

Statement of Purpose

My name is Ram Bharadwaj. I am writing to express my interest in pursuing a Ph.D. in Artificial Intelligence, with a focus on interpretability, reasoning, and 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 interests include interpreting and improving the faithfulness of chain-of-thought (CoT) reasoning in LLMs, developing robust evaluation methodologies for LLM-based agents, and exploring other critical areas related to improving our understanding and trustworthiness of modern AI systems. 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, I joined Mu Sigma Business Solutions, 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, built on top of Microsoft's Autogen, which was successfully sold to two major Fortune 500 companies. This platform was designed to assist entry-level and mid-level data analysts throughout 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* [?], examines how filler tokens affect the faithfulness of CoT explanations. This work offers insights into improving the interpretability of CoT reasoning and highlights a relatively underexplored approach in the interpretability domain.

My interest in AI alignment and safety has grown significantly over the past few years. I have received recognition through prizes in the **AI Alignment Awards** [?] and the **ELK challenge**, which reflect my strong motivation and ability to undertake independent research in this critical area. These experiences have inspired me to deepen my understanding of AI systems and their impact on society.

Research Interests

While I am open to exploring a wide range of problems related to improving our understanding and trustworthiness of modern AI systems, my current research interests focus on the following key areas:

- **Faithfulness in Chain-of-Thought Explanations:** Chain-of-thought reasoning is a simple yet highly effective technique for enabling LLMs to perform complex reasoning tasks. CoT explanations also provide a window into understanding the underlying reasoning mechanisms of these models. However, it remains unclear how faithful these explanations truly are. Recent work, such as “**Let’s Think Dot by Dot**” [?], “**Language Models Don’t Always Say What They Think**” [?], and other studies on measuring faithfulness in CoT reasoning, has shown that we cannot fully rely on these explanations. My recent paper explores the problem of CoT explanations involving filler tokens. I believe there is significant value in deeply understanding and addressing the limitations of CoT explanations.

- **Evaluations for LLM Agents:** LLM agents represent a natural evolution of highly capable LLMs, and we are already witnessing their widespread use. However, evaluating LLM agents before deployment in production is a complex challenge. Compared to non-agentic scenarios, agents have a larger action space and limited corrigibility, which increases their risk profile. It is therefore crucial to develop robust evaluation frameworks for identifying dangerous capabilities such as sandbagging, deception, and self-awareness. I see tremendous potential in this area and am eager to contribute to building more reliable and comprehensive evaluation methods.

Why a Ph.D. and Future Goals

I have long been intrigued by the fundamental aspects of intelligence, often spending my free time contemplating questions related to processes behind reasoning. Over the years, I realized the value of expert guidance and structured mentorship, especially in a field as timely and significant as AI alignment and safety. A Ph.D. program provides the ideal environment, culture, and guidance to refine my research skills and evolve from an enthusiastic amateur into a full-fledged researcher. I am particularly interested in working with Assistant Professor David Bau whose research in Interpretability I find really close to my spirit. I am committed to pursuing a Ph.D. at Northeastern University to grow intellectually and contribute meaningfully to cutting-edge research.

Conclusion

My professional background, independent research experience, and passion for addressing challenging AI problems make me a strong candidate for your Ph.D. program. I am excited about the opportunity to join Northeastern University and contribute to its vibrant academic community. I am eager to collaborate with peers and mentors, broaden my research capabilities, and make meaningful contributions to the field of AI. 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] 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>
- [4] 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>