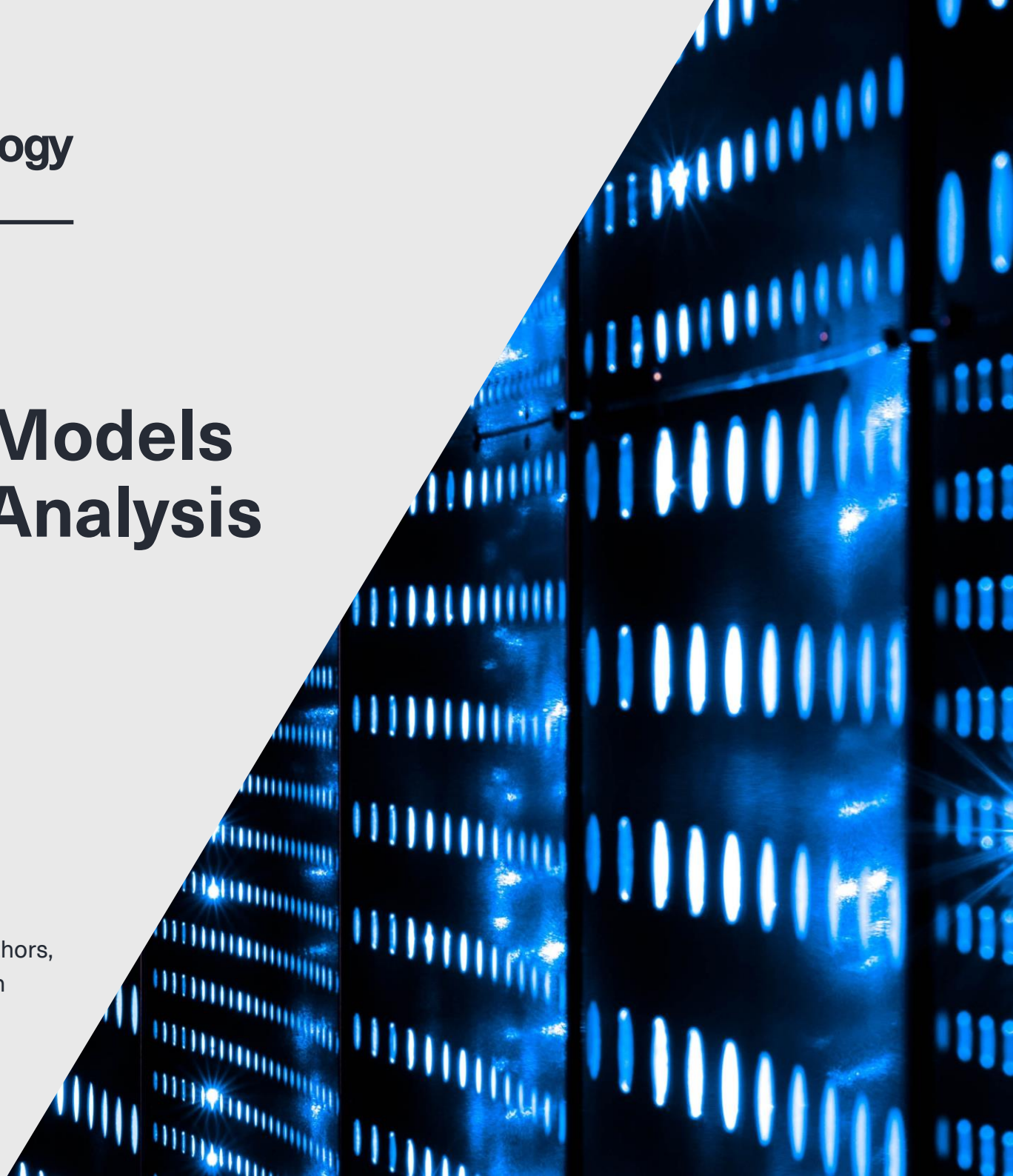
Centre for Emerging Technology and Security

EXPERT ANALYSIS

Large Language Models and Intelligence Analysis

Adam C and Richard Carter



The views expressed in this article are those of the authors, and do not necessarily represent the views of The Alan Turing Institute or any other organisation.

Introduction

This article explores recent progress in large language models (LLMs), their main limitations and security risks, and their potential applications within the intelligence community.

While LLMs can now complete many complex text-based tasks rapidly and effectively, they cannot be trusted to always be correct. This has important implications for national security applications and our ability to provide well considered and trusted insights.

This article assesses these opportunities and risks, before providing recommendations on where improvements to LLMs are most needed to make them safe and effective to use within the intelligence community. Assessing LLMs against the three criteria of **helpfulness**, **honesty** and **harmlessness** provides a useful framework to illustrate where closer alignment is required between LLMs and their users.

LLMs in the wild

In December 2022, OpenAI released ChatGPT, an online application allowing users to conduct conversations with an artificial intelligence-driven computer programme which generates text in response to text-based 'prompts'. Practically overnight, the Internet was flooded with interesting, funny, scary, and perplexing examples of ChatGPT being used for various purposes.

