Ramesh Arvind Naagarajan

PhD Researcher • Explainable AI, LLM Reasoning, Mechanistic Interpretability

ramesh.naagarajan@etit.tu-chemnitz.de | ramesh-arvind.github.io | Scholar | GitHub | LinkedIn | ORCID

RESEARCH

I work on making large language models reason *faithfully* over structured systems. My PhD targets the gap between optimization-based controllers and the humans who run them: building LLM layers that explain a controller's decisions in language an expert can audit, grounded in the optimizer's own constraints rather than free-form generation. I am moving toward mechanistic interpretability of these reasoning circuits, with the longer-term goal of contributing to alignment research on deployed reasoning systems.

Interests: mechanistic interpretability • LLM reasoning and faithfulness • tool-using agents over symbolic systems • alignment of domain-adapted models • causal reasoning under distribution shift.

SELECTED PUBLICATIONS

Enhancing greenhouse management with interpretable AI: A natural language interface for advanced and optimization-based control

R. A. Naagarajan, K. K. Sathyanarayanan, N. Bauer, S. Streif. *Smart Agricultural Technology*, 2025 | doi.org/10.1016/j.atech.2025.101041

Natural-language layer over a running MPC controller; answers operator questions grounded in the optimizer's active constraints, multipliers, and per-step cost contributions, with a domain knowledge graph constraining hallucination.

Automated analysis and textual summarization of time-varying references in advanced greenhouse climate control

R. A. Naagarajan, K. K. Sathyanarayanan, N. Bauer, S. Streif. *Frontiers in Agronomy*, 2025 | doi.org/10.3389/fagro.2025.1536998

Automated decomposition of time-varying reference trajectories into operator-readable components (diurnal, seasonal, event-driven), enabling auditable conversation about setpoint design.

EXPERIENCE

PhD Researcher (Wissenschaftlicher Mitarbeiter), *Chair of Automatic Control and System Dynamics, TU Chemnitz* Sept 2024 – present

Advised by Prof. Dr.-Ing. Stefan Streif. Lead author on the LLM-reasoning + interpretability track of the group's controlled-environment-agriculture program. Prior research-assistant role in the same group (Sept 2023 – Aug 2024) produced both 2025 publications above. Currently extending the work toward mechanistic interpretability of domain-adapted reasoning models.

Software Engineer (Java / Data), *Vaillant Group* • *Tata Consultancy Services* 2018 – 2024

Six years of production software experience prior to research transition. At Vaillant: data acquisition pipelines from IoT sensors and a monolith-to-microservices migration in Spring Boot (∼60% latency reduction). At TCS: scalable JVM services on Elasticsearch / MongoDB.

EDUCATION

MSc, Digital Transformation, *University of Applied Sciences and Arts Dortmund (FH Dortmund)* 2022 – 2024

Focus: data science, business intelligence.

MTech, Computer Software Engineering, *Vellore Institute of Technology (VIT), India* 2013 – 2018

CGPA 8.44 / 10.

SKILLS

Research	Mechanistic Interpretability • Explainable AI • LLMs • Natural Language Generation • Model Predictive Control • Causal Reasoning
ML / Tools	Python • PyTorch • Hugging Face • scikit-learn • pandas • NumPy • Git • Linux
Engineering	Java • Spring Boot • REST APIs • SQL • Elasticsearch • MongoDB
Languages	English (professional) • German (A2) • Tamil (native)