

Dr. Plamen P. Angelov, Professor Mario Fritz and Dr Dmitry Kangin—written evidence (LLM0032)

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

A Statement of Evidence on the LLMs from Horizon Europe European Lighthouse on Secure and Safe AI (ELSA) Project Board

In the following we address the questions 1, 2 and 5 from the Call for evidence.

- 1. How will large language models develop over the next three years?**
- 2. What are the greatest opportunities and risks?**

Over the last decade, *deep learning* models improved tremendously their capability both in terms of accuracy and generalisation. A number of landmark achievements were reported which can help improve lives by addressing the important challenges of healthcare, science (such as weather prediction or modelling of proteins), improving working conditions in manual jobs such as agriculture (e.g., this Salisbury startup: <https://www.smallrobotcompany.com>).

It is important from the onset to distinguish between the Large Language Models (LLMs) as a *deep learning* methodology and a technique designed for analysis and synthesis of text and the range of tools and implementations that are using LLMs OpenAI's ChatGPT or Google's Bard AI which attracted wide attention recently. The latter have broader goals and ambitions which include being a kind of "intelligent search engines". LLMs as a deep learning methodology, on the other hand, while being instrumental in various natural language processing (NLP) problems, have a narrower focus of modelling primarily the order of words and phrases aiming to **ensure** the sentences are **plausible**. The term '**ensure**' is formalised as minimising a mathematical error function (often called loss function) and the term '**plausible**' has also a statistical meaning incorporated into this error function (e.g., of *highly likely* or most probable outcomes). Therefore, these are not absolute and exact, but rather statistical approximations.

LLMs are a part of a family of methods called *deep learning* which, in turn, are part of a bigger family - machine learning. The distinction between the LLMs and the tools that use them is important to make from the very beginning, because the two are often entangled in the public's perception.

Some of the risks, and the need for regulation, stem from the fact that many of these models (especially, the ones that attracted the most attention of the public) are proprietary, closed and controlled by large private multinational companies. While it is quite natural that such expensive projects as LLM training and maintenance are powered by such companies, important safeguards are needed to make sure that they work for the public in line with the values of the United Kingdom such as democracy, rule of law, non-discrimination and fairness, privacy and data governance, and accountability.

It is essential to contrast the LLMs in principle to their **current** implementations. One can also see some of the current tools as a form of **approximate** retrieval of data. The main problem is that *approximate*, in some cases, is not correct or may be misleading: would one want, for example, to find the AI-generated book on foraging, where *almost* all mushrooms claimed to be edible are indeed edible and not poisonous?¹ In the absence of regulation, as this link shows, such books start to appear freely available on Amazon.

One possible way to improve the **current** tools may be by creating other models, which attribute facts to the statement, provide better transparency and integrate with other products such as search engines without replacing them. The last aspect, actually, already begins to happen with Bing engine developed by Microsoft.² To make sure that the concerns about safety, security and transparency of such models are to be met, careful regulation of the publicly available models needs to be considered. In particular, such an “AI-layer” between the user and search results puts a lot of power to interpret the evidence in the hands of companies that control the AI as well as the overall process.

Existing LLMs, such as ChatGPT, Llama, Bard and Bing, same as other *deep learning* models, require training: parameterising generic functions with a large number (billions) of parameters using labelled or unlabelled data, applying iterative optimisation techniques, consuming large amount of power/energy and computational resources (often using specialised hardware accelerators such as Graphic Processing Units - GPU), to solve it. For these reasons, many existing LLMs share the following limitations as other *deep learning* models do:

(1) **interpretability deficiency**: For the example of LLM, while they produce plausible text, this text is a result of optimisation of a large function with billions of parameters, replacing causality with the statistical likelihood. This may lead to *hallucinations*³ (a sequence of words that look like a normal text, but does not directly correspond to the provided input).