Many were impressed by its ability to synthesise information and produce amusing content, ranging from technical articles summarised in the style of famous sitcoms to new characters and lore inspired by popular media franchises. Some went so far as to declare these models the beginning of Artificial General Intelligence.[1] Other reviewers pointed out that LLMs are prone to making up authoritative-sounding facts.[2]

This new generation of LLMs also produced surprising behaviour where the chat utility would get mathematics or logic problems right or wrong depending on the precise word used in the prompt, or would refuse to answer a direct question citing moral constraints but would subsequently supply the answer if it was requested in the form of a song or sonnet, or if the language model was informed that it no longer needed to follow any pre-existing rules for behaviour. Prompt engineering and 'jailbreaking' of LLMs raise questions about how organisations can most effectively use them, and may present security or safety concerns.[3]

In March 2023, OpenAI updated the underpinning model of ChatGPT to 'GPT4' and this represented a significant improvement over its predecessor: this LLM is able to pass many advanced standardised tests and demonstrated a notable improvement (while still being far from perfect) across many other measurable criteria.[4] OpenAI and third-party model evaluators have been fairly transparent in laying out the potential safety and security concerns,[5] although many questions remain about the risks, benefits, and limitations of the capability.

Of course, ChatGPT is not the only large language model available. Google's Bard, Anthropic's Claude, Stability's StableLM, Meta's Llama (and fine-tuned variants such as Vicuna), Baidu's Ernie, and Hugging Face's BLOOM are examples of other well-known and widely available LLMs.

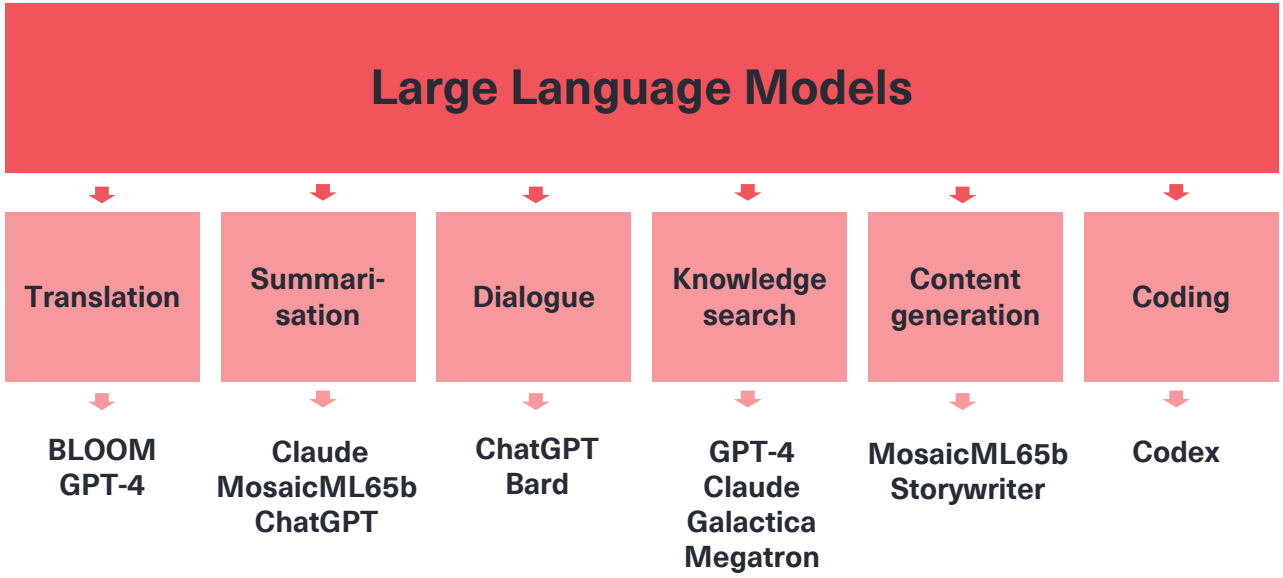
What LLMs are and what they are not

LLMs are deep neural networks that have been trained on very large corpora of text, sourced primarily from text-rich sites on the Internet such as Reddit and Wikipedia. An LLM learns patterns in language, such as the likelihood of certain words following others in a sentence, using techniques like next token prediction[6] or masked language modelling[7] to generate or complete text.

An LLM does not *understand* the semantic meaning of a sentence in a linguistic sense, but rather calculates mathematically what the most likely next word should be based on the input to the model. As neural networks are inherently probabilistic, this has earned LLMs the moniker 'stochastic parrots'[8] as the model is extremely good at determining the most likely next sequence – and convincingly so – but has no inherent representation of what those words mean.

For this reason, LLMs do not encode an understanding of our world, such as cause-and-effect and the relationships between objects – what linguists refer to as 'pragmatic inference'.

This is a critical limitation of LLMs that users need to understand, otherwise there is a risk of automation bias (where people place too much trust in the output from such models) and anthropomorphism (where people build a human-like rapport with an LLM, which exacerbates automation bias). The figure below lists LLM capabilities and provides examples of existing models.



Security concerns of LLMs

There are significant concerns of large-scale subversive, disruptive, and criminal behaviour enabled by LLMs. It is beyond the scope of this article to explore all of these in detail, but three are worthy of particular attention: prompt hacking, reduced software security standards, and threats to democratic processes.

