Generative AI

KEY OPPORTUNITIES AND RESEARCH CHALLENGES

June 2023

Carnegie Mellon University
Software Engineering Institute

Executive Summary

Generative AI (GenAI) has been around for decades, but the latest leap in progress, fueled by high-capability large language models (LLMs), image and video generators, and AI pair programmers highlights the need for further investigation by the defense and national security community. On June 9, 2023, USD(R&E) and the Carnegie Mellon University Software Engineering Institute (SEI) convened a workshop to elicit DoD use cases for GenAI. Workshop attendees included representatives from the Services and from federally funded research and development centers. Key opportunities included use cases in business, wargaming, coding, and simulation. Needs and challenges included understanding "hallucinations," investing in guardrails and responsible AI amid a race to capability, and enabling GenAI at the edge.

Opportunities

Some of the most in-demand use cases for generative AI involve boosting productivity and could bring about a change in the way users interact with this technology, similar to the uptake of mobile devices. Some foresee the rise of GenAI as a personal assistant. Generating and tracking tasks, responding to both emails and queries like FOIA requests, writing contracts, drafting forms and even creating slides to present data and prepare principals for meetings are being investigated and, in some cases, piloted. For soldiers in the field including maintenance and logistics teams, using LLMs as very specific search engines can offer tremendous productivity benefits by helping them sift through policy documents such as uniform regulations, rules of engagement, manuals, and doctrine. A bespoke search engine that could be realized through the integration of LLMs with traditional information retrieval technologies has the advantage of mitigating hallucinations by directing users to the source of the data. However, this approach raises the challenge of multi-security classification boundaries: as statistical generators, would these models be capable of producing results at the proper classification level?

Adjacent to the productivity-boosting potential of GenAI as a personal assistant are pair programming and other coding-related and software engineering use cases. Generating unit tests from feature descriptions of code can save time and effort, but for high-stakes situations, another option is iterative prompting.[1] AI-based pair programmers could offer significant cost savings associated with DoD software modernization efforts. In the cyber domain, finding and fixing bugs and understanding what bugs might arise from vulnerabilities are powerful use cases.

For wargaming, GenAI and reinforcement learning could potentially analyze past data and provide recommendations on courses of action (COAs) and could enable teams to conduct wargaming in a rapid manner that permits the testing of multiple COAs. LLMs in particular play great stand-ins for humans in simulating an agent in the world. Teams can use them for learning and development scenarios and to learn to think like an adversary. There are opportunities for improving warfighter performance similar to how human players of the game Go have been shown to improve dramatically when they play against and learn from AI-based Go players.

Sensors are frequently used to observe and record multi-modal data (e.g., hyperspectral images, radio frequency [RF], radar, and sonar). Based on recent advances in multi-modal foundation models (the models behind GenAI capabilities), there is an opportunity for DoD to develop specialized foundation models that support applications in DoD-specific sensing domains. These foundation models could support advanced domain awareness and force protection applications at the tactical edge and aid wargaming by creating unique maps, imagery, spectrum phenomena, and realistic operating environments. Conversely, multi-modal data may be converted to textual representations to classify and describe ongoing scenarios. Finally, sensors may work together with LLMs to give autonomous systems instructions for non-standard tasks.

It is well known that data—having enough data and having the right kind of data—is a challenge. GenAI could be used to generate synthetic data for research or training without disclosing sensitive information; examples include network data and patterns of life. Generating synthetic data raises the issue of bias (using neural nets to generate data amplifies biases that exist in training data) among other challenges that would need to be addressed.

Research Challenges

While there are many possible DoD use cases for GenAI and LLMs, there remain significant open questions and challenges about their responsible and appropriate use in many DoD applications.

One significant challenge with large language models is the concept of hallucinations, a term used to signify when large language models seemingly "make things up." In a recent highly publicized example, an attorney submitted a defense brief written by an LLM and cited previous cases that do not exist.[2] Hallucinations occur because LLMs have no world model; they simply generate the most likely sequences of text based on the body of text they were trained on. Although hallucinations can be mitigated to some extent by reinforcement learning with human feedback (RLHF), they continue to pose a number of problems, especially with regard to trust in models.

One way hallucinations can be mitigated is through proper attribution. For example, there are LLM training techniques that

1 See Wang et al., 2023: https://arxiv.org/abs/2305.16291
2 https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html

"encourage" the model to answer questions with exact quotes from input text. Other forms of attribution focus on answering questions with summary, citation pairs. The problem becomes more difficult as LLMs are asked to summarize larger and greater numbers of texts.

Associated with the problem of attribution is that of document classification. There is much interest in creating an LLM that can follow security classification guides. The ability to use data across security levels is an even greater challenge: How does aggregating data raise security levels? How can we ensure generated documents are at the appropriate classification level? Can models generate outputs for both classified and unclassified users?

