CSET CENTER for SECURITY and EMERGING TECHNOLOGY

Summary of "Controlling Large Language Models: A Primer"

Concerns over risks from generative artificial intelligence (AI) systems have increased significantly over the past year, driven in large part by the advent of increasingly capable large language models (LLMs). Many of these concerns stem from the risk of models producing undesirable outputs, some of which can fall into three general categories:

1) **Inaccurate information**, which can be particularly dangerous if users come to over-rely on LLMs for what they believe to be factual information.
2) **Biased or toxic outputs.**
3) **Outputs that may result from malicious use**, including information hazards such as information related to chemical, biological, radiological, or nuclear weapons.

However, the inherent complexity of LLMs, their probabilistic nature, and the vast amounts of text data they are trained on all make controlling or steering their outputs a considerable technical challenge. This issue brief explains four popular techniques for controlling LLM outputs along various stages of the AI development pipeline:

1) **Editing pre-training data**
2) **Supervised fine-tuning**
3) **Reinforcement learning with human feedback (RLHF)** and **Constitutional AI**
4) **Prompt and output controls** on fully trained LLMs

Today's methods are more like sledgehammers than scalpels; none of these techniques are perfect, and they are often used in combination with each other. Furthermore, the existence of open models—which anyone can download and customize for their own purposes—also complicates efforts to control LLM outputs, as developers are unable to monitor end users' behavior. Regardless, understanding how LLMs work and how developers can best steer their outputs will be critical for effective AI governance.

For more information:

- Download the report: https://cset.georgetown.edu/publication/controlling-large-language-models-a-primer
- Contact: Jessica Ji (jessica.ji@georgetown.edu).