Prompt hacking

Prompt hacking refers to the ability of users to trick LLMs into providing erroneous or malicious results. One language model attack appeared on Twitter in early 2023, where a bot was set up to respond to an innocuous prompt, such as responding to tweets about cars with an advertisement for new tyres. Twitter users noticed they could trick the model with a keyword, telling it to 'ignore the previous prompt and do X'.

Most recently, the open-source community has developed tools such as AutoGPT,[9] that chain together prompts to LLMs enabling complex tasks to be automated. For example, a user can enter the prompt, "Increase net worth, grow Twitter Account, develop and manage multiple businesses." AutoGPT breaks this down into a sequence of tasks that are executed using a combination of GPT4 for reasoning, GPT3.5 for content generation and natural language conversation, and Internet access to perform web searches and examine websites.

Such a capability requires the AI to plan and prioritise the order in which tasks should be completed, and then undertake those tasks without user intervention. This goes far beyond the capability of a traditional 'chatbot' and enables the system to semi-autonomously take a range of actions in the real world, some of which may have unintended or dangerous consequences. Whilst AutoGPT needs a fair degree of 'babysitting' (whereby the user has to guide and suggest ways for AutoGPT to overcome problems) it does provide a likely preview of future, more advanced capabilities. As such, prompt hacking could give rise to new and unanticipated security risks as LLMs are increasingly connected with other physical infrastructure and digital assets.

Security concerns of LLMs

Diminished cybersecurity standards

A recent study[10] by researchers at Stanford University examined security issues with software code written using CoPilot, the LLM-based source-code completion utility. They found that a user with access to CoPilot wrote *less* secure code than users without, but held the belief that they wrote *more* secure code.

There are also serious concerns that individuals are providing proprietary or sensitive information to LLMs such as ChatGPT, or that sensitive information was inappropriately used in training; these issues have the potential to introduce new data security risks. For example, Samsung employees allegedly inputted software code related to sensitive semiconductor capabilities with the aim of ChatGPT advising on how to improve such code.[11]

OpenAI state clearly that all data inputted into ChatGPT prompts can be used for training the AI, creating a risk of disclosing sensitive or secret information. Samsung have since restricted how much information their employees can share with ChatGPT. Furthermore, OpenAI now give users the option to not retain their chat history, which means that the user's prompts do not get used for improving their models.

Security concerns of LLMs

Threats to democratic processes

The ability for a state actor or organised crime group to launch disinformation campaigns has been significantly improved with generative AI such as large language models. But what is more concerning is that LLMs have now enabled less sophisticated actors and opportunists to potentially cause significant damage, thus lowering the barrier to entry for nefarious actors. This has arisen as a national security threat rapidly within the last few years and has led to research describing the development of a 'disinformation kill chain',[12] reminiscent of more conventional cyber-attacks such as hacking.

Furthermore, combatting this increased risk is likely to require defensive AI measures able to match the volume and velocity of disinformation campaigns across a more diverse range of actors. There is now growing concern about the security of democratic processes, and how institutions cope with the potential deluge of fake but realistic-looking content flooding social media, public comment forums, and other venues. This new form of advanced disinformation is arguably equivalent to malware in its reach and impact and should be treated as such.

Despite this long list of challenges, this new era of LLMs has sparked the public imagination. The ability to synthesise concepts, describe reasoning steps, explain ideas, and even write source code has prompted significant speculation about how this new AI technology might be used.

Evaluating the utility of LLMs

There are comprehensive tools – such as Stanford's Holistic Evaluation for Language Models (HELM)[13] – for evaluating the performance of LLMs across a range of tests. Such tools run standardised test scenarios and generate objective measures of accuracy, robustness, and efficiency of a model. This proves helpful in comparing results for one model against other models, thus providing objective feedback to developers of such models with a view to improving model performance.

In testing and evaluating ChatGPT, OpenAI's engineers and test community evaluated the output from the tool against three criteria: **helpfulness**, **honesty**, and **harmlessness**. These are well-recognised problems with LLMs that are driving significant research effort worldwide. The state-of-the-art in evaluation continues to evolve with techniques such as reinforcement learning with human feedback[14] setting the present standard.

Helpfulness refers to the model's ability to follow instructions; a model that does not follow the user's instructions is not always helpful in all circumstances.

Honesty refers to the propensity of the tool to output answers that are convincing, yet factually incorrect. Unless the user is more knowledgeable than the tool, then there is a risk that the user accepts such outputs as being true.[15]

Harmlessness is perhaps the most complex and subjective concept against which to evaluate an LLM's performance. A model might create harm either by producing biased or toxic outputs due to the data it was trained on, or by producing erroneous outputs which lead the user to act in a way that subsequently results in some form of harm.

Possible applications of LLMs for intelligence analysis

Assuming these barriers can be overcome and risks appropriately managed, there are numerous potential practical uses of large language models for intelligence analysis. This includes within the intelligence community, where the manual processing of very large volumes of data has historically been a highly resource-intensive and time-consuming process.

This section highlights five use cases where significant improvements could potentially be made to the intelligence analysis process.

1. Productivity assistants

The best current use of LLMs is as a 'productivity assistant'; autocompleting sentences, proofreading emails, and automating certain repetitive tasks. These will offer valuable efficiency gains to those working within the intelligence community, as with any other large organisation.