The investment in capability over guardrails, assurance, and testing highlights the need for transparency and dialog between government and industry as well as funding research into how GenAI capabilities work. Silos and a lack of skills in the DoD workforce make this challenge especially difficult. For applications of LLMs as productivity tools, including for pair programming in software development, successful use of LLMs for boosting productivity requires the human partner to have a high level of expertise in order to spot errors produced by LLMs and to craft prompts that reduce errors and fix incorrect solutions. We also need explainability to better understand the innerworkings of these models, especially where the weights live. Processes like LoRA (low rank adaptors) reduce the weight space while allowing researchers to train on new tasks. Studies of this sort can help show where important knowledge "is" within these models.

Related to furthering investment in guardrails is the need for further research into operationalizing responsible AI (RAI) to prevent adversarial or unintentional misuse of GenAI. A baseline approach to responsible AI principles is to pair GenAI with humans to ensure model results are factually correct and accurate, but even this basic practice surfaces challenges: How do we assess the efficacy of human-AI teams for tasks such as generating intelligence reports and evaluating personnel? How do we support collaboration between humans and GenAI that does not take place in a single moment but can rely on continuous feedback? Another risk of the rush to capability is that models get larger and have more data poured into them without regard for explainability, a key to building human users' trust and understanding of their GenAI models.

As mentioned above, there is a real opportunity for DoD to invest in the development of foundation models for domains and modalities that are specific to DoD needs and operations. Foundation models for domains like hyperspectral imaging, RF and electromagnetic spectrum sensing, various space-based sensing capabilities, and other modalities could provide advances in capability across many warfighting missions and applications. Similarly, there is an opportunity for DoD to advance the use and integration of physics-based models with LLMs and other foundation models including emerging capabilities in physics-inspired neural networks, which could lead to improved capabilities for advanced material and weapon-system design.

Without explicit requirements for GenAI, technology readiness levels (TRLs) beyond 5 are out of reach. Furthermore, readiness levels for GenAI are highly dependent on the tool's context of use. More work is needed to understand not only the requirements for GenAI but how to assess the readiness of GenAI for particular uses. Recent and ongoing work in trust readiness levels[3] and calibrated trust are beginning to address this challenge.

Understanding and meeting compute requirements in the enterprise and at the edge are key to empowering the warfighter. Although strides have been made to effectively trim LLMs[4] to require less computational power, GenAI models are generally quite large and present challenges in DDIL environments, where cost, size, weight, and power constrain capability. More work is needed in this area, but two approaches being investigated include deploying pretrained models to low-power single-GPU laptops and distributing LLMs over multiple GPUs. At the enterprise level, considerations should be made for supporting the necessary infrastructure to fine-tune LLMs and to train and maintain DoD-unique and domain-specific foundation models to support warfighting needs.

Risks

A number of risks exist for both DoD applications of GenAI and in the defense and national security space in general.

Information warfare raises a notable risk related to the application of GenAI. For adversaries who seek to spread mis- or dis-information, the quantity of information created and the speed with which it can be disseminated is an advantage. Because precision is less important in such scenarios, the challenge of obtaining accurate responses from GenAI is less of a barrier.

Beyond the potential to generate a large quantity of false information rapidly, GenAI is weak to data poisoning and prompt engineering. These attacks can corrupt results and allow models to be "tricked" into providing results that reveal sensitive information or violate security or ethics policies. The risks of data leakage and information spillage are exacerbated by the fact that users may put sensitive information or code into an LLM or pair programmer.

Other risks posed by GenAI include the difficulty of creating robust codebases that are secure amid rapid change as well as the costs to train models and buy and set up hardware.

Conclusion

As LLMs, AI art, and AI pair programming captivate public audiences, high-stakes DoD scenarios for generative AI require research and investment into not only how reliably these technologies work but also safety mechanisms that allow humans to trust them.

3 See Hobbs et al., 2022: https://arxiv.org/pdf/2210.09059.pdf

4 See Meta, 2023: https://research.facebook.com/publications/llama-open-and-efficient-foundation-language-models/

About the SEI

Always focused on the future, the Software Engineering Institute (SEI) advances software as a strategic advantage for national security. We lead research and direct transition of software engineering, cybersecurity, and artificial intelligence technologies at the intersection of academia, industry, and government. We serve the nation as a federally funded research and development center (FFRDC) sponsored by the U.S. Department of Defense (DoD) and are based at Carnegie Mellon University, a global research university annually rated among the best for its programs in computer science and engineering.

Contact Us

CARNEGIE MELLON UNIVERSITY
SOFTWARE ENGINEERING INSTITUTE
4500 FIFTH AVENUE; PITTSBURGH, PA 15213-2612

sei.cmu.edu
412.268.5800 | 888.201.4479
info@sei.cmu.edu

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM23-0748