

---

Foundation models can draw on knowledge acquired during training, retrieve external information, and process inputs across multiple modalities. Yet access to this information does not ensure that a response is supported by the evidence a task requires. Language models can struggle to recover less common factual knowledge [4] and to use relevant passages within long contexts [3], while video models can answer correctly without identifying the supporting moments [9]. These limitations motivate my interest in how models identify relevant evidence and reason from it. I want to develop methods that diagnose these limitations and improve how models identify and reason from evidence.

During my master's at New Jersey Institute of Technology, I compared how different models, training data, and additional input signals affected performance on crisis-informatics tasks [6, 7, 8]. Stronger models or additional information did not consistently improve performance when that information was not sufficiently relevant or effectively incorporated. These experiments showed me why the choice of information and how it enters a model deserve attention alongside the model itself. At Pacific Northwest National Laboratory, I contributed to developing scientific datasets and benchmarks, and later to training and evaluating language models on chemistry tasks [2, 1]. These evaluations showed that models could not always answer chemistry questions reliably using knowledge acquired during training, and the information supporting their answers remained difficult to verify.

Supplying documents at inference time provides an explicit source against which answers can be checked. At PNNL, I contributed to evaluating retrieval-augmented and long-context models for question answering over environmental reports [5]. Through this work, I learned that making the source available was only part of the problem, as retrieving relevant passages produced better answers than supplying entire reports.

Video makes this distinction harder because the relevant evidence may occupy only a few moments within a much longer sequence and must first be located before it can be used. At the UKP Lab at TU Darmstadt, I have studied this problem through video fact-checking. I developed GroundInc to examine whether model decisions were actually supported by the relevant parts of a video. By comparing faithful and misleading captions that shared the same underlying evidence, I found that misleading language made this evidence harder to identify. Correct verdicts also frequently coincided with incorrect evidence localization. I then developed GroundSnip to improve the evidence selection, which in turn improved an off-the-shelf verifier. This showed that changing the evidence supplied to a verifier could improve its decisions without modifying the verifier itself.

Across these projects, I learned that the information available to a model and its ability to use that information must be considered separately. Richer inputs did not consistently improve performance, and knowledge acquired during training did not always yield reliable answers. In document question answering, I learned that supplying supporting information at inference time did not eliminate the need to select relevant evidence. This distinction informed my approach to video-caption verification, where I found that correct verdicts did not necessarily mean models had identified the supporting evidence. I want to build on this work to understand why multimodal models fail to find or use relevant evidence, and how to address those failures.

During my PhD, I want to develop evaluations that distinguish failures to identify relevant evidence from failures to use it once provided. Studying how responses change when evidence is supplied, withheld, or altered would help determine which errors better selection can resolve and which require improvements in how models reason from that evidence. In long videos, I am interested in methods that help models determine when to search for additional evidence, revisit earlier moments, or combine evidence across time and modalities. I want to pursue these questions across VideoQA, long-video understanding, and multimodal agents, using the limitations revealed through evaluation to guide method development. A PhD would give me the sustained mentorship and methodological depth to extend these investigations beyond individual tasks into a broader research program.

## References

- [1] Anurag Acharya, Shivam Sharma, Robin Cosbey, Megha Subramanian, Scott Howland, and Maria Glenski. Exploring the benefits of domain-pretraining of generative large language models for chemistry. *arXiv preprint arXiv:2411.03542*, 2024.
- [2] Sameera Horawalavithana, Ellyn Ayton, Shivam Sharma, Scott Howland, Megha Subramanian, Scott Vasquez, Robin Cosbey, Maria Glenski, and Svitlana Volkova. Foundation models of scientific knowledge for chemistry: Opportunities, challenges and lessons learned. In *Proceedings of BigScience Episode#5-Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 160–172, 2022.
- [3] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.
- [4] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [5] Hung Phan, Anurag Acharya, Sarthak Chaturvedi, Shivam Sharma, Mike Parker, Dan Nally, Ali Janesari, Karl Pazdernik, Mahantesh Halappanavar, Sai Munikoti, et al. Rag vs. long context: Examining frontier large language models for environmental review document comprehension. *arXiv preprint arXiv:2407.07321*, 2024.
- [6] Shivam Sharma and Cody Buntain. Improving classification of crisis-related social media content via text augmentation and image analysis. In *TREC*, 2020.
- [7] Shivam Sharma and Cody Buntain. Method comparison for crisis pipelines. In *Proceedings of the Text REtrieval Conference (TREC)*, 2021.
- [8] Shivam Sharma and Cody Buntain. Bang for your buck: Performance impact across choices in learning architectures for crisis informatics. In *ISCRAM*, pages 719–736, 2022.
- [9] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can I trust your answer? visually grounded video question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13204–13214. IEEE, 2024.