Statement of Purpose:

My name is Aryasomayajula Ram Bharadwaj. I am writing to express my interest in pursuing a Ph.D. in Artificial Intelligence, with a focus on AI safety. With over six years of professional experience in machine learning (ML) and MLOps, as well as a strong background in software engineering, I have developed a solid foundation in creating scalable ML solutions. My current research interests include developing methods to better interpret and evaluate LLMs to improve trustworthiness and avoid harmful behaviors. I will elaborate on these interests in the following sections.

Professional Background and Experience:

After completing my B.Tech in Electronics and Communications in 2019 at GMR Institute of Technology, India, I joined Mu Sigma Business Solutions, a leading decision sciences firm, where I played a key role in the innovation and development lab. During my tenure, I contributed to various development projects. Notably, I led the development of an LLM-agent platform, which was successfully deployed for internal use in two research labs of major Fortune 500 companies. This platform was designed to assist entry-level and mid-level data analysts with their projects. Additionally, I was instrumental in developing the company's proprietary trading platform and made substantial contributions to the in-house MLOps platform.

Currently, as a Technical Architect at Quantiphi Analytics, I specialize in building LLM-agent-based chatbot solutions for a diverse range of clients. My ability to integrate technical innovation with practical deployments has been recognized through multiple awards, including the *Star Performer of the Team* and various *Impact Awards* throughout my career. These experiences have provided me with a unique perspective on the challenges and opportunities in deploying advanced AI systems in real-world settings.

Independent Research and Contributions:

In addition to my professional responsibilities, I have actively pursued independent research. My recent paper, *Understanding Hidden Computations in Transformer Language Models: Filler Tokens in Chain-of-Thought Reasoning* [1], examines how filler tokens affect the faithfulness of CoT explanations. This work offers critical insights into improving the interpretability of model reasoning by highlighting a relatively underexplored approach in understanding chain-of-thought mechanisms.

I have also received prizes in the **AI Alignment Awards** [2] and the **ELK challenge** [3] which reflects my strong motivation and ability to undertake independent research in AI safety. These experiences have deepened my commitment to understanding the complex interactions within AI systems.

Research Interests:

My current research interests are in two domains that are critical to enhancing the trustworthiness of modern AI systems:

- **Understanding Reasoning Mechanisms in Large Language Models:** Recent studies like **"Let's Think Dot by Dot"** [4] and **"Language Models Don't Always Say What They Think"** [5] have exposed significant limitations in our understanding of how these models do complex reasoning. My research seeks to develop robust methodological frameworks to quantify and improve explanation faithfulness of reasoning mechanisms in LLMs

- **Evaluations for LLM Agents:** The proliferation of LLM agents, which combine powerful language models with tools to act and plan in real-world environments, demands sophisticated evaluation methodologies. My research will address the critical challenge of assessing LLM agents before deployment by developing comprehensive risk assessment approaches, creating metrics to detect potential harmful behaviors, and designing adaptive evaluation frameworks that can anticipate emergent agent capabilities like self awareness, deception.

Why a Ph.D. and Future Goals:

I have actively pivoted my career toward becoming an AI safety researcher over the past few years, driven by my belief that this is one of the most pressing and impactful challenges of our time. Through this journey, I have come to appreciate the immense value of expert guidance and structured mentorship, particularly in a field as dynamic and significant as AI safety. I view a Ph.D. program as the ideal platform to refine my research skills and transition from an enthusiastic practitioner to a dedicated researcher. I am especially interested in working with **Assistant Professor Tianyi Zhou** on understanding limitations of reasoning methods in current LLMs or **Assistant Professor Yizheng Chen** on evaluating LLM Agents for harmful behaviors. I am committed to pursuing a Ph.D. at University of Maryland to grow intellectually and make meaningful contributions to cutting-edge advancements in AI safety.

Conclusion:

My professional background, independent research experience, and passion for addressing challenging problems related to AI safety make me a strong candidate for your Ph.D. program. I am excited about the opportunity to join University of Maryland and contribute to its vibrant academic community. Thank you for considering my application.

Regards,

Aryasomayajula Ram Bharadwaj

References

[1] Aryasomayajula Ram Bharadwaj. (2024). *Understanding Hidden Computations in Chain-of-Thought Reasoning*. https://arxiv.org/abs/2412.04537

[2] AI Alignment Awards Winners. https://www.alignmentawards.com/winners

[3] ELK challenge. https://www.lesswrong.com/posts/zjMKpSB2Xccn9qi5t/elk-prize-results

[4] Jacob Pfau, William Merrill, and Samuel R. Bowman. (2024). *Let's Think Dot by Dot: Hidden Computation in Transformer Language Models*. https://arxiv.org/abs/2404.15758

[5] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. (2023). *Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting*. https://arxiv.org/abs/2305.04388

[6] Paul Christiano, Ajeya Cotra, and Mark Xu. (2021). *Eliciting Latent Knowledge: How to Tell If Your Eyes Deceive You*. Alignment Research Center, December 2021. https://docs.google.com/document/d/1WwsnJQstPq91_Yh-Ch2XRL8H_EpsnjrC1dwZXR37PC8/edit