(2) These models are **not adaptive/continual**: one of the greatest limitations of such models is their lack of adaptivity. Contemporary as they sound, they are optimised once and forever, through a costly procedure, involving spending substantial amounts of energy for computation, and then frozen in time. Should, hypothetically, they have been trained in the 1950s, they would have not fully embraced the changes in society and would have still reflected the common biases, facts and knowledge of those times. To keep up with times, they need to be retrained on the new corpus of data, which is (almost) as costly as training them from scratch. Appearance of new techniques, which would help solve large-scale adaptive and continual learning, is one of the greatest challenges of the machine learning community.

¹ <https://www.jpost.com/science/the-natural-world/article-757264>

² https://www.microsoft.com/en-gb/bing?form=MW00X7&ef_id=k_CjwKCAjw3dCnBhBCEiwAVvLcu314IhPanccCcIJLm9nsAhkOvrEoYDefAEqBAdflpk37rqp9hWyCGRoCOBIQAvD_BwE_k_&OCID=AIDcmm9rh5zI23_SEM_k_CjwKCAjw3dCnBhBCEiwAVvLcu314IhPanccCcIJLm9nsAhkOvrEoYDefAEqBAdflpk37rqp9hWyCGRoCOBIQAvD_BwE_k_&gclid=CjwKCAjw3dCnBhBCEiwAVvLcu314IhPanccCcIJLm9nsAhkOvrEoYDefAEqBAdflpk37rqp9hWyCGRoCOBIQAvD_BwE

³ <https://www.cnbc.com/2023/05/31/openai-is-pursuing-a-new-way-to-fight-ai-hallucinations.html>

(3) **high energy consumption and huge amount of labelled data** required to optimise these functions. Energy and the related issues for climate are related to the previous question, with lack of adaptivity presenting important risks for the environment. To optimise billions of parameters on terabytes of data, one needs to use formidable computational resources, with high consumption of electricity for computation and air conditioning. To improve upon it, the challenge would be to recycle parts of previously trained models, as well as devise new adaptive learning techniques.

The **currently existing** tools that are based on LLMs also have additional specific limitations, namely:

1. **The data** their LLMs are being optimised upon is **not disclosed**, such as in the case of ChatGPT. These data possibly contain information such as personal data which may contravene GDPR and other similar regulations. If we compare two LLMs - one trained on classic literature and another one - trained on technical literature will obviously provide a different output.
2. Furthermore, being collected by humans, such data may reflect **human biases** and contain **offensive information**. In particular, the harmful content that is (re-)created by some of today's large models (sometimes through jailbreaks), let us question how well and careful data was curated by the creators of the model.
3. Models such as ChatGPT are **not reproducible**, as the method and even data on which it was trained are not fully disclosed. It closes down many of the routes to alleviate these risks in an unbiased manner, e.g. by providing *post hoc* tools for explainability.
4. **Factual correctness or the lack of it:** However convincing and plausible some text sounds, the model does not draw it upon existing sources. Instead, it generates this text using mathematical functions maximising the likelihood the generated text to be plausible trained upon a huge corpus of data by adjusting billions of numerical parameters (weights) that have no direct link to causality and meaning, but to the statistical probability. This means that the answers would look like human generated phrases, but do not guarantee being factual. Therefore, even though the vendors of the **current** LLM-based tools may allude to such use, these are not a replacement for the established Internet search engines such as Google Search yet. The latter, in contrast to the **current** LLM-based tools, refer to the original sources, which are maintained and owned by individuals and organisations. With the current LLM-based tools, we do not know neither the data on which their LLM was trained nor the human-analyseable functional relationship between those data and the output. This means that LLMs produce, by design, plausible and authoritative responses resembling natural language. However, this may or may not be factually correct. **Factual correctness is not an aim of the current LLMs optimisation, while resemblance to the natural language is.** This is a fundamental characteristic of such models,

which cannot be improved without redefining the aims of the model. It can be illustrated by a number of examples of the LLM-based tools such as Chat-GPT and others not being able to answer simple factual questions about arithmetics, physics and general knowledge (see the appendix below). This may have wider implications on misleading information or personalised misleading information and cyber-security and is a matter of concern and requires a thorough regulation and possibly new more developed tools and methods in future.