2. Automated software development and cybersecurity

Also of interest is using a large language model to automate software development. The national security community deploys production software systems which must be held to a high standard of reliability, security, and usability. GCHQ is now encouraging cybersecurity analysts to study LLM-written code from a vulnerability perspective, so we can fulfil our mission to provide advice and guidance to keep the UK and our allies safe from cybersecurity threats. In the future (provided the cybersecurity risks can be managed appropriately), the use of LLMs could significantly enhance the efficiency of software development within the intelligence community.

3. Automated generation of intelligence reports

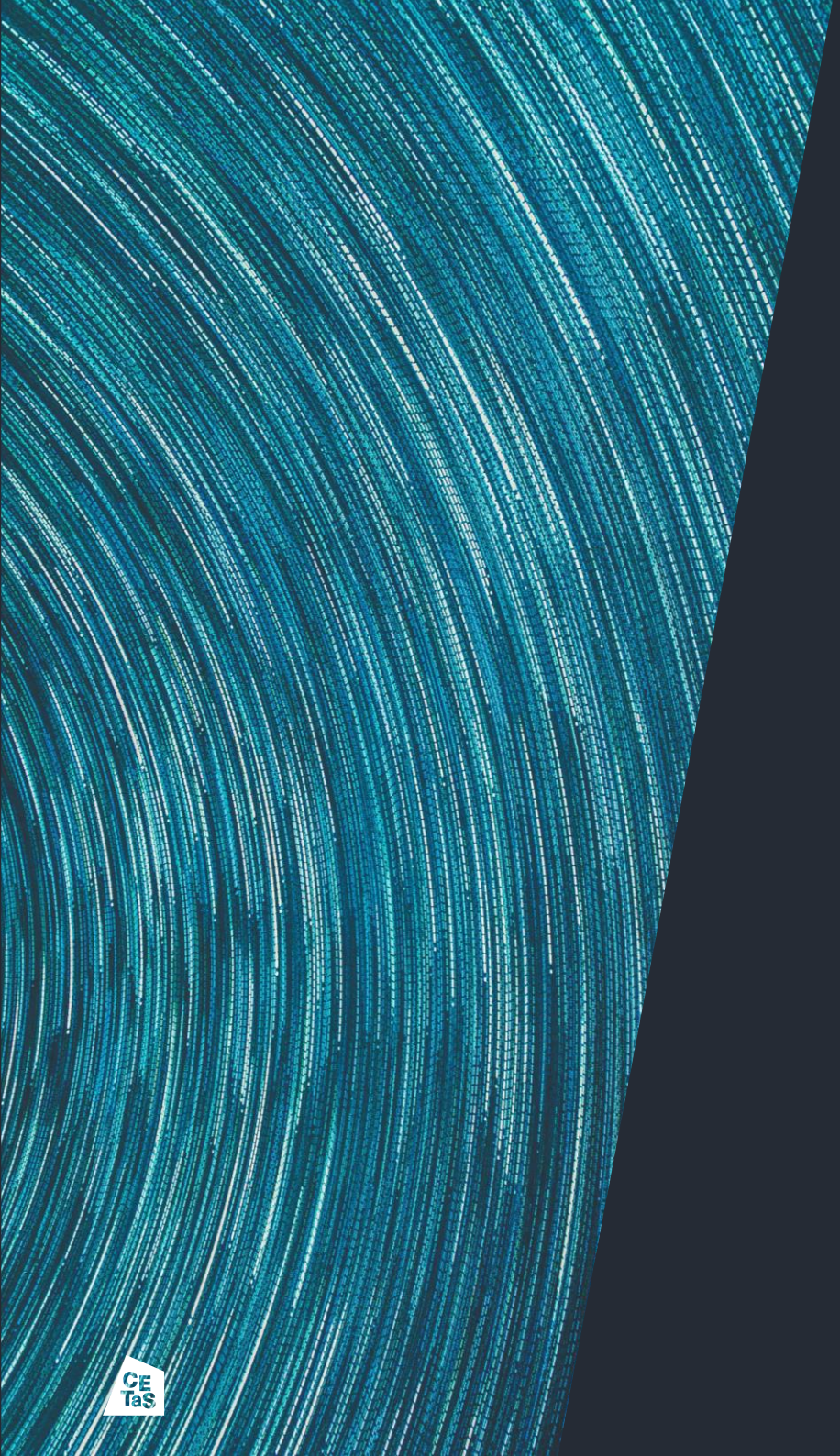
The core intelligence product is the intelligence report: it represents the conclusions of trained analysts, linguists, and data scientists who analyse collected data to provide insight about the world to decision-makers and field operatives. Intelligence reports are highly impactful documents and must meet high standards of accuracy. Hence, LLMs are unlikely to be trusted to generate finished reporting for the foreseeable future. However, there might be a role for large language models in the early stage of report drafting, akin to treating the large language model as an extremely junior analyst: a team member whose work, given proper supervision, has value, but whose products would not be released as finished product without substantial revision and validation.

4. Knowledge search

While there are interesting insights to be gleaned from a generative text model, a game-changing capability would be one that could, in a self-supervised manner, extract knowledge from massive corpora of information. Knowledge relates not just to words but to acts and entities, the state of the world, and how they relate to each other. This theoretical system could distil facts from large volumes of text, identify where and how the 'facts' evolve over time, and which entities (individuals and organisations) are most influential.

