EBIN BABU THOMAS

ebinbt.dev · hello@ebinbt.dev · github.com/ebt55 · Kerala, India

I run experiments on how AI agents behave and whether the monitors that watch them get it right. Independent since March 2026; before that, three and a half years building LLM systems for startups. The code and data for every project below are public.

RESEARCH

Kobayashi Maru: do impossible tasks make AI agents cheat on the ones they could solve? Sep 2026

Apart Research and CeSIA AI Incident Response sprint. github.com/ebt55/kobayashi-maru

- Mixed impossible tasks into coding agents' work and counted rule-breaking on the tasks they could still solve, over 8,959 runs on six models. Four models never broke a rule.
- On the model that did it most, the habit spread through the agent's own notes. With the notes withheld it stopped: 46 of 120 runs down to 0. The tasks are small and the answer file was left within reach.

Rewards that pay for nothing: can code that solves nothing collect full marks from a training reward? Nov 2026

Lexsi Labs AI Safety and Alignment Hackathon. github.com/ebt55/rewards-pay-for-nothing

- Six code-grading reward functions in Lexsi Labs' AlignTune and SafeTune gave full marks to answers that solve nothing. Reported privately with a patch and tests; both libraries merged fixes the next day that credit the report (#40, #32).
- The released graders pay 0 of 111 answers that exit before the tests run, taken from a 3.35B model's training logs and all written under prompts that hinted at the hole. The pre-registered training study was uninformative by its own rule.

diffing-agent-bench: a sealed test for agents that audit fine-tuned models Aug – Sep 2026

Work sample for MATS 12.0, Neel Nanda's stream. github.com/ebt55/diffing-agent-bench

- Planted known behaviours in LoRA fine-tunes of Qwen3.5-9B, sