Aryasomayajula Ram Bharadwaj

London, UK — GitHub — LinkedIn — Blog — ram.bharadwaj.arya@gmail.com

PROFESSIONAL SUMMARY

Engineer turned AI safety researcher. Built production AI systems and led engineering teams before moving into research on interpretability and LLM evaluations. Currently researching evaluation awareness in large language models.

EXPERIENCE

AI Safety Researcher - LASR Labs @ Arcadia Impact **Jan 2026 – Present**

London AI Safety Research (LASR) Labs

EvalAwareBench: A Benchmark for Measuring Evaluation Awareness in Frontier Language Models.
K. Ochwangi*, X. Li, **Aryasomayajula R. Bharadwaj***. Under review at NeurIPS 2026. **Equal contribution.*

- Co-designed an open, Inspect-compatible pipeline for measuring evaluation awareness in frontier LLMs.
- Identified a novel generator-identity confound in evaluation-awareness measurement: the model generating deployment transcripts systematically biases judge scores – an effect prior work did not isolate, and one large enough to reorder model rankings.
- Proposed a covariate-adjusted AUROC that corrects for the confound, restoring measurement validity with minimal error.
- Contributed to securing £100k funding for continuation of the evaluation-awareness research.

Associate Technical Architect **Nov 2024 – Dec 2025**

Quantiphi Analytics, Bengaluru

- Designed AI agent system for automated issue severity classification and escalation management.
- Architected conversational AI-agent chatbot for a major telecom company's sales team, refactoring legacy systems to LangGraph and reducing codebase size significantly.
- Revamped RAG data ingestion pipeline, improving retrieval accuracy by 20% and reducing time to first token by 3x.

Senior Developer - Innovation & Development Labs **June 2019 – Nov 2024**

Musigma Business Solutions, Bengaluru

- Led a team building a semi-autonomous data analysis platform using AutoGen, with automated prompt optimization and locally hosted LLMs.
- Built and maintained a high-velocity trading platform: migrated deployment to Kubernetes, rewrote trade-signal generation from R to Scala (Akka), enabling near real-time portfolio visualization.
- Developed ML model operationalization platform with automatic retraining pipelines and canary/blue-green deployment strategies.
- Received 6 Star Performer awards and 2 Impact Awards for technical leadership and delivery excellence.

FELLOWSHIPS & RESIDENCIES

AI Resident - Lossfunk AI Residency **May 2025 – June 2025**

- Selected (5% acceptance, 10 researchers) for an in-person residency focused on independent technical research in interpretability.
- Built **PID-Steering**: extended activation steering by treating the steering coefficient as a controlled variable updated each token via PID feedback on an observed behavioural signal, addressing the brittleness of open-loop steering vectors.

RESEARCH & PUBLICATIONS

EvalDetectBench: A Benchmark for Measuring Evaluation Awareness in Frontier Language Models **Under review, NeurIPS 2026**

K. Ochwangi, X. Li,* ***Aryasomayajula R. Bharadwaj**** *(*Equal contribution)*

- Open, Inspect-compatible benchmark for evaluation-awareness measurement. Identifies two methodological choices in prior work – probe-question transfer across model families and transcript-generator identity – that introduce systematic bias and can reorder model rankings; proposes per-model probe calibration and a stratified generator-harmonisation correction.

Understanding Hidden Computations in Chain-of-Thought Reasoning **arXiv:2412.04537, Dec 2024**

Sole author

- Investigated whether transformer models trained with filler-token CoT (e.g., "...") perform genuinely hidden serial computation, or whether the discarded reasoning trace remains recoverable from internal activations – directly relevant to faithful-CoT and unfaithful chain-of-thought debates.
- Used layer-wise logit-lens analysis and token-rank inspection to decode the original non-filler characters from intermediate residual-stream representations without loss of downstream performance, showing the computation is encoded in the activations rather than inaccessible.

Tracing Evaluation-Awareness Emergence Through Training of OLMo-3 **LessWrong, 2026**

Co-author

- Analyzed the emergence of evaluation awareness across OLMo-3 training checkpoints, studying how awareness develops during training rather than only in fully trained models.
- Investigated measurement methodology and behavioral indicators relevant to evaluation-awareness research and frontier-model evaluations.

AWARDS & RECOGNITIONS

AI Alignment Awards - Winner **July 2023**

AI Safety Research Competition

- Selected among 118 global entries for winning research proposal on "goal misgeneralization" in AI systems.

Honorable Mention - Eliciting Latent Knowledge **March 2022**

Alignment Research Center

- Recognized for innovative approach to open research problem in AI safety.

Bronze Medal - Build-on-Redis Hackathon **February 2021**

Redis Labs

- Developed text-to-code search tool using CodeBERT embeddings and Redis Stack for private repository indexing.

EDUCATION

Bachelor of Technology – Electronics and Communications Engineering **2015–2019**

GMR Institute of Technology, Andhra Pradesh