RESEARCH STATEMENT

Haolin Deng
hldeng028@gmail.com https://harlyndn.github.io

My research interest lies at the intersection of **natural language processing** and **machine learning**. I am particularly intrigued by the potential of foundation models to revolutionize the way we acquire knowledge. My long-term objective is to develop innovative AI systems that can aid in human's **knowledge acquisition** process, ultimately empowering individuals' thinking, learning, and decision-making capabilities.

My recent research focuses on **source attribution in large language models (LLMs)** (Deng et al., 2024). Though LLMs are able to distill an extraordinary amount of world knowledge from trillions of training tokens, they face notable challenges regarding hallucination and ethical or legal responsibilities associated with their outputs (Bommasani et al., 2021, §5.1). These challenges undermine their reliability as knowledge providers, calling for feasible source tracing methods to attribute and verify the information generated by LLMs.

1 ACCOMPLISHED WORKS

As a research intern at Tencent, I have been working on the exploration of AI search engines powered by LLMs. The core idea is to leverage the capabilities of LLMs to synthesize web search results into direct responses that accurately address user queries. The responses also include in-line citations, allowing users to verify the information via the cited sources. This innovative approach eliminates the need for users to sift through multiple web pages, significantly accelerating the knowledge acquisition process. However, ensuring accurate source attribution in the form of citations presents a unique challenge. To address this, my work (Deng et al., 2024) formulated the task of attributed query-focused summarization (AQFS) and created a high-quality Chinese dataset grounded in real-world web search scenarios. We also developed a robust evaluation framework and a cost-effective automatic evaluator, achieving substantial agreement with human assessments. These contributions paved the way for exploring practical system implementation techniques which have later been integrated into Tencent's web search services, including Sogou [1] and QQ browser [2].

Prior to the era of LLMs, I have also worked in various fields of NLP. My work on open event extraction made contributions to the development of automatic trending news identification system used in Tencent's web search services (Deng et al., 2022). During my undergraduate studies, I actively engaged in lab projects on topics like dialog systems Gao et al. (2021).

2 VISION FOR THE FUTURE

My future research agenda encompasses three primary dimensions:

2.1 TOWARDS ACCURATE AND UNIFIED SOURCE ATTRIBUTION.

Though citation generation provides a promising path for source attribution, several challenges remain to be addressed. First, most existing systems such as perplexity.ai[3] cite entire web pages including lengthy articles and research papers. This approach makes information verification time-consuming and cumbersome for users, defeating the system's primary advantage of accelerating knowledge acquisition. A more user-friendly solution is to pinpoint exact text fragments within source documents as fine-grained citations. Also, accurate attribution becomes more challenging as the context size increases (Deng et al., 2024). With current LLMs expanding their context windows to millions of tokens, further studies on source attribution in such long-form contexts are essential.

[1] https://sogou.com
[2] https://browser.qq.com
[3] https://perplexity.ai

Therefore, I plan to explore practical solutions for **accurate attribution in long-context settings**, thereby laying the foundation for more user-friendly and effective source attribution techniques.

Moreover, model outputs are not only influenced by context. A parallel line of research has been studying the impact of training data on model behavior. (Koh & Liang, 2017; Grosse et al., 2023; Park et al., 2023). Enabling LLMs to cite their training data is also a promising direction for future research. By doing so, we could establish **a unified framework that encompasses both in-context and training data attribution**, providing valuable insights into the interplay between the model's contextual and parametric knowledge.

2.2 Towards Multi-modal Knowledge Retrieval and Presentation.

Knowledge dwells not only in plain text but also in structured data like tables, charts, and various multimedia formats. Incorporating informative non-textual data into system responses also helps users understand complex concepts. As such, I intend to enhance the capabilities of current text-only AI search engines by **enriching the supported modalities for knowledge sourcing and response generation**. I will also explore techniques for accurate **cross-modal attribution** to ensure the verifiability and trustworthiness of these systems.

2.3 Towards Autonomous Knowledge Engine.

When people interact with AI systems for knowledge, they may struggle to formulate clear and specific queries. In some cases, they might only have vague ideas or superficial concepts in mind. In other instances, they may simply provide lengthy documents, expecting the system to offer its own insights and analysis. To be truly beneficial in both cases, the AI systems should emulate a "researcher": formulating targeted questions, retrieving and analyzing information, distilling knowledge, generating new insightful questions, and iterating this process while dynamically adjusting the strategy as needed. Given the potential lack of explicit human instructions or supervision, the system should also possess **self-planning and self-improving capabilities**. I aim to contribute to the development of such autonomous AI systems, enhancing their ability to assist users effectively, even when initial queries are less than ideal.

In conclusion, these future research paths present exciting challenges, and I believe that the efforts made in these directions will have a significant impact on the advancement of human-centered artificial intelligence.

References

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *ArXiv*, 2021. URL `https://crfm.stanford.edu/assets/report.pdf`.

Haolin Deng, Yanan Zhang, Yangfan Zhang, Wangyang Ying, Changlong Yu, Jun Gao, Wei Wang, Xiaoling Bai, Nan Yang, Jin Ma, Xiang Chen, and Tianhua Zhou. Title2Event: Benchmarking open event extraction with a large-scale Chinese title dataset. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6511–6524, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.437. URL https://aclanthology.org/2022.emnlp-main.437.

Haolin Deng, Chang Wang, Xin Li, Dezhang Yuan, Junlang Zhan, Tianhua Zhou, Jin Ma, Jun Gao, and Ruifeng Xu. Webcites: Attributed query-focused summarization on chinese web search results with citations, 2024.

Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. Improving empathetic response generation by recognizing emotion cause in conversations. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 807–819, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.70. URL https://aclanthology.org/2021.findings-emnlp.70.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*, 2023.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale. *arXiv preprint arXiv:2303.14186*, 2023.