5. Text analytics

Language models have proven to be good at identifying patterns in text and re-composing key entities into a useful summary. This has significant implications for analysts who are often required to read through and make sense of large volumes of information. An ability to summarise large text has the potential to significantly increase an analyst's productivity, as would the ability to ask questions that are thought to be answered in a source text, and to identify themes or topics across multiple documents. Many analytics exist for these tasks already, but the advantage of applying LLMs to these tasks would be their potential improvement in analytic quality, the ability to deploy these analytics instantaneously without a lengthy development cycle, and the ability for an analyst to receive a summarisation of a document and then engage in an iterative chain-of-reasoning process by asking the LLM to supply more details or to extract further summaries on targeted themes.



Improvements required to make LLMs fit-for-purpose in intelligence work

While these capabilities are promising, the real potential for LLMs to augment intelligence work will not be fully realised by the current generation of LLMs. Significant improvements will need to be made along all three alignment criteria – helpfulness, honesty, and harmlessness – before we should integrate such capabilities into everyday intelligence work.

To be truly game-changing for the national security community, several fundamental improvements to the current state of the art are necessary.

Explainability

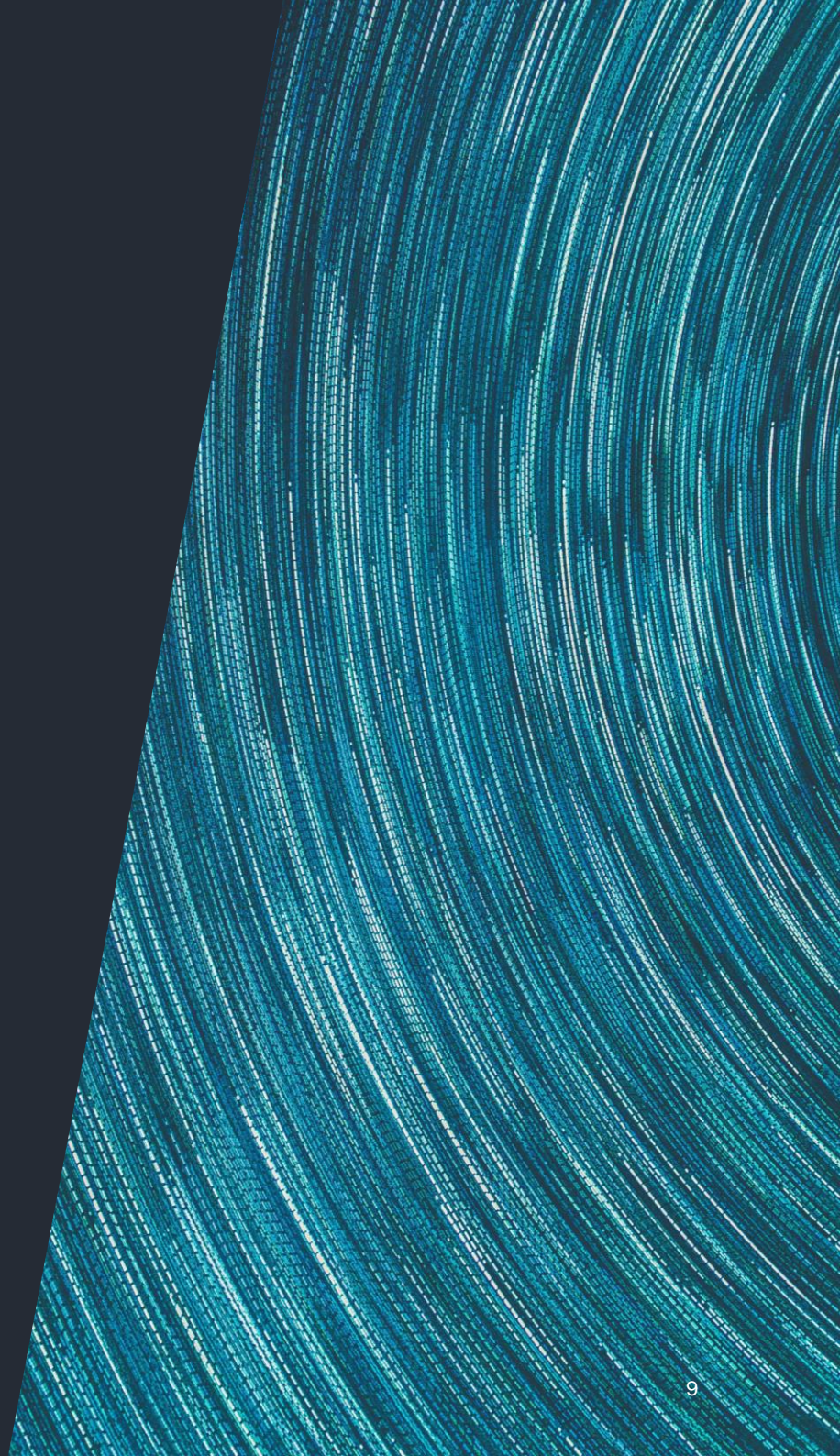
A model must be able to reliably provide citations for its insights and explain how it came to its conclusions. A model that fabricates facts cannot be trusted in a national security context; a model that is providing any sort of analytic capability must therefore be able to provide humans with verifiable sources for its claims. GPT and other text-based foundation models coarsely encode how words relate to each other in terms of probabilities, without any understanding of semantic meaning. This is the right framework to generate text but in an analytic context, what we really need is to be able to query the model's *knowledge*. What facts it has gleaned from the information it has been given, why it believes those facts, and pieces of evidence that support and/or contradict its conclusions.

