An Introduction to Capabilities and Limitations of Large Language Models and Generative AI Technology

Machine learning tools open a plethora of opportunities in content creation. This summary clarifies the capabilities and limitations of these tools and how human guidance is still needed, specifically when considering ethics, credibility and safety.

Many approach the rise of machine learning tools like large language models (LLM) and multimodal generative artificial intelligence (GAI) with great expectations. These systems have multiple capabilities and diverse applications. For example, they can answer questions, analyze sentiment, generate images from text, and follow instructions. Common applications include content creation, translation, code generation, cybersecurity, candidate screening, storytelling, and virtual assistants. LLM and GAI capabilities are growing at an enormous rate, with major new systems and applications announced each week.

This storm of development inspired IDA researcher Dr. Daniel Shapiro to conduct an assessment of what these tools can achieve in principle, with the goal of tying readers' expectations for their capabilities and limitations to a core understanding of the technology. This report explains why generative AI and large language models are able to demonstrate "a level

of intelligence that has never been seen before in a computing system", and are at the same time, "impaired models of cognition". Given this understanding, they draw implications for future development.

LLM is a statistical model of a large training corpus that creates new content from prompts by generalizing past examples. Training an LLM on every sentence in every book ever written results in a statistical model of what comes next given what was seen before. Because that model reflects the knowledge expressed in training texts, LLM prompts can mine it to produce new output.

Because statistical models capture regularities in data, but not the principles that shape the data, GAI output only incidentally reflects properties we value and associate with human cognition. LLMs hallucinate because they cannot distinguish fact from fiction. They have trouble with inference and logical consistency because statistical prediction is a poor substitute for algorithmic reasoning. Because LLMs and GAIs lack norms, they are infamous for producing socially inappropriate content (the Tay Chatbot generated 95,000 misogynistic, racist and anti-Semitic tweets in one day).

The application community is working to address these limitations. For example, text-to-image systems become more socially acceptable after post-training with human feedback or training on a corpus that reflects "acceptable"

imagery. However, these solutions will always be partial. The required effort will only be expended in high-value applications, and the results will have the character of engineering efforts; performance will incrementally improve over time but the statistical behavior of the underlying LLM or GAI will always shine through.

The IDA team concludes that generative AI is not a substitute for human presence. GAI tools can be intelligent, versatile and ubiquitous, but their underlying technology makes them very high risk for safety-critical decision tasks even with substanital mitigation efforts.

This summary is based on IDA Product 3002626.

Daniel G. Shapiro (dshapiro@ida.org) is a researcher in the Information Technology and Systems Division of the Systems and Analyses Center, an IDA managed federally funded research and development center. Daniel holds a doctorate and master's degree in management science and engineering from Stanford University, and a master's in EECS from MIT. His areas of expertise are artificial intelligence, machine learning, cognitive systems, and games.