Operationalizing Deep Learning: A Sociotechnical Perspective (CSE598) MSCS Portfolio Report *

Adarsh Saripalli
Arizona State University
Tempe, US
asaripa3@asu.edu

I. Introduction

This is a portfolio report for Operationalizing Deep Learning: A Sociotechnical Perspective (CSE598) course taken in Spring 2024 semester. For this course we chose to bridge the gap between the non-transparently trained input data and the outputs generated by LLMs through the integration of robust citation mechanism, aiming to improve the accuracy, fairness, and reliability of outputs.

The increasing autonomy of Large Language Models (LLMs) such as ChatGPT and Stable Diffusion in generating content has sparked widespread interest and debate. Our research began by examining the foundational question of Generative AI's role in the modern digital ecosystem: Is it merely an assistant, or does it function as an independent creator? As our investigation progressed, we focused specifically on language models, the most prevalent type of generative models, analyzing their tendencies to produce hallucinated or otherwise unreliable outputs. In an era marked by escalating concerns over data privacy, content validity, and intellectual property rights, the urgency of developing robust validation methods has become clear. Our project aims to fortify the validation processes for LLM outputs through the implementation of effective citation practices. By emphasizing the importance of citations in LLM outputs, we seek to enable users to verify the truthfulness of the claims made by LLMs and to support these claims with evidence, thereby fostering the generation of accurate and reliable information. Through this project, the team wanted to create an end-to-end system capable of handling natural language questions, generating responses, and providing citations for all retrieved information.

II. Explanation of the Solution

Our approach involves the development of an end-to-end system capable of handling natural language questions, generating responses, and providing citations for all retrieved information. We utilize several datasets for testing our model's referencing capabilities and employ a combination of inline search and closed book models to optimize the generation of verified content. Key components include retrieval systems, synthesis, and post-editing stages to ensure the integration of citations during the text generation phase.

A. Data Collection

We utilized several question-answering (QA) datasets to test our model's ability to reference and synthesize accurate information:

- **ASQA** (A Dataset Of Long-Form Answers For Ambiguous Questions)[2]: Introduced by Stelmakh et al. in 2022, this dataset features long-form factoid questions that require multiple short answers to cover various aspects of the questions. Its novelty lies in its comprehensive long-form answers, which also include translations for short concerns. The dataset primarily consists of information from Wikipedia, specifically from the snapshot dated 2018-12-20.
- **QAMPARI** (An Open-domain Question Answering Benchmark for Questions with Many Answers from Multiple Paragraphs)[3]: Developed by Rubin et al. in 2022, this dataset focuses on answers in the form of lists of entities selected from multiple passages, presenting unique challenges in establishing connections between facts from different contexts. It utilizes the same Wikipedia snapshot as ASQA, ensuring consistency in reference material.
- **ELI5** (Explain Like I'm Five)[15]: Created by Fan et al. in 2019, this dataset is based on queries from the Reddit forum ELI5, where users seek explanations that could be understood by a five-year-old. The questions in ELI5 are diverse and require long, detailed answers with high certainty and multiple references to facts. For this dataset, the Sphere 2021—a filtered version of the Common Crawl corpus—was selected as the basis for the masked language model (LM).

By dividing the corpus into 100-word passages, we enable users to verify the information and also control the low-contextuality of the current LLMs: because each snippet can be traced to the single paper and also independently verified for correctness, it increases transparency of the model's outputs.

B. Evaluation Metrics

The evaluation of the language model outputs is conducted using the following metrics:

- **Fluency:**[4] We assess the naturalness and coherence of the generated text across various contexts using the MAUVE metric. This measure helps determine how smoothly the text flows and how closely it resembles human language.
- **Correctness:**[7] We have developed tailored metrics for each dataset to evaluate how accurately and comprehensively the model responds compared to known facts. This involves checking the factual correctness of the responses.
- **Citation Quality:**[6] With the help of the TRUE NLI model, we grade the citations used in the generated text based on their relevance. This ensures that the content generated is not only accurate but also based on trustworthy sources, which enhances the credibility of the text.

C. Implementation with Inline Search and Closed Book Model

This section delves into three key modeling components of an ALCE system: retrieval, synthesis, and post-editing.

Retrieval: [16] For retrieval, we employ straightforward, readily available retrievers such as GTR and DPR for Wikipedia, and BM25 for Sphere. For each query, the top 100 relevant passages are retrieved to facilitate further processing.

Synthesis: [17] Our primary objective is to investigate methods for guiding a Large Language Model (LLM) to effectively engage with the retrieval system, as well as to synthesize and properly attribute the collected evidence, without the need for fine-tuning the model's internal parameters. A significant obstacle in this endeavor is the constrained context window of current LLMs, which limits the number of passages that can be processed simultaneously.

