Using Energy and Probing to Enhance Academic Research Capabilities of LLMs
(Technical Topic)

Addressing Risks of Misinformation, Plagiarism, and Inequality of LLMs as Research Assistants for Students
(STS Topic)

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Ganesh Nanduru

May 1, 2024

Technical Team Members: Alexi Gladstone

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Advisors

Kathryn A. Neeley, Department of Engineering and Society

Jundong Li, Department of Electrical and Computer Engineering

Prospectus

Introduction

A state-of-the-art large language model developed by Meta AI recently scored 40.9% on a test for basic factual accuracy (Semnani et al., 2023, p. 2393). How can we trust language models to give us accurate information when they can be so unreliable? Large language models (LLMs) are computer models meant to interpret natural language that are highly effective at understanding text and answering questions (Brown et al., 2022, p. 1878). Their functionality is mainly derived from a process called pre-training, where models will process terabytes of text in order to learn the meaning of language. Because of their enhanced abilities to process large corpora of text and quickly answer questions on the texts, LLMs have enormous potential as research assistants for students. While some educators endorse LLMs to enhance student learning by providing them access to vast knowledge and personalized question answering, others believe LLMs stunt learning by acting as a crutch for students to avoid critical thinking (Oravec, 2023, p.230). Finally, LLMs raise concerns of unequal access among students of different income brackets and ethnicities (Sidoti & Gottfried, 2023).

LLMs such as ChatGPT are spiking in popularity recently; 1 in 5 US teens who've heard of ChatGPT use it for their schoolwork (Sidoti & Gottfried, 2023). While proven to be effective at answering new research questions that students may have via few-shot learning (Brown et al., 2022, p. 1877), LLMs are still not completely reliable, and will sometimes hallucinate evidence when generating responses (Semnani et al., 2023, p. 2387). Additionally, educators find LLMs have invented a new form of cheating on assignments, so they are currently grappling with the extent to which generative language models should be involved in student learning (Oravec, 2023, p. 228).

LLMs are also raising concerns of racial and class inequality; white teens are more likely to be familiar with ChatGPT than their peers, and there is a staggering 34% increase of students familiar with ChatGPT from households making over $75,000 vs. students from households making under $30,000 annually (Sidoti & Gottfried, 2023, “Teens’ awareness of ChatGPT”). It is crucial for LLM developers to consider the social effects of their work, even though they may not feel directly relevant to the development process. A recent advancement in AI is the energy-based model.

Energy-based models have been studied to be effective at generating autoregressive responses in a visual context, such as frame prediction (Wang et al., 2023, p. 3). An autoregressive model assumes that its output is a function of the previous outputs, treated as a timeseries. I propose to adapt the autoregressive strengths of EBMs to natural language processing by pre-training a language model with energy as the objective function, which can result in more rational patterns of thought for answering research questions students may have. Additionally, I will address model hallucination by implementing hidden-state probing for improved factual robustness, as opposed to using typical model querying (Liu et al., 2023, 4792). The technical deliverable of this prospectus will be a language model architecture that implements energy-based training and is resistant to hallucination, so it can be used to reliably aid students with research for their coursework. I will publish this project open-source to help address issues of unequal accessibility.

Despite some professors viewing all uses of LLMs for classwork as cheating, the successful incorporation of AI in education has been studied to actually increase student engagement in the classroom (Bhosale et al., 2023, p. 627). I will produce a report detailing the extent to which LLMs should be used to enhance research while avoiding plagiarism and

reductions in student engagement based on the historical effectiveness of AI implementations in schools. I will include research done on novel interpersonal capabilities of LLMs, so that educators and school boards can have a better sense of whether or not they should allow students to use LLMs to conduct research for their school assignments. The STS deliverable will also include a section on the demographics of LLM users and the barriers of entry to using this technology.

