we would like to see this principle applied in the UK.

2.1.11 Statistics has an important role to play in improving the evaluation of AI systems – just as the discipline previously did in the context of healthcare evaluation. New statistical approaches can provide more informative, nuanced, and reliable assessments of models’ performance and behaviour. Statisticians are also expert at uncertainty estimation – offering users more reliable confidence intervals and uncertainty measures for AI predictions. Statistical methods can also systematically quantify biases across different demographic groups, facilitating more robust and fair AI models Statisticians are also expert at communicating the outcome of evaluations appropriately for their audience.

2.1.12 It will be important for the CAIE to develop methodologies through practical work. For example, this might mean offering practical help to public sector teams considering whether and how to deploy AI models effectively. The investment in such a centre could be relatively modest -- £10 million would be enable it to have significant impact – and would allow the UK to establish world-leading practice.

3 Domestic regulation

3.1.1 The RSS is concerned that the AI white paper does not deal especially well with generative AIs – including those using LLMs. On one level this is a question of how the government plans to engage with the professions that interface with AI. Without engaging with external experts with deep expertise, the government's AI bodies are not in a position to comprehensively identify emerging opportunities and risks. This seems to have already happened with LLMs. In 2020, in our engagement with the AI Roadmap, the RSS argued that scaling LLMs was the most pressing challenge facing the UK's AI industry and that an important element of the AI strategy that was missing was investment in the open source software development required to do this. Large AI models require hardware that is out of reach of all but large US and Chinese companies – to be competitive the UK needs to learn how to scale up LLMs in a cost-effective way. It is our view that a significant investment in open source software development – of the order of £250 million over three years – is the best way to bridge this gap.

3.1.2 This would

- Support Government-funded unit of open source developers who would:
  - Contribute to open source software projects of strategic importance to the UK.
  - Provide internships to help build and transfer skills.
  - Support the use of open source in the UK through knowledge sharing, training and building the community.
  - Collaborate internationally with similar projects such as OLMo at the Allen Institute for AI.
  - Disseminate knowledge about LLMs within the UK tech community.
- Introduce a new Fund – modelled on Germany's Sovereign Tech Fund and Prototype Fund – to support open source projects outside government.
- Provide funding to increase compute power and build data sets.

3.1.3 If the government provided £50 million per year for three years that would enable good progress to be made (it requires around £20 million just for the compute power to train a reasonable LLM). Adding an extra £100 million over the three years would enable the UK to use model sizes 5 times larger than is currently possible – which would really push forward the state of the art for open source models.

3.1.4 On the specific matter of regulation, the RSS is concerned that splitting responsibilities for regulating the use of AI between existing regulators does not meet the scale of the challenge – and will also not help businesses in the UK to act effectively in this space. Any regulatory standards, existing or new, need to be rapidly and robustly applied to new entrants and this ideally requires a combination of industry specialist regulatory bodies and a centralised body with technical and practical expertise. At a minimum, we think it is important that the regulators being given new powers should be given more money. The Online Safety Bill led to Ofcom's budget increasing by £50 million – yet, so far, it seems that regulators are expected to make AI safe and easy for business and government with no extra money.

3.1.5 The responsibility for regulating the use and the content of foundation models sits across many government departments and regulators. Given the technical complexity of these models, there is a real concern about capacity, especially since many of the smaller or specialised regulators (CQC, Ofsted, ICO etc) do not employ many data scientists. It is critical that these regulators interact with the appropriate stakeholders – including industrial practitioners and not just academic or public sector data scientists. The RSS could have a role in helping to build capacity – and we would be happy to help in this however possible.

3.1.6 We would also stress that it is important to think about the way that statistical methods are regulated in this context – the issues around the planned use of an algorithm to award pupils A-Level and GCSE grades during Covid was an early warning of some of the issues that we might face. The report produced by the OSR on this topic is a helpful starting point.

3.1.7 Above all, central leadership is required to give a clear, coherent and easily communicable framework that can be applied across sectors. The speed of change within AI means that maintaining a consistent and coherent system across multiple existing regulators is challenging and inevitably some regulators will lag behind others: consistency, coherence and horizon scanning are vital to deal with this.