Given a query q and a corpus of text passages D, the system is required to return an output S, which consists of n statements $s_1, \ldots, s_n$, and each statement S_i cites a list of passages $C_i = \{c_{i,1}, c_{i,2}, \ldots\}$, where $c_{i,j} \in D$.

- **ClosedBook:** We introduce a straightforward closed-book baseline, in which the model receives only the instruction and the question as input, without the inclusion of any retrieved passages. As a result, this particular variant does not provide citations for any supporting evidence.
- **InlineSearch:** In the INLINESEARCH approach, LLMs are granted the ability to invoke a "search" operation during the generation process (Yao et al., 2023; Press et al., 2022; Jiang et al., 2023). At each step, the model can choose from three possible actions: "Search: query" to perform a search among the top 100 passages using GTR; "Output" to generate text; and "End" to conclude the interaction, similar to the INTERACT method. Whenever a "Search" action is executed, the most relevant retrieved passage is displayed within the context. To conserve context space, the passage is removed after a single action.

Post Editing: We employ two techniques to enhance the output quality.

- **RERANK:** To generate multiple candidate responses for each question, we randomly sample nsample = 4 responses. The best response is then selected based on the automatic citation recall score. By employing this RERANK strategy, we anticipate an improvement in the quality of the citations provided by the system.
- **POSTCITE:** To attribute each statement generated by the model, we search for the most relevant passage among the top 100 retrieved passages using GTR and cite it accordingly. In our experiments, we integrate this citation approach with the CLOSEDBOOK baseline to enhance the system's ability to provide evidence for its generated content.

D. Human Evaluation

We have conducted human evaluations of randomly selected outputs by both subject matter experts and laypeople to validate the results from our automated metrics. This data was used for statistical analysis to determine the correlation between human judgments and our automated metrics. This dual approach of using both human evaluation and automatic metrics has confirmed the validity of our methodology while also highlighting potential areas for improvement.

III. Description of the Result

Our methodology has demonstrated significant improvements in the accuracy of LLM outputs, primarily through the inclusion of robust citation mechanisms and the implementation of a closed book model with inline search capabilities. However, our evidence also indicates that maintaining this consistency between the two is still challenging, with nearly half of the outputs lacking adequate citation support. This trend underscores the need for continued development and improvement in our citation integration methodologies.

- **RERANK** consistently improves citation quality on ASQA and ELI5 datasets, as confirmed by both automatic scores and human evaluation.
- **CLOSEDBOOK+POSTCITE** achieves strong correctness but struggles with citation quality. While CLOSEDBOOK outperforms VANILLA in correctness on ELI5 and QAMPARI and has a minor 2 percent gap on ASQA, it cannot provide citations on its own, and the combination with POSTCITE still results in inadequate citation quality.

	CLOSED BOOK	INLINE SEARCH
LENGTH	0.998	0.998
FREQUENCY (MAUVE)	1.56	1.57
CORRECTNESS (Claim)	1.27	1.27
CITATION (REC)	0/1	1/1
CITATION (PRE)	0/1	1/1

TABLE I
LLAMA2 EVALUATION RESULT

IV. DESCRIPTION OF MY CONTRIBUTION

A. Research, Paper Gathering

I was responsible for the critical task of gathering relevant research papers and journals. Given the project's reliance on comprehensive, up-to-date academic resources to inform our approach, I sifted through over 50 articles, papers, and journals. This extensive literature review allowed us to gather pertinent topics and insights from various sources, including multiple device manufacturers, which was essential for understanding the current landscape and integrating these insights into our project..

B. Planning Pipeline for Contextual Entailment to base Llama2

I have visualized the pipeline for contextual entailment to llama2 outputs, which is crucial for assessing whether LLMs function merely as assistive tools or as independent creators. The final layer of the neural network was supplied with contextual entailment filter this component of the project was inspired by the recent discourse surrounding OpenAI's legal challenges, specifically regarding copyright issues, which underscored the relevance of our study. The pipeline was built to enable the model to evaluate the context of questions posed and generate appropriately cited outputs, ensuring both relevance and legal compliance.

C. Report Writing and Review

My role also encompassed the writing and reviewing of our project report. I ensured that the documentation was clear, well-structured, and thorough, reflecting all critical findings and methodologies.

V. New Skills Acquired

This project proved to be exceptionally enlightening, offering a unique opportunity to tackle an organic problem statement with tailored, innovative methodologies. The challenge of integrating robust citation mechanisms into LLM outputs not only addressed a pressing issue in the field of AI—enhancing trust and transparency—but also provided a concrete, practical solution to a problem that spans legal, ethical, and technical domains.

The hands-on experience with cutting-edge technologies and the application of complex problem-solving strategies significantly deepened my understanding of both the potential and the limitations of large language models. It also underscored the importance of interdisciplinary approaches in AI research, combining insights from law, ethics, technology, and user-centered design to develop solutions that are not only effective but also responsible and user-friendly.