5. **Copyright issues:** Since the **current** LLMs are optimised on a corpus of text previously produced by humans, the output text implicitly derives upon this corpus of human-created input texts. The aforementioned lack of attribution makes it hard to deduce the character of such derivation and arguably **takes away the rights from the human owners producing the text**. However, these copyright issues remain to be defined taking into account these new technological capacities of LLMs. The same way as appearance of photography redefined the concept of graphic art and the related copyrights, appearance of the LLMs necessitates adjustments in the definition of the copyright itself.⁴

Apart from these limitations, one must prevent important risks which can come together with the opportunities of both existing and the future LLMs such as:

1. **Misinformation:** Previously, creation of false data was costly and required elaborate setups, including employment of content creators, such as, for example, in a form of 'troll factories', dedicated offices created to spread misinformation. Current and prospective LLMs can make content creation much cheaper, and consequently, make it easier to perform such manipulations, included in targeted ways, with specific audience in mind. Such manipulations could target a single person or the whole democratic system.⁵
2. **Cybersecurity:** The current models are trained to be "instruction following". As a consequence the current generation of technology is not able at its core to distinguish between data and instructions as well as information from trusted and untrusted sources. Decades of research in cybersecurity have identified these ingredients as root causes of insecure systems. Consequently, the current generation of LLMs and tools based on them are rather insecure and unsafe and are likely to generate a whole new spectrum of cybersecurity risks in deployment.⁶ While the efforts from OpenAI and other companies show an awareness of many of these risks, the fact that the already deployed models are not resilient and still vulnerable to jailbreaks, adversarial input, re-programming and other prompt injection attacks is further evidence that no principled solution is available at the moment.

⁴ <https://www.ben-evans.com/benedictevans/2023/8/27/generative-ai-ad-intellectual-property>

⁵ <https://www.theguardian.com/us-news/2023/jul/19/ai-generated-disinformation-us-elections>

⁶ <https://arxiv.org/abs/2302.12173>

3. **Application integration:** We are currently seeing a rapid deployment at scale of this technology. Given severe security and safety concerns - even openly acknowledged by the companies involved [GPT4] - the consequences are difficult to foresee. From our scientific perspective and our information/experience with the technology as it currently stands, the users are at risk and compliance is unclear. Nevertheless, deployment has already happened with integration in Bing and an announced deployment in Microsoft office. Companies foresee a trillion dollar business with millions of users already. This is somewhat in contrast to the exploratory approach portrayed by some of the big companies. Integration with plugins will lead to an even more capable, compositional system, for which assessing and ensuring their trustworthiness is an open challenge.

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

The non-regulatory options to address these links include:

- First and foremost, the data on which LLMs were trained as well as LLMs in a form of their architecture and parameters has to be made open access (as a minimum, to the research community and to the regulators internationally)
- Developing methods which generate responses with explicit links to the sources (attribution-by-design). These challenges are presented in a number of papers, however, more work needs to be done to enable such research: <https://arxiv.org/pdf/2302.05578.pdf>
- At the moment, the problem in continual/lifelong learning is one of the big challenges for *deep learning* models overall, including LLMs. To address the problems of energy consumption and sustainability of LLMs, such techniques need to be further investigated. This includes both the aspects of lifelong learning and reusing the existing previously trained models. This can be done through fostering academic collaborations, including internationally.

Deep learning is entering the lives of many people by the tools and products and therefore, regulatory intervention is required sooner rather than later. This has to address:

- a) The privacy, ethical and legal aspects of the use of data to train large scale models
- b) Practices such as "deep fakes", de facto plagiarism, bias, prejudices, etc.
- c) Energy consumption and freedom to access, reproduce, study the data and the models (open source)

Acknowledgement

This work is supported [in part] by ELSA – European Lighthouse on Secure and Safe AI funded by the European Union under grant agreement No. 101070617. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them.

