Ajay Pravin Mahale

AI Researcher | Machine Learning Engineer

Trier, DE | +49 176 88247239 | mahale.ajay01@gmail.com | linkedin.com/in/ajay-mh | github.com/designer-coderajay

PROFESSIONAL SUMMARY

AI Researcher specializing in foundational models, multi-agent systems, and mechanistic interpretability. Built a causal circuit-analysis pipeline for GPT-2 that beats attention-based baselines on 75% of prompts, revealing confidence is near-uncorrelated with faithfulness (r = 0.009). Positioned to push boundaries in next-generation AI agents and machine learning at the van der Schaar Lab.

SKILLS

Programming: Python, SQL

Tools & Frameworks: PyTorch, TransformerLens, Hugging Face, scikit-learn, LangGraph, Model Context Protocol (MCP), Qdrant, LlamaIndex

Expertise: Machine Learning, Foundational Models (LLMs), AI Agents (Multi-Agent Systems), Mechanistic Interpretability, Reproducible ML Pipelines, Quantitative Research

Languages: English (C1), German (B1), Hindi (Native), French (Beginner)

WORK EXPERIENCE

Machine Learning Engineer (Research Collaboration) **Nov 2025 – Apr 2026**
BaseMotion AI, Berlin, Germany

- Evaluated foundational model faithfulness by applying mechanistic circuit analysis on GPT-2 Small, establishing baseline reliability metrics using ERASER methodologies.
- Built a reproducible circuit-to-language pipeline in PyTorch and TransformerLens with an automated test suite and CI, so evaluation runs were auditable and repeatable.
- Measured a near-zero correlation (r = 0.009) between foundational model confidence and explanation comprehensiveness, demonstrating confidence is an unreliable proxy for model faithfulness.
- Showed that fluent LLM-written explanations raised perceived quality from 60% to 99% (ERASER) without any corresponding gain in actual mechanistic explanation faithfulness.

Independent Researcher (Project Thesis & AI Safety) **Apr 2024 – Sep 2025**
Hochschule Trier, Trier, Germany

- Synthesized 25+ research papers across AI alignment, scalable oversight, and mechanistic interpretability to map current boundaries in foundational model safety.
- Authored structured technical reports cataloguing frontier LLM failure modes and compared robustness methodologies to establish baseline quantitative metrics for interpretability.

AI Engineer **Feb 2021 – Mar 2023**
ICEICO Technologies Pvt. Ltd., Nagpur, India

- Engineered custom data pipelines and tuned hyperparameters in PyTorch to train predictive models, establishing reproducible machine learning baselines for internal testing.
- Implemented foundational generative AI capabilities by constructing a Retrieval-Augmented Generation pipeline using LangChain to automate knowledge retrieval across operational datasets.

PROJECTS

Causally-Grounded Mechanistic Interpretability (M.Sc. Thesis) **Dec 2025 – May 2026**
GitHub: github.com/designer-coderajay/Causally-Grounded-Mechanistic-Interpretability

- Designed a causal circuit-to-language evaluation pipeline where the discovered model circuit beat the attention baseline on 75% of prompts evaluated.
- Applied ERASER sufficiency and comprehensiveness metrics systematically to mechanistic interpretability on the Indirect Object Identification task in GPT-2 Small.
- Proved 100% sufficiency yet 22% comprehensiveness, showing sufficiency alone overstates evaluation reliability in language models.

Glassbox – Mechanistic Interpretability & EU AI Act Toolkit **Feb 2026 – Mar 2026**
GitHub: github.com/designer-coderajay/glassbox-mech

- Published an open-source mechanistic interpretability toolkit automating EU AI Act Annex IV documentation, successfully achieving 1,000+ PyPI downloads/month in deployment.
- Discovers causal circuits behind transformer predictions via activation patching, providing vendor-independent evaluation tooling deployed with automated test suites.
- Live HuggingFace demo, passing CI/CD badge, and arXiv preprint (2603.09988) demonstrating practical evaluation tool deployment.

Enterprise Agentic AI Platform **May 2025 – Jun 2025**
GitHub: github.com/designer-coderajay/enterprise-agentic-ai-platform

- Designed a multi-agent system with LangGraph (0.3) orchestrating 3 MCP servers (Postgres, document, notification) and hybrid RAG to transition MVPs into enterprise solutions.
- Built robust data pipelines via FastAPI streaming over WebSockets to a Next.js UI, structured with Celery for asynchronous processing.
- Instrumented end-to-end observability (Langfuse, Prometheus/Grafana) and containerized via Docker Compose to ensure deployment scalability.

EDUCATION

M.Sc. Interdisciplinary Engineering (Focus: AI & Machine Learning) **Apr 2023 – Jun 2026**
Hochschule Trier, Germany | Overall grade: 1.8, Thesis: 1.0

B.Tech in Automotive Engineering **May 2018 – Jun 2021**
Government College of Engineering and Research, India | Grade: 1.8

CERTIFICATIONS AND PUBLICATION

- **Microsoft Certified:** Azure AI Engineer Associate (AI-102), Azure AI Fundamentals (AI-900)
- **BlueDot Impact (Fellowships):** Technical AI Safety (30-hour cohort, Jan 2026)
- **Google:** AI Essentials, UX Research (Professional Certificate)
- **Explainable AI for LLMs: A Causally Grounded Pipeline.** Preprint, submitted to the Workshop on Mechanistic Interpretability. arxiv.org/abs/2603.09988