Statement of Purpose - Jeffrey Cheng

I aim to conduct research addressing the *transparency* and *accessibility* issues of large language models (LLMs) – characterizing aspects of closed industry models and ensuring models and innovations remain computationally tractable. My existing work in this area, on examining knowledge cutoffs, has been honored by an Outstanding Paper award at the inaugural Conference on Language Modeling. As a PhD student at Princeton, I will continue to pursue a research agenda focusing on **the relationship between pretraining data and model behavior** and **efficiency-reasoning tradeoffs during inference**.

Pretraining Data and Transparency

Understanding the relationship between pretraining data and model behavior is crucial for improving LLMs in areas such as continued learning and memorization. Because pretraining datasets for closed models are often withheld for competitive and legal reasons, I am interested in *reducing the transparency gap between open and closed models* by measuring these relationships under the closed data setting. For my first NLP research project, I investigated the effects of data on the temporal alignment of various LLMs' domain knowledge.

In place of pretraining datasets, model providers often report a knowledge cutoff date. Logically, the model's knowledge should be aligned to this cutoff. To test this assumption, we constructed time-varying datasets consisting of monthly versions of Wikipedia documents. We identified the model's alignment as the dates when the perplexity of the versions was minimized, and observed drastic misalignments between the model's knowledge and the reported cutoff, often by periods of many years [1]. To verify our results, we indexed over four trillions tokens from open datasets and showed correlations between the distribution of versions in the datasets and the measured alignment of models trained on them. Though base LLMs are often augmented by retrieval systems, these temporal misalignments are still important considerations when resolving knowledge conflicts between parametric and nonparametric knowledge.

Motivated by the results of this project, I am considering various future research directions, one of which is the effect of data order. Recent studies have shown that data quality matters during the cooldown phase of pretraining [2]; would it be possible to amend the knowledge misalignments by performing continued pretraining on recent versions? Another related direction is how to determine the text inside pretraining datasets. Even when the datasets are open, efficient querying requires the use of data structures such as bloom filters [3] or suffix arrays [4]. I am more interested in methods for the closed data setting such as membership inference attacks [5, 6] and adversarial prompting techniques [7].

Even though our method did not require access to a model's underlying pretraining data to determine knowledge alignments, our results would not have been verifiable without open pretraining datasets. What other fundamental issues of LLMs remain undetected due to the lack of data transparency? In addition to providing me with a deeper understanding of pretraining pipelines and how to efficiently work with large amounts of data, this project **reaffirmed my commitment to conduct research advocating for open science.**

Efficient Language Modeling and Accessibility

LLMs perform better on reasoning tasks when prompted to “think step by step” [8]. However, these performance gains come at the cost of increased decoding latency due to the need to generate the tokens in the reasoning chain. Generating these extra tokens is expensive; the recent o1 reasoning traces provided by OpenAI are hundreds of times longer than the initial queries. Moreover, processing these reasoning chains often require prohibitively large amounts of memory due to the quadratic complexity of attention mechanisms. My second research project focused on *making reasoning more accessible in models* by investigating methods to generate compressed representations of the reasoning chains.

Given a query, we obtained its reasoning chain and passed it through the model to select a learned subset of the hidden states to be its compressed representation. Treating the compressed representations as gold labels, we trained an adapter through teacher-forcing to autoregressively generate them. We then decoded the answer by conditioning on the unmodified query and generated representations. Our work shows that we can view reasoning through the lens of an efficiency-performance tradeoff. Rather than choosing between directly decoding an answer or generating the full reasoning chain, we can adaptively choose a desired balance by controlling the size of the selected subset [9].

This project inspires future research interests. Our project focused on generating representations that are a priori unknown; however, I am also interested in the many lines of work focusing on compressing known context [10, 11, 12]. Another interest of mine involves reasoning with embeddings and knowledge distillation [13, 14], as the compressed representations in this project constitute reasoning in continuous space. To this end, I am currently pursuing a project in improving the efficiency of character-based language models by clustering kernelized character embeddings.

