Knowledgebase > Policy Frequently Asked Questions > Generative AI as a Research Tool

Generative AI as a Research Tool

Tuesday Mueller-Harder - 2023-08-28 - Comments (0) - Policy Frequently Asked Questions

The rapid emergence of generative AI technology using Large Language Models (LLMs) offers great potential for researchers as they design, conduct, support, and present their research. However, there are also a number of challenges and risks to understand when using these tools in the research domain.

Development of Code for Research

A foundational component of many research computing tasks is the development of code that will drive analyses, simulations, data collection, or visualizations. Generative AI can serve as an engine to expedite this development process. In particular, tools like ChatGPT, Bard, and GitHub's Copilot can be useful in automating some of the routine tasks involved in writing scientific software. Three areas are already emerging as strong use cases: writing functional boilerplate code, writing test cases, and documenting code.

- Writing **boilerplate code** is often a challenge, especially when using certain languages and libraries that are considered to be especially verbose. Most modern integrated development environments (IDEs) have included plugins for years that fill in boilerplate code in the form of code snippets. Tools like GitHub's Copilot represent the next generation of these IDE plugins for scaffolding code, since they can be asked to create new code specific to a given purpose. However, the code can include errors in design, syntax, or optimization. It's critical to deeply review all code generated with these tools for quality and efficiency.
- Writing **test cases** is not a popular task, but it contributes a lot to maintaining high-quality code. Fortunately, this work of writing tests is something that generative AI tools generally perform quite well. Given the code for a function, tools like ChatGPT will usually write effective unit tests or integration tests that establish the given function is working properly. Note that this process becomes more involved if your function needs to interact with external resources (such as a SQL database). When interacting with additional resources, it may be appropriate to create a mock version of the resource and reference it when prompting the AI. Again, spend time thoroughly validating the test cases produced by AI tools.
- AI can also be very useful for **documenting code**. Researchers can share portions of their code with generative AI tools, which can then create added documentation in the form of code comments or higher-level descriptions of the functionality of the given piece of code.

Note that when interacting with generative AI tools like ChatGPT or Bard, one should exercise a high degree of caution with respect to any information you share with the tool. These tools are generally hosted by third parties, and any data included in your interactions with the tool may be stored for further machine learning, and even possibly shared back to other unrelated users of these services. Do not share any Level 2 or 3 data, or any algorithms or code artifacts that are proprietary, with AI tools. For reference, the Office of Information Technology (OIT) has shared additional guidance about protecting information when using generative AI tools. The Office of the Vice President for Research (OVPR) also offers advice on public disclosure and ownership of intellectual property when using AI .

Development of Large Language Models

If you are interested in developing your own large language models, or extending existing models that have been open-sourced, there are resources available at OIT’s Center for Computation and Visualization (CCV). This class of models is typically built using machines with discrete GPUs. There are nearly 500 GPUs on Brown’s “Oscar” supercomputer. Any member of the Brown University community can request an Exploratory account to use Oscar, which is provided at no cost. Moreover, the languages and software for building these models (e.g., Python, Julia, C++, PyTorch, JAX, TensorFlow, CUDA, CUDNN) are either already installed on Oscar, or can be easily installed by users.

Researchers interested in extending existing open-source large language models may consider platforms such as Hugging Face. These platforms provide free libraries and APIs that make it trivially simple to fetch pre-trained models and integrate them into an existing application or fine tune a particular task.

For researchers interested in training a large language model “from scratch,” this will require considerable time as well as a large amount of training data. There are many open-source datasets for training these models. A very popular text data set is “The Pile”, which is a collection of datasets including Common Crawl and BookCorpus, among many others.

Attribution

It is important to ensure proper attribution of material generated by AI tools in any research documents. Researchers using LLM tools should document their use in the Methods or Acknowledgements sections of their manuscripts. The Brown Library has included a guide to citation of AI content in their Library Guide about generative AI content and scholarship. Library staff are also available for consultation on questions of citation.

Glossary of Generative AI Terms

- **Bard**: Bard is a conversational chatbot released by Google. It has similar functionality to OpenAI’s ChatGPT.
- **ChatGPT**: ChatGPT is a conversational chatbot released by OpenAI. It has very broad functionality, ranging from creative storytelling to writing code that is used in computer programming.
- **Deep Learning**: Deep learning models—also called deep neural networks, artificial neural networks, or simply neural networks—are a family of machine learning models

for prediction and classification. Deep learning models are extremely powerful; they derive their strength from their flexibility, and particularly their ability to approximate any function. Deep learning models have proven to be particularly powerful for computer vision and natural language processing tasks.

