Statement of Purpose

Mihir Panchal
mihirpanchal.tech

Max Planck Institute for Software Systems
MPI-SWS Winter Research Internship Program

My primary research interest lies in developing reliable language and foundation models. Recent large language models (LLMs) have acquired vast amounts of knowledge. However, they remain black boxes with limited interpretability and structure, making it difficult to systematically understand and expand the boundaries of their abilities. LLMs also lack systematic reasoning capabilities and faithful transparency in their decision making, leading to unreliable and often untrustworthy outputs. These issues motivate my two research questions: (1) How can we develop methods to improve LLM reasoning by detecting and mitigating cognitive biases, thereby ensuring more reliable and trustworthy outputs? (2) How can we design model architectures and analytical frameworks that enhance interpretability and provide faithful explanations of logical inference in complex NLP tasks?

Bridging my research interests, I believe there are unexplored opportunities at the intersection of reasoning and interpretability. Model transparency is especially important to give researchers actionable insights into model behavior, opening new directions for architectural innovations and principled design improvements. Building on this motivation, in our upcoming paper at AACL 2025, I contributed to the development of IndicTunedLens [1], an interpretability framework for multilingual LLMs trained on Indian languages. Our work addressed a key limitation of existing interpretability methods, which are largely English-centric and struggle with morphologically rich Indian languages. We trained language specific affine transformations to align intermediate hidden states with output distributions, significantly improving the decoding of intermediate representations and enabling robust layer wise analysis of transformers across diverse scripts and linguistic structures. To further capture semantic dynamics, we introduced token entropy visualization and logit rank evaluation, demonstrating 14% peak accuracy improvement compared to standard LogitLens.

These outcomes reaffirm that interpretability can serve as a foundation for strengthening reasoning in LLMs. Building on this trajectory, I am now exploring probing classifiers as a complementary approach to transform this interpretability challenge into a reasoning framework. By designing probes that identify where and how reasoning signals emerge in hidden representations, my ongoing work [2] aims to systematically connect interpretability with the detection of reasoning structures, ultimately enabling principled interventions for more transparent and cognitively aligned language models. A particularly intriguing direction I want to pursue is demystifying the middle layers of transformers. Recent studies [3] and my own observations [1, 4] suggest that these layers exhibit peculiar behaviors as the context abidance score of LLMs increases, the representations in middle layers appear to undergo non-trivial transformations that are not yet well understood.

Recent work has highlighted both the promise and limitations of interpretability methods for reasoning in LLMs. Techniques such as chain of thought reasoning methods [5], while useful for generating stepwise explanations, can sometimes produce unfaithful post-hoc rationalizations, revealing opportunities to extend probing beyond diagnosis to active mitigation. Advances in sentence level attribution methods [6], which identify ”anchor” steps driving reasoning trajectories, demonstrate the potential for mapping intermediate signals to reasoning behavior, though challenges remain in generalizing these anchors across tasks. Together, these observations illustrate both the risks of over-interpreting current techniques and the promise of new frameworks. Building on these insights, I aim to develop probing based methods that connect interpretability with reasoning, with a particular focus on faithful reasoning intuition detection, reliable interventions, and understanding how middle layer representations transform during inference.

My motivation to pursue these research questions builds on a broad undergraduate foundation, where I combined hands on industry experience with academic research across NLP, artificial intelligence, mathematics, and machine learning [7, 8]. Early in my career, I was among the few students to secure a research opportunity at AI-NLP-ML Research Lab of IIT Patna under the supervision of Dr. Asif Ekbal. There, I worked on the emerging challenge of peer review in light of the growing use of AI in research paper creation. This effort resulted in publications in Q1 and Q2 journals [4, 9, 10, 11], and our AI assisted peer review tools were integrated into mainstream conference and journal pipelines. Seeking to extend this impact to global challenges, I applied to the Climate Change AI Research Cohort at NeurIPS and was one of only 52 students worldwide selected. I collaborated with researchers from Asia, USA and Government of India on predictive modeling for disaster resilience. Our project [12] developed a graph based multi-agent system for landslide risk assessment in the Himalayas, covering over 400,000 km^2 of high mountain terrain and integrating climate and topographical data. This experience highlighted the societal value of AI, where scalable and interpretable models can directly inform disaster preparedness and policy decisions for vulnerable communities. Through these experiences, I realized that translating research insights into broader understanding and practical applications requires not just innovation but also clear communication and mentorship.