I believe techniques that balance performance and efficiency are important considerations for NLP research. They allow resource-limited devices to effectively allocate computational power and also enable scaling to larger datasets and models without needing additional hardware. I aim to continue doing **research to make models more accessible to users and researchers without a large resource budget.**

PhD at Princeton and Beyond

I am excited about the opportunity to join the Princeton CS program as a PhD student. I would like to work with **Prof. Danqi Chen** on long-context modeling as well as data attribution. These areas are closely related to my interests in LLM efficiency. **Prof. Karthik Narasimhan’s** research in reasoning and efficient generation also intrigues me. Lastly, I am interested in **Prof. Sanjeev Arora’s** work in LLM reasoning and data-efficient finetuning.

Ultimately, my goals are to work as a research scientist in industry or as a PI in an academic lab. With every release of a new LLM, I see these technologies becoming more widespread. Doing research with the potential for global impacts excites me. I want to be a part of the group ushering in the transition to the AI age – a transition not driven by profit, but instead by scientific curiosity and the common good. Pursuing a PhD at Princeton would be a tremendous next step in pursuing my goals.

References

[1] Cheng, J., Marone, M., Weller, O., Lawrie, D., Khashabi, D., and Van Durme, B. “Dated Data: Tracing Knowledge Cutoffs in Large Language Models.” *Conference on Language Modeling*, 2024.

[2] Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K.R., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., Khot, T., Merrill, W., Morrison, J.D., Muennighoff, N., Naik, A., Nam, C., Peters, M.E., Pyatkin, V., Ravichander, A., Schwenk, D., Shah, S., Smith, W., Strubell, E., Subramani, N., Wortsman, M., Dasigi, P., Lambert, N., Richardson, K., Zettlemoyer, L.S., Dodge, J., Lo, K., Soldaini, L., Smith, N.A., and Hajishirzi, H. “OLMo: Accelerating the Science of Language Models.” *ACL*, 2024.

[3] Marone, M., & Van Durme, B. “Data Portraits: Recording Foundation Model Training Data.” *NeurIPS*, 2023

[4] Liu, J., Min, S., Zettlemoyer, L.S., Choi, Y., & Hajishirzi, H. “Infini-gram: Scaling Unbounded n-gram Language Models to a Trillion Tokens.” *Conference on Language Modeling*, 2024.

[5] Duan, M., Suri, A., Mireshghallah, N., Min, S., Shi, W., Zettlemoyer, L.S., Tsvetkov, Y., Choi, Y., Evans, D., & Hajishirzi, H. Do Membership Inference Attacks Work on Large Language Models? *Conference on Language Modeling*, 2024.

[6] Mozaffari, H., & Marathe, V.J. (2024). “Semantic Membership Inference Attack against Large Language Models.” *NeurIPS*, 2024.

[7] Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T.B., Song, D.X., Erlingsson, Ú., Oprea, A., & Raffel, C. “Extracting Training Data from Large Language Models.” *USENIX Security Symposium*, 2024.

[8] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E.H., Xia, F., Le, Q., & Zhou, D. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. *NeurIPS*, 2022.

[9] Cheng, J., & Van Durme, B. “Compressed Chain of Thought: Efficient and Adaptive Reasoning through Dense Representations.” *Preprint*

[10] Qin, G., Rosset, C., Chau, E.C., Rao, N., & Van Durme, B. “Dodo: Dynamic Contextual Compression for Decoder-only LMs.” *ACL*, 2023

[11] Kumari, L., Wang, S., Zhou, T., Sarda, N., Rowe, A., & Bilmes, J. “BumbleBee : Dynamic KV-Cache Streaming Submodular Summarization for Infinite-Context Transformers.” *Conference on Language Modeling*, 2024

[12] Zhang, Z., Sheng, Y., Zhou, T., Chen, T., Zheng, L., Cai, R., Song, Z., Tian, Y., Ré, C., Barrett, C.W., Wang, Z., & Chen, B. “H2O: Heavy-Hitter Oracle for Efficient Generative Inference of Large Language Models.” *NeurIPS*, 2023.

[13] Deng, Y., Prasad, K., Fernandez, R., Smolensky, P., Chaudhary, V., & Shieber, S. “Implicit Chain of Thought Reasoning via Knowledge Distillation.” *ArXiv, abs/2311.01460*, 2024.

[14] Teehan, R., Lake, B., & Ren, M. (2024). CoLLEGe: Concept Embedding Generation for Large Language Models. *arXiv preprint arXiv:2403.15362*.