This project, with its focus on real-world applications and implications, has been a pivotal step in my professional development, sharpening my skills and broadening my perspective on the role of AI in society.

1) Advanced Research Techniques: Improved ability to conduct extensive literature reviews, critical analysis of academic papers, and synthesis of information from diverse sources relevant to the project.
2) Technical Writing: Enhanced skills in drafting scientific reports, ensuring clarity, coherence, and adherence to academic standards. Learned to effectively communicate complex technical content to varied audiences.
3) Data Pipeline Development: Gained hands-on experience in designing and implementing data pipelines for processing and analyzing large datasets, specifically for contextual entailment in language models.
4) Legal and Ethical Understanding: Developed a deeper understanding of the legal and ethical considerations in AI, particularly in the context of copyright laws and the operational boundaries of LLMs.
5) Collaborative Teamwork: Strengthened abilities in collaborative work environments, learning to coordinate with team members on shared goals, and contributing effectively to a group project.
6) Technical Proficiency in NLP: Acquired practical skills in natural language processing, specifically in the application of LLMs for generating text with citations. Learned to use various NLP tools and techniques to enhance model performance.
7) Problem-solving in AI Applications: Enhanced problem-solving skills by addressing complex issues related to the integration of citations in AI-generated content, and developing solutions that improve transparency and trustworthiness of AI outputs.

VI. Team Members

The group consisted of 5 members which included: Adarsh Saripalli, Aditya Rao, Ameya Shahu, Ayushi Rajshekhar and Harshit Sharma.

References

[1] Gao, Tianyu, Ho-Ching Yen, Jiatong Yu and Danqi Chen. "Enabling Large Language Models to Generate Text with Citations." ArXiv abs/2305.14627 (2023): n. pag.

[2] Sun, Haitian, William W. Cohen and Ruslan Salakhutdinov. "Answering Ambiguous Questions with a Database of Questions, Answers, and Revisions." ArXiv abs/2308.08661 (2023): n. pag.

[3] Amouyal, Samuel Joseph, Tomer Wolfson, Ohad Rubin, Ori Yoran, Jonathan Herzig and Jonathan Berant. "QAMPARI: A Benchmark for Open-domain Questions with Many Answers." IEEE Games Entertainment Media Conference (2022).

[4] Pillutla, Krishna, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi and Zaïd Harchaoui. "MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers." Neural Information Processing Systems (2021).

[5] Nakano, Reiichiro, Jacob Hilton, Suchir Balaji, Jeff Wu, Ouyang Long, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess and John Schulman.

"WebGPT: Browser-assisted question-answering with human feedback." ArXiv abs/2112.09332 (2021): n. pag.

[6] Honovich, Or, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim and Yossi Matias. "TRUE: Re-evaluating Factual Consistency Evaluation." ArXiv abs/2204.04991 (2022): n. pag.

[7] Liu, Nelson F., Tianyi Zhang and Percy Liang. "Evaluating Verifiability in Generative Search Engines." ArXiv abs/2304.09848 (2023): n. pag.

[8] Bohnet, Bernd, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov and Kellie Webster. "Attributed Question Answering: Evaluation and Modeling for Attributed Large Language Models." ArXiv abs/2212.08037 (2022): n. pag.

[9] Stelmakh, Ivan, Yi Luan, Bhuwan Dhingra and Ming-Wei Chang. "ASQA: Factoid Questions Meet Long-Form Answers." ArXiv abs/2204.06092 (2022): n. pag.

[10] Worledge, Theodora, Judy Hanwen Shen, Nicole Meister, Caleb Winston and Carlos Guestrin. "Unifying Corroborative and Contributive Attributions in Large Language Models." ArXiv abs/2311.12233 (2023): n. pag.

[11] Transcript from the FTC's October 4, 2023, roundtable on the Creative Economy and Generative AI.

[12] Choudhury, Avishek and Hamid Shamszare. "Investigating the Impact of User Trust on the Adoption and Use of ChatGPT: Survey Analysis." Journal of Medical Internet Research 25 (2023): n. pag.

[13] Ho, Calvin Wai-Loon. "Generative AI and the Foregrounding of Epistemic Injustice in Bioethics." The American Journal of Bioethics 23 (2023): 99 - 102.

[14] Adam Clark Estes (January 18, 2024). "How copyright lawsuits could kill OpenAI". VoxTechnology. https://www.vox.com/technology/2024/1/18/24041598/openai-new-york-times-copyright-lawsuit-napster-google-sony

[15] Fan, Angela, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston and Michael Auli. "ELI5: Long Form Question Answering." ArXiv abs/1907.09190 (2019): n. pag.

[16] Karpukhin, Vladimir, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen and Wentau Yih. "Dense Passage Retrieval for Open-Domain Question Answering." ArXiv abs/2004.04906 (2020): n. pag.

[17] Yao, Shunyu, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan and Yuan Cao. "LANGUAGE MODELS." (2023).