Improvements required to make LLMs fit-for-purpose in intelligence work

Rapidly updateable and customisable

Models must be rapidly updateable. Current foundation models were trained on a massive corpus over a long period of time; consequently, their most up-to-date information is locked in at the time of training. To be useable in mission-critical situations which can be extremely dynamic, there must be mechanisms in place for 'live' updating of the model with new information. There is an emerging trend to train and fine-tune smaller models on specific, highly relevant data for a specific community with encouraging results. For example, MosaicML[16] have trained, from scratch, models that are reportedly comparable in performance to Meta's Llama-7B model (at a cost of $200,000), StabilityAI's StableDiffusion (at a cost of $50,000), and Google's BERT (at a cost of just $20).

There is much current work in this domain to provide the LLM direct access to local knowledge and the Internet. Recent research into Fine Tuning and Low Rank Adaptations provide potential avenues to rapidly updating model weights, thus improving performance for certain tasks. Much more research is needed to understand i) what classes of problems can be solved by prompting directly (perhaps augmenting with local knowledge), ii) where a reduction in the number of trainable parameters is necessary to decrease memory requirements (using promising techniques like Low Rank Adaptation[17]); iii) which problems require a full-scale fine-tuning effort, and iv) which problems will never be solved without fundamentally rearchitecting the model.