Using Energy and Probing to Enhance Research Capabilities of LLMs

A major innovation of OpenAI's pre-training process was its ability for its models to adapt to new information based on a few (or sometimes even zero) examples, known as shots: "Broadly, on NLP tasks GPT-3 achieves promising results in the zero- and one-shot settings, and in the few-shot setting is sometimes competitive with or even occasionally surpasses state-of-the-art" (Brown et al., 2022, p. 1878). Pre-training was a colossal advancement in AI, greatly enhancing capabilities of computer models to reason and understand semantics of text. The addition of few-shot learning increased pretrained models' adaptability to new information. An innovation by Microsoft in the field of visual processing was energy-based pre-training, where Microsoft researchers were able to pre-train an AI with a new objective function known as energy, which is the compatibility of an output to a dataset, previous response, or any other desired baseline (Wang et al., 2023, p. 2). Figure 1 below demonstrates an example application of energy used to reconstruct a corrupted image.

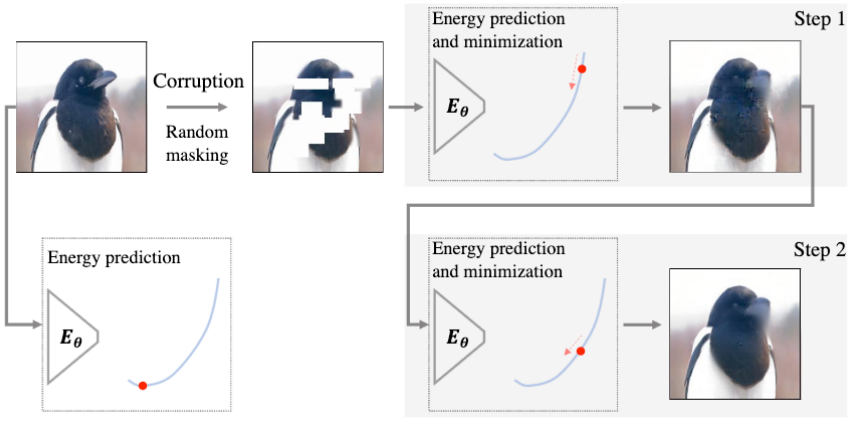

Figure 1. A flowchart detailing the process of gradient descent from the energy metric with a visual example. Lower energy means the picture matches the target class - in this case, a bird. Higher energy means the picture is incompatible with the model's representation of a bird. In the diagram, a picture is masked to increase its energy, then through the minimization of energy, the original image is iteratively regenerated. Adapted from Wang et al., 2023, p. 4.

Despite LLMs recently adapting well to new tasks with few-shot learning, they still occasionally malfunction when faced with unknown information in a process that MIT researchers described, "disagreements between internal and external representations of truthfulness appear predominantly attributable to different prediction pathways" (Liu et al., 2023, p. 4795). When a model comes across information that it did not cover during pre-training, it can struggle with generating an accurate representation of it, which leads to instability in its responses, sometimes causing it to make up evidence, or "hallucinate." Model hallucination is a major source of factual inaccuracy that threatens providing students with misinformation during their research, and even state-of-the art language models face issues with generating misinformation.

LLaMA, Meta AI's most popular language model, had an egregious 40.9% factual accuracy when queried on common knowledge assembled from Wikipedia pages (Semnani et al., 2023, p. 2393). If language models are going to be feasible for academic research, they need to

be more factually reliable, otherwise students will lose credibility when relying on LLMs for knowledge or argumentation.

My deliverable will be a large language model pre-trained with an energy-based objective function specialized for present-future compatibility estimation, building off of Microsoft's results proving the energy model's effectiveness at frame prediction (Wang et al., 2023, p. 9). A challenge I may encounter is model hallucination, which I hope to address using hidden-state probing, a method of model response verification proven to increase factual accuracy (Liu et al., p. 4792).

Addressing Risks of Misinformation, Plagiarism, and Inequality of LLMs as Research Assistants for Students

Education is a constantly evolving field that has recently been adapting innovations in computer technology, such as tools for Artificial Intelligence and Internet-of-Things devices (Bhosale et al., 2023, p.626). Figure 2 below demonstrates how students can leverage LLMs in a multi-step process to improve the quality of their research. However, as with any new change in the sociotechnical system of education, adapting LLMs comes with its own challenges. Teachers commonly attribute student use of language models to cheating (Oravec, 2023, p.213). Language models are also not perfect, sometimes hallucinating evidence to create misinformation (Semnani et al., 2023, p. 2388). Finally, not everyone has equal access to LLMs with barriers such as paywalls restricting use of state-of-the-art chatbots like GPT-4 (Sidoti & Gottfried, 2023, "Teens' awareness of ChatGPT").