- **DALL-E**: The DALL-E family of models are text-to-image generative models created by OpenAI starting in 2021. These models take some text input and produce an image.
- **Fine Tuning**: It is often desirable to take an existing deep learning model that has already been trained, and repurpose it for a slightly different application. Suppose that there is a pre-trained model that classifies images into one of 1000 different categories (e.g., dog, cat, human, table, potato). Fine tuning would classify images of some subset of those 1000 categories (e.g., only dogs and cats). Creating such a specialized model can improve classification accuracy and performance for its specific purpose.
- **Generative Artificial Intelligence**: Generative AI models are tools for learning the structure of some training data, and then being able to generate wholly new output examples based on the learned structure. The current state-of-the-art generative AI models rely on deep learning models using the transformer architecture.
- **GitHub Copilot**: GitHub’s Copilot is a tool that aids in computer programming. It can be integrated as a plugin to several commonly used text editors and IDEs (e.g., VS Code, Neovim, and JetBrains).
- **GPU**: Deep learning models are generally trained on graphics processing units (GPUs). GPUs were originally developed for rendering graphics—especially 3-dimensional graphics—on computers. In the early 2000s researchers began experimenting with using GPUs for scientific computing. Use of GPUs for scientific computing is extremely effective because GPUs can perform many operations in parallel. Being able to parallelize linear algebraic operations has been particularly transformative for deep learning, and is why GPUs are such powerful tools to drive the development of generative AI technology.
- **Hallucinations**: The term hallucination is frequently used to describe those instances in which ChatGPT or similar conversational chatbots provide output that is factually incorrect, either in part or in whole. It is very common for generative AI tools to invent inaccurate “facts” that sound convincing, but have no basis in reality.
- **Inference**: The inference phase involves using models to make predictions. Inference is conducted when a model is trained. When ChatGPT produces a reply to a question, that is the model running inference. When an iPhone’s camera places square bounding boxes around human faces, that is a model running inference.
- **Large Language Model (LLM)**: Language models are probabilistic models for predicting the next word in a sequence of words. Large language models are essentially hyper-parametric versions of language models—often having millions or even billions of parameters. ChatGPT, Bard, and other popular generative AI tools are

powered by large language models.

- **LLaMA**: The LLaMA—short for Large Language Model, Meta AI—family of models are large language models (LLMs) similar to the GPT family of models. The original LLaMA models were created by Facebook, who then released the models' weights to the research community. These were subsequently leaked via BitTorrent.
- **Model Collapse**: Large language models (LLMs) are trained on massive text data sets. One mechanism of assembling these training sets is crawling popular sites on the internet. The concept of model collapse has been proposed by researchers as a mechanism by which LLMs may gradually degrade in quality because of the nature of the text on which they are trained. Specifically, model collapse describes the degradation of LLMs by a "feedback loop" that will be created as a result of training an LLM on data that was generated by another LLM—rather than the human-generated data that currently makes up most of the text on the internet. As LLMs are used more and more, their output will become an increasingly large proportion of the text on the internet. And as such LLMs trained in the future will be trained less on human-generated content, and more on content generated by earlier LLMs.
- **OpenAI**: OpenAI is an artificial intelligence startup responsible for ChatGPT and the various generations of the GPT family of large language models (e.g., GPT 3, GPT 4) that have powered ChatGPT.
- **Pre-Trained Model**: A deep learning model is considered "pre-trained" when the process of estimating the model's parameters is complete. This is a very computationally expensive process, and can take hours, days, or weeks to complete, even when running on a supercomputer.
- **Training**: Broadly speaking, the process of training a model amounts to estimating the model's parameters. In the context of machine learning, training can be a very time consuming process—often taking hours, days, or weeks. In the case of deep learning specifically, this involves many "passes" through a training data set in small batches. The results of each batch are reviewed for the quality of its results, to tune the model. It is common to make dozens of passes through the data.
- **Training Data**: When building a statistical or machine learning model, the model needs to be "trained" on some input data. This essentially amounts to showing a model many examples of the sort data it should specialize on. For example, if a model is being built that classifies images of animals as either a cat or dog, the training data would consist of many labeled images of cats and dogs.
- **Transformers**: In the context of machine learning, transformers are a species of layer within a deep learning model. Specifically, they use an "attention" mechanism first proposed in 2014 by Bahdanau (et al.). The transformer architecture has proven to be extremely flexible in deep learning models working with both text and image data. The original paper describing transformers is *"Attention is all You Need"* (Vaswani, et al., 2017).

Related Content

- Protecting Information When Using AI Tools