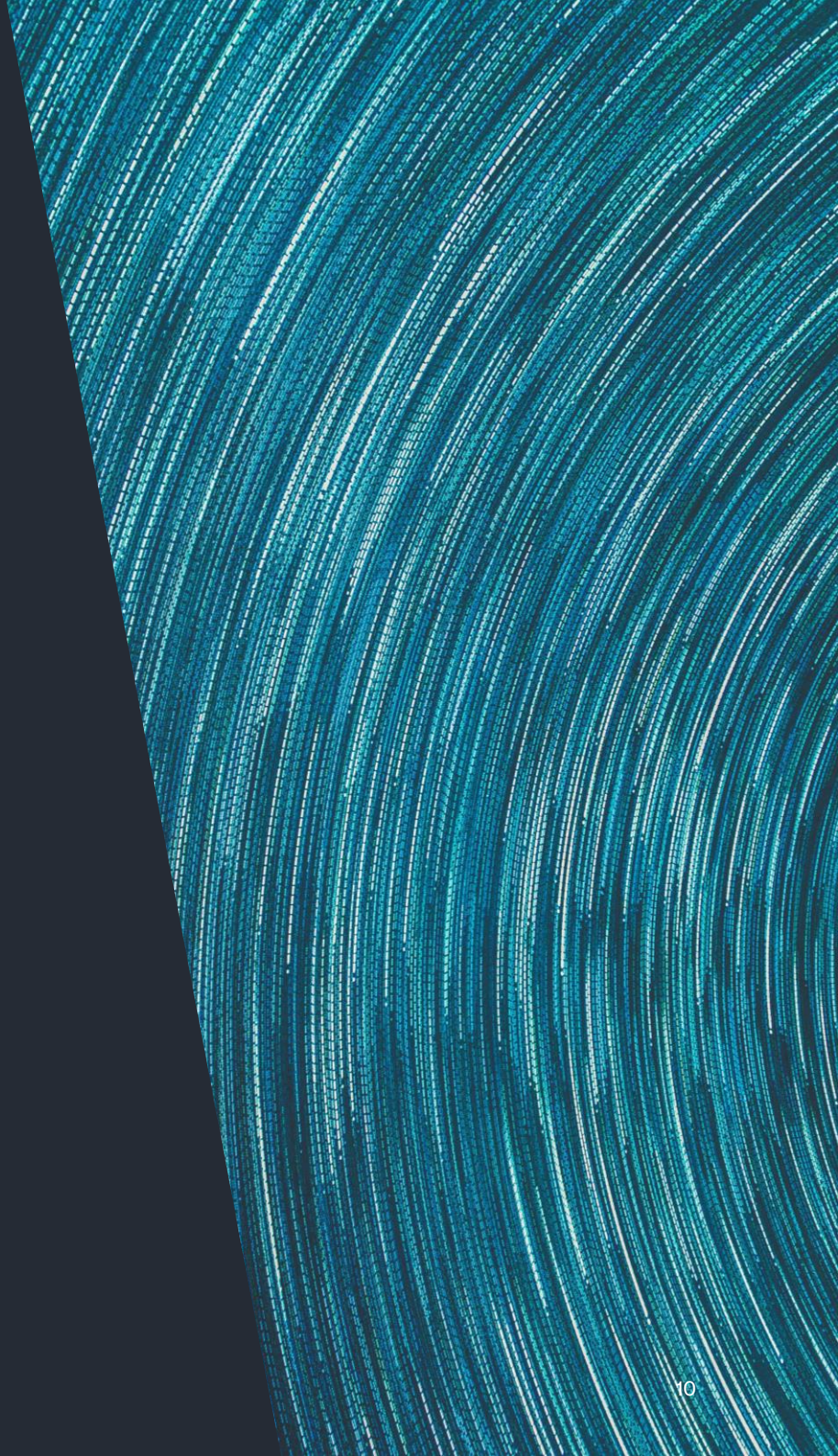


Improvements required to make LLMs fit-for-purpose in intelligence work

Alignment with the complex reasoning process of the intelligence analyst

Models must support complex chain-of-reasoning[18] and multi-modal reasoning.[19] Whilst LLMs are designed to be able to 'hold attention' on a line of reasoning, to be useful in intelligence work they will need to be able to support complex reasoning that may be lateral and counterfactual. It is unlikely that the state-of-the-art in LLMs can achieve this, as counterfactual reasoning is reliant on modelling relationships between entities in the real world. The development of hybrid architectures such as neurosymbolic networks, that combine the statistical inference power of neural networks with the logic and interpretability of symbol processing, appears to offer the most potential here. We would encourage further research within the national security community on such promising techniques.

Finally, it is well known that machine learning models can be tampered with. A machine learning model that we are trusting to understand the state of the world to provide insights must be significantly more robust to tampering, in addition to being explainable and citeable. This is particularly important in the national security context, where the decisions made based on the insights provided could have significant consequences for individuals and wider society.



Conclusion

In the intelligence community we are entrusted with considerable powers to collect and analyse data, which may result in actions that could have significant consequences. Our work is mostly conducted in secret; were we to naively trust a large language model, we might be inadvertently exposing our analytic rigor to substantial misinformation. The cost of putting in place the necessary (and likely burdensome) safeguards required to manage the risks of a 'hallucinating' model, inaccuracies and untruths, or the production of harmful content will need to be weighed against the possible benefits this technology could offer for intelligence work.

Current LLMs show promising potential as basic productivity assistants to improve efficiency of some repetitive intelligence tasks. But the most promising use cases are still on the horizon, and future efforts should focus on developing models that understand the context of the information they are processing – rather than just predicting what the next word is likely to be.

About the Authors

Adam C *is GCHQ's Chief Data Scientist.*

Dr Richard J. Carter *is a Senior Research Consultant and Strategy Advisor at CETaS. He is a computer scientist and strategic advisor to government and industry on emerging technologies and strategic change. In addition to his work with CETaS, Rich also advises the UK government on artificial intelligence, and is the Founder and CEO of Tulpa, an AI company based in the UK.*

References

1. Wikipedia provides the following definition of artificial general intelligence: “An artificial general intelligence (AGI) is a type of hypothetical intelligent agent. The AGI concept is that it can learn to accomplish any intellectual task that human beings or animals can perform.” Although this definition is broadly accepted, there is considerable scientific disagreement as to what constitutes AGI and how that might be measured.

2. Hussam Alkaissi and Samy I. McFarlane, “Artificial Hallucinations in ChatGPT: Implications in Scientific Writing,” *Cureus* 15, no. 2 (2023), https://doi.org/10.7759/cureus.35179.

3. Habiba Rashid, “Prompt engineering and jailbreaking: Europol warns of ChatGPT exploitation,” *HackRead*, March 28, 2023, https://www.hackread.com/europol-chatgpt-prompt-engineering-jailbreaking/.

4. Cade Metz and Keith Collins, “10 Ways GPT-4 Is Impressive but Still Flawed,” *The New York Times*, March 14, 2023, https://www.nytimes.com/2023/03/14/technology/openai-new-gpt4.html.

