**Statement of Purpose: Arijit Gupta**

With the exponential advancements made in large language models (LLMs) in recent years [3], there has been a sharp rise in their integration into different professional settings and individual lives. This brings about a need to verify the outputs produced by these models, ground them to factual knowledge, and prevent the generation of misinformation. My main research interests are:

- **Trustworthy AI for Science**: How can we detect and mitigate hallucinations in LLMs for high-stakes applications (healthcare & scientific research)?
- **Explainability in NLP**: How can we use linguistic research to make model reasoning interpretable, verifiable, and grounded in factual sources?

My objective is to pursue a Ph.D in Computer Science in order pariticpate in their project to build LLM systems that address the tensions between the capabilities and trustworthiness of modern LLMs.

**Background**

**Trustworthy AI for Science**. Hallucinations are factually incorrect outputs from LLMs, and are a prevalent issue with modern LLMs [2]. It is crucial to address this problem especially for high stakes contexts like healthcare that need accuracy and traceability. I worked for 2 years as a **Data Scientist** in the pharmaceutical industry. We utilized cutting edge language models to interpret **unstructured healthcare data** from the wild and provide targeted insights for our clients. I led a project on the development of a chatbot for medical insurance agents and customers which utilised information from thousands of insurance documents from different providers. To counter the model's tendency to hallucinate I built a pipeline where the LLM was just used for orchestration and formatting, the responses were grounded with cited documents, and any **errors were always traceable**. This taught me that even if we provide LLMs with the right context, they don't always give the correct answers. I also developed the ability to independently develop large scale code pipelines, and design explainable AI systems with error tracing for real-world applications.

For my **master's dissertation** I worked with **Prof. Carolina Scarton** at **The University of Sheffield**, in order to investigate **hallucination span detection** [6] methods in LLMs, as part of the VIGILANT[1] project. Specifically, I explored the viability of smaller, open-source LLMs for hallucination detection by focusing on three aspects. I first benchmarked their performance against state-of-the-art architectures like GPT-4 to assess their accuracy. I also concurrently evaluated their effectiveness as context retrievers from their internal knowledge. Finally, I performed comprehensive error analysis to identify common mistake patterns in order to propose effective mitigation strategies and future directions. We found that smaller, open-source LLMs are comparably capable for all these tasks. This work integrates theoretical understanding with practical problem-solving, and has strengthened my ability to design methods that improve model reliability and interpretability.

[1] https://www.vigilantproject.eu/

**Explainability in NLP**. While LLMs and AI agents perform multi-step reasoning for complex tasks [7], there is a lack of transparency and it is difficult to assess the reliability of their performance. Recent research has shown that these explanations are often superficial and reflect a structured inductive bias that vanishes when LLMs are pushed beyond training distribution [8]. During my undergraduate studies I conducted research on **modelling lexical development** in children using diachronic distributed word representations, an experience that honed my skills in experimental design, data-driven linguistic analysis, and scholarly communication, and I presented our work at the **ACL 2022** Student Research Workshop [1]. To deepen my knowledge further, I worked with **Prof. Kenny Smith** at **The University of Edinburgh**'s Centre for Language Evolution where I investigated the reliability of contextual versus formal word embeddings for noun gender prediction **based on existing linguistic research**, which improved my understanding of embedding evaluation and the interaction between linguistic theory and computational modelling.

**Research Alignment**.

My proposed research area includes creating self improving systems that learn from each other. One of the main challenges of my research direction is model collapse, where LLMs get worse over time by reinforcing errors or biases that stem from hallucinations. With my experience in hallucination span detection techniques and integration of linguistic findings in NLP, I would like to create responsible frameworks for LLM learning. I am particularly keen to make LLMs more involved in decision-making in critical domains like healthcare and scientific exploration. I hope to continue my work in combining language research with computational methods by using existing research on human language transmission to mitigate hallucination-based model collapse [5]. In recent work [4], it has been shown that LLMs can enhance the quality of complex explanations in different domains. It is another direction that I am excited about, investigating symbolic methods to ground LLM explanations to facts, and hence preventing transfer of incorrect information even in an unsuper-vised environment.

**Future Plans**. My career goals upon completing my Ph.D. are to join a leading industry lab as a Research Scientist, contributing towards the development of usable and verifiable models in high-impact, specialized domains, such as analyzing clinical notes in healthcare, where trustworthy AI can have the most significant and positive societal impact.

## References


[1] Arijit Gupta, Rajaswa Patil, and Veeky Baths. Towards Using Diachronic Distributed Word Representations as Models of Lexical Development, 2022. URL https://underline.io/lecture/51950-srw-157.

[2] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025. doi: 10.1145/3703155. URL http://dx.doi.org/10.1145/3703155.

[3] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey, 2025. URL https://arxiv.org/abs/2402.06196.

[4] Xin Quan, Marco Valentino, Louise A. Dennis, and André Freitas. Verification and refinement of natural language explanations through llm-symbolic theorem proving, 2024. URL https://arxiv.org/abs/2405.01379.

[5] Kenny Smith, Simon Kirby, Shangmin Guo, and Thomas L. Griffiths. AI model collapse might be prevented by studying human language transmission. *Nature*, 633(8030):525, 2024. doi: 10.1038/d41586-024-03023-y. URL https://www.nature.com/articles/d41586-024-03023-y.

[6] Raúl Vázquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sánchez-Vega, Alessandro Raganato, Jindřich Libovický, Jussi Karlgren, Shaoxiong Ji, Jindřich Helcl, Liane Guillou, Ona de Gibert, Jaione Bengoetxea, Joseph Attieh, and Marianna Apidianaki. Semeval-2025 task 3: Mu-shroom, the multilingual shared task on hallucinations and related observable overgeneration mistakes, 2025. URL https://arxiv.org/abs/2504.11975.

[7] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

[8] Chengshuai Zhao, Zhen Tan, Pingchuan Ma, Dawei Li, Bohan Jiang, Yancheng Wang, Yingzhen Yang, and Huan Liu. Is Chain-of-Thought reasoning of LLMs a mirage? A data distribution lens, 2025. URL https://arxiv.org/abs/2508.01191.