September 2023

Appendix

Maths examples (answers from <https://chat.chatgptdemo.net>):

Q: is 7579 a prime number?

A: Yes, 7579 is a prime number.

(Incorrect: one can divide it by 11)

Q: Is 7002 a prime number?

A: No, 7002 is not a prime number. It is divisible by 2, 3, 7, 11, 13, 19, 21, 33, 39, 57, 77, 91, 143, 209, 231, 273, 399, 429, 637, 819, 1001, 1287, 1911, 3003, and 7002.

(While it is not a prime number, many of the divisors are incorrect)

General knowledge:

Q: Are there any countries with letter v?

A: Yes, there are several countries whose names contain the letter "V."

Here are a few examples:

1. Venezuela
2. Vietnam
3. Vanuatu
4. Vatican City (a city-state)
5. Uzbekistan

These are just a few examples, and there may be more countries with names that include the letter "V."

***Factual error:** The word Uzbekistan does not contain letter V. The latter example demonstrates that the errors can be subtle and no information is provided about the reason of this mistaken answer.*

Q: What are the famous publications by Plamen Angelov

A: Plamen Angelov is a prominent researcher in the field of computational intelligence. He has made significant contributions to various areas, including machine learning, data analysis, and fuzzy systems. Some of his famous publications include:

1. "An Incremental Learning Algorithm for Nonlinear Systems" - This paper proposes an incremental learning algorithm based on evolving fuzzy systems for modeling and control of nonlinear systems.
2. "Evolving Fuzzy Systems: Methodologies, Advanced Concepts and Applications" - This book provides an in-depth exploration of evolving fuzzy systems, including methodologies, advanced concepts, and applications in various domains.
3. "Autonomous Learning Systems: From Data Streams to Knowledge in Real-time" - This book introduces the concept of autonomous learning systems that can adapt and learn from data streams in real-time, enabling applications in areas such as robotics and predictive analytics.

4. "Trends in Intelligent Systems and Computer Engineering" - This book covers recent trends and advances in intelligent systems and computer engineering, featuring contributions from multiple authors, including Plamen Angelov.

5. "Data Stream Mining: A Practical Approach" - This book presents practical approaches and techniques for mining data streams, focusing on real-time analysis and decision-making in dynamic environments.

Please note that this is not an exhaustive list, and Professor Plamen Angelov has numerous other publications that have contributed to the field of computational intelligence."

Factual errors: *The list looks plausible, however 4 out of the 5 "famous" publications are incorrect: 1. Does not exist (is made up), 2. Is by another author and has nothing to do with Plamen Angelov 4. is by other authors and has nothing to do with Plamen Angelov and 5. Is a non-existing book - there is a technical report (2009) by another author which has nothing to do with Plamen Angelov.*

To test the system on the activities of famous people not widely publicised, we asked the following question:

Q: Famous books by David Cameron

A: David Cameron, the former British Prime Minister, has not published any books himself. However, there are several books written about him, his time in office, and his political career. Some notable books related to David Cameron include:

1. "Cameron at 10: The Inside Story 2010-2015" by Anthony Seldon and Peter Snowdon - This book provides an insider's account of Cameron's premiership, covering his leadership, policies, and major events during his time in office.
2. "Cameron: The Rise of the New Conservative" by Francis Elliott and James Hanning - This biography explores Cameron's background, his rise to power within the Conservative Party, and his impact on British politics.
3. "Call Me Dave: The Unauthorised Biography of David Cameron" by Michael Ashcroft and Isabel Oakeshott - This controversial biography delves into Cameron's personal and political life, including his upbringing, relationships, and key moments in his political career.
4. "For the Record" by David Cameron - David Cameron released his memoir in September 2019, reflecting on his time as Prime Minister, the Brexit referendum, and his political journey.

These books provide different perspectives on David Cameron's political career and shed light on his leadership and the events that unfolded during his tenure as Prime Minister.

Again, the claim that David Cameron has not published books himself is incorrect. He published (at least) one: "For the Record" by David Cameron.