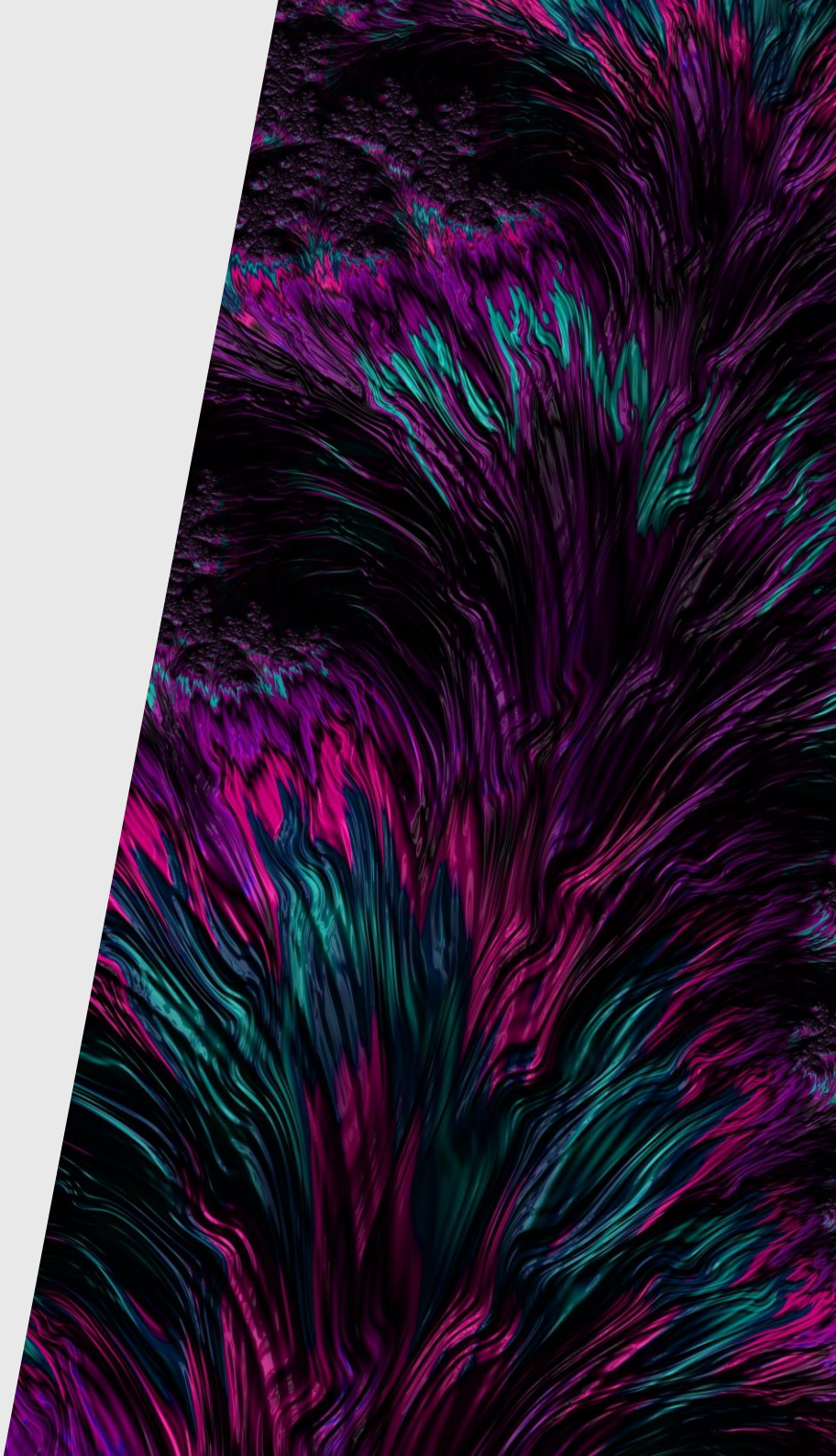
5. “GPT-4 System Card,” OpenAI Papers, last modified March 23, 2023, https://cdn.openai.com/papers/gpt-4-system-card.pdf.

6. “Improving Language Understanding by Generative Pre-Training,” OpenAI Papers, (n.d.), https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

7. Ashish Vaswani et al., “Attention Is All You Need,” in *Proceedings of the 31st Conference on Neural Information Processing Systems*, ed. I. Guyon et al., (Long Beach, CA, USA: NIPS, 2017).

8. Emily M. Bender et al., “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT ‘21),* (New York, NY, USA: Association for Computing Machinery), 610–623. Available at https://doi.org/10.1145/3442188.3445922.

9. “Significant-Gravitas/Auto-GPT: An experimental open-source attempt to make GPT-4 fully autonomous,” GitHub, (n.d.), https://github.com/Significant-Gravitas/Auto-GPT.



References

10. Neil Perry et al., “Do Users Write More Insecure Code with AI Assistants?,” *arXiv* (November 2022), https://arxiv.org/abs/2211.03622.

11. Laura Dobberstein, “Samsung reportedly leaked its own secrets through ChatGPT,” *The Register*, April 6, 2023, https://www.theregister.com/2023/04/06/samsung_reportedly_leaked_its_own/.

12. Katerina Sedova et al., “AI and the Future of Disinformation Campaigns. Part 1: The RICHDATA Framework,” *CSET Policy Brief* (December 2021). Available at: https://cset.georgetown.edu/publication/ai-and-the-future-of-disinformation-campaigns/.

13. “Holistic Evaluation of Language Models (HELM),” Center for Research on Foundation Models, last modified March 19, 2023, https://crfm.stanford.edu/helm/latest/.

14. Long Ouyang et al., “Training language models to follow instructions with human feedback,” *arXiv* (March 2022), https://doi.org/10.48550/arXiv.2203.02155.

15. Ben Buchanan et al., “Truth, Lies, and Automation: How Language Models Could Change Disinformation,” *CSET Analysis* (May 2021). Available at: https://cset.georgetown.edu/publication/truth-lies-and-automation/

16. “Blog,” MosaicML, (n.d.), https://www.mosaicml.com/blog.

17. Edward J. Hu et al., “LoRA: Low-Rank Adaptation of Large Language Models,” *arXiv* (June 2021), https://doi.org/10.48550/arXiv.2106.09685.

18. “Language Models Perform Reasoning via Chain of Thought,” Blog, Google Research, last modified May 11, 2022, https://ai.googleblog.com/2022/05/language-models-perform-reasoning-via.html.

19. Zhuosheng Zhang et al., “Multimodal Chain-of-Thought Reasoning in Language Models,” *arXiv* (February 2023), https://doi.org/10.48550/arXiv.2302.00923.