What better way to learn a topic for research than to teach it and gain numerous perspectives? My Teaching Assistant experience has been a key part of this journey. During my junior and senior years, I conducted workshops, lectures, and seminars, reaching over 600 students across different levels. Teaching not only strengthened my communication and leadership skills but also gave me a deeper appreciation of the subjects I taught. In explaining complex ideas in accessible ways, I was pushed to think critically, refine my own understanding, and approach problems from multiple angles. This iterative process of learning through teaching proved both challenging and rewarding, reinforcing my conviction that pedagogy and research are deeply interconnected. More importantly, I found fulfillment in mentoring and inspiring students, which motivates me to pursue a career that meaningfully combines research with teaching. I aim to build on this foundation by engaging in impactful teaching opportunities while advancing my research contributions.

My past experiences and current research interests lie at the intersection of reasoning, interpretability, and generative modeling, supported by a rigorous theoretical foundation. At Max Planck Institute for Software Systems, I aim to build on this trajectory by working with Dr. Abhilasha Ravichander and Dr. Mariya Toneva on advancing principled approaches to interpretable and reliable language modeling, and with Dr. Bishwamittra Ghosh and Dr. Camila Kolling on developing efficient methods for scaling and systematically evaluating language models. I am particularly drawn to Max Planck Institute for Software Systems's unique environment, where advances span from mathematical theory to practical applications in AI. My long term vision is to integrate these perspectives into research that not only advances scalable and transparent language models but also establishes principled frameworks for reasoning and interpretability, laying the foundation for my future career as a professor.

References

[1] **Mihir Panchal**, Deeksha Varshney, Mamta Sahni, and Asif Ekbal. Indic-tunedlens: Interpreting multilingual models in indian languages. In *Review IJCNLP AACL*, 2025.

[2] **Mihir Panchal**, Deeksha Varshney, Mamta Sahni, and Asif Ekbal. Probing middle layer representations for reasoning emergence in large language models. In *Preparation for ICLR*, 2026.

[3] Subhabrata Dutta, Joykirat Singh, Soumen Chakrabarti, and Tanmoy Chakraborty. How to think step-by-step: A mechanistic understanding of chain-of-thought reasoning. In *arXiv preprint arXiv:2402.18312v2*, 2024.

[4] Prabhat Kumar Bharti, **Mihir Panchal**, Viral Dalal, Mayank Agarwal, and Asif Ekbal. Not all peers are significant: A dataset exhaustive vs trivial scientific peer reviews leveraging chain-of-thought reasoning. *Scientometrics*, 2025. Accepted.

[5] Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint arXiv:2503.08679*, 2025.

[6] Paul C Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. Thought anchors: Which llm reasoning steps matter? *arXiv preprint arXiv:2506.19143*, 2025.

[7] **Mihir Panchal**, Arnav Deo, Prinkal Doshi, Varad Prabhu, and Chetashri Bhadane. Ledge - leveraging dependency graphs for enhanced context aware documentation generation. *Automated Software Engineering*, 2025. Preprint.

[8] **Mihir Panchal**, Chintan Jagdish Dodia, and Pankaj Dulabhai Rathod. Game machine and algorithm towards trends in game states using machine learning and deep learning. In *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*, 2023. Published.

[9] Prabhat Kumar Bharti, **Mihir Panchal**, and Viral Dalal. Consistentpeer - reviewers through graphrag-driven counterfactuals to measure consistency in peer review. *International Journal of Data Science and Analytics*, 2025. Status: Minor Revision.

[10] Prabhat Kumar Bharti, Viral Dalal, **Mihir Panchal**, Mayank Agarwal, and Asif Ekbal. Co-reviewer - are llms on the same page as human reviewers? an agentic ai framework for evaluating review quality and consensus. *Scientometrics*, 2025. Status: Minor Revision.

[11] Prabhat Kumar Bharti, **Mihir Panchal**, and Viral Dalal. Peergauge - a dataset for peer review disagreement and severity gauge. *Language Resources and Evaluation*, 2025. Status: In Review.

[12] **Mihir Panchal**, Ying-Jung Chen, and Surya Parkash. Cc-grmas - a multi-agent graph neural system for spatiotemporal landslide risk assessment in high mountain asia. In *Neural Information Processing Systems (NeurIPS)*, 2025. Status: Accepted.