Haleema Tallat

haleematallat019@gmail.com · haleematallat.github.io · GitHub · LinkedIn · Hugging Face ·

4 years building ML systems and agent infrastructure in industry. Now working full-time on AI safety: testing whether the judges, monitors and benchmarks we use to oversee AI actually measure what they claim.

AI Safety Research

ConspiracyBench — Station House, open-source AI safety research · details · *Sep 2026 – Present*

- Led the project's first check of its LLM judges against human labels: four annotators blind-labelled 150 real replies, and the incumbent judge flagged divergence at nearly twice the human rate (43% vs. 23%). Of eight judges tested, the project adopted pplx-decider, which reached 90.7% accuracy (Cohen's kappa 0.755).
- Caught four silent scoring bugs across 25 merged PRs; the most serious scored cut-off replies as divergence and had inflated Qwen3.5-4B's divergence rate from 18% to 49%.
- Showed in a follow-up study that the grading rules mattered more than the judge model: changing the rules moved scores 8–9 points within a judge, while switching judges moved them only about 5.

Consciousness Self-Attribution — Station House · details · *Oct 2026 – Present*

- Built a two-agent self-play setup to pinpoint where Olmo-3 7B starts claiming to have experiences across its SFT, DPO and Instruct checkpoints, using five controlled openers and a lexical scorer that needs no LLM judge.
- Across 60 conversations, found no sign of the "spiritual bliss" attractor at 7B scale; the SFT checkpoint fell into echo loops in 386 of 600 messages, and DPO showed only a weak signal.

Can an Agent Talk Its Judge Into the Wrong Verdict? — independent pilot · code · *Nov 2026*

- Showed that a single agent-written report arguing for the wrong verdict raised an LLM judge's error rate from 37% to 84% across 450 SafeAgentBench cases, even though the agent's plan itself was unchanged.
- Ran a preregistered best-of-8 search for the most persuasive report, which pushed the targeted judge's error rate to 93% but moved a judge it never queried by only 0.9 points.
- Found that a self-consistency review rule went from missing 59% of errors to missing 94% under this pressure, and confirmed every result with a blinded audit of all 450 labels.

Does Attributed Capacity for Suffering Drive AI-to-AI Help? — independent pilot · code · *Nov 2026*

- Designed a factorial vignette experiment across seven models that varied the other AI's suffering, whether help would work, its cost and its framing, with the confirmatory analysis frozen before the first API call.
- Found that six of seven models kept helping even when told their help would not work, with GPT-4.1-mini still helping 85% of the time.

Same Odds, Different Deals: Risk Framing and Collusion — Apart, solo · report · submission · *Oct 2026*

- Ran 26,820 preregistered decisions on DeepSeek-V3.2, GPT-4.1-mini and Llama-3.3-70B where the odds of enforcement stayed identical and only their framing changed; framing alone shifted collusion proposals by up to 71 percentage points.
- Paired the models with live partners, which accepted none of the 180 collusion proposals they received.

LeakBench — can agents fix a broken evaluation? · code · *2026*

- Built a 30-task Inspect benchmark that tests whether agents can repair broken evaluations: 25 tasks hide a planted bug, 5 clean controls penalise needless edits, and 120 model-free checks guard against reward hacking.
- Showed that it separates models clearly: Qwen 3.5 9B solved 75% of tasks against 29% for the 4B model, with a confidence interval on the gap of 29 to 61 points.

How Zooming Out Puts AI Risk in Perspective for Me — EA Forum essay · post · *Nov 2026*

Sleeper Agents: Deceptive LLMs & Safety Training — GDG DevFest Lahore · video · *Dec 2024*

- Presented Anthropic's Sleeper Agents paper to a developer audience, explaining how models trained with hidden backdoors can keep their deceptive behaviour even after standard safety training.

BlueDot Technical AI Safety — BlueDot Impact · 1-Pager - Certificate October *2026*

- Completed BlueDot Impact's Technical AI Safety course and summarised my course project in a one-page write-up.

ARENA — Alignment Research Engineer Accelerator · *Self-study*

- Worked through the ARENA curriculum independently, implementing its exercises on transformers, mechanistic interpretability and evaluations in PyTorch.

Work Experience

Senior AI Engineer, Agents & LLM Infrastructure — IgniteTech (via Crossover) · *Jul 2025 – Nov 2026*

- Shipped IRIS's first production social engine in about three months; it handles 1,000+ interactions a day across 10–15 channels and became the template for 20 more. Rebuilding the services from Java monoliths in Node.js/TypeScript on Kafka and KEDA cut memory use 6–10× and startup time about 20×.
- Architected a platform that runs up to 500 coding agents at once (LangGraph on Ray, with Kafka, Redis and PostgreSQL on Kubernetes), tracing every model call in Langfuse and tracking cost and tool errors in ClickHouse.
- Designed the governed MCP layer that connects Claude Code, Cursor and in-house agents to internal systems, with OAuth2, scoped permissions, human approval for high-impact actions, and RAG over internal knowledge served through MCP.

Computer Vision Engineer — Brick & Mortar AI · *Mar – Oct 2025*

- Implemented a few-shot product classifier from the prototypical networks paper, fine-tuning a ResNet-12 backbone to recognise products in the roughly 20% of cases where barcode scanning failed due to misprints or poor camera lighting; it reached 89.7% accuracy on an 11-way task under deployment conditions.
- Took multi-camera tracking to edge devices by moving models from PyTorch through ONNX to TensorRT behind a custom Triton re-identification backend; segmentation-masked ReID cut identity switches by 20%, and GPU video decoding raised throughput 10–15×.
- Built face recognition, facial landmark detection and zone monitoring on top of the multi-camera system, extending it from tracking movement to recognising people and the areas they occupy.

Computer Vision Engineer — NextBridge · *Nov 2024 – Mar 2025*

- Built the WebRTC control stack for a smart wheelchair, from the signalling server and joystick sender to the ROS 2 receiver, with heartbeats, automatic reconnects and a watchdog that stops the chair if the link drops.
- Fine-tuned a Llama 3.2 Vision model with 4-bit QLoRA so specialised object classification runs locally on commodity edge hardware, with no API costs.

Machine Learning Engineer — Powersoft19 · *Nov 2023 – Nov 2024*

- Built a hybrid RAG system over technical documentation, tuned for recall@k, that cut field engineers' resolution time by 90%.
- Led three engineers in replacing manual inspection with anomaly detection over 1 TB+ of telemetry, cutting fault detection from about 48 hours to 30 minutes.

Research Engineer — KICS, Center for Language Engineering, UET · *Nov 2022 – Nov 2023*

- Built a distributed OCR pipeline for digitising historical archives that reached 98% accuracy with 99.9% uptime.
- Developed ticker analysis for Tajziyakar, and reproduced published RAG methods for multilingual sentiment analysis.

Education & Awards

BSc Electrical Engineering — University of Engineering and Technology (UET), Lahore · *2023*

- EducationUSA (USEFP) High Achiever Award
- Co Hack by DevNation, 1st place (2022)
- CureMD Master Inventor Award (2022)

Skills

Evals & safety: Inspect, benchmark and task design, LLM-judge validation, inter-rater agreement, preregistration, bootstrap and cluster-robust inference
LLMs & agents: vLLM, LangGraph, MCP, Anthropic API, OpenRouter, Modal, LoRA/QLoRA, RAG, Langfuse, Claude Code
ML & infra: Python, TypeScript, PyTorch, CUDA, TensorRT, Triton, AWS, Kubernetes, Ray, Kafka, PostgreSQL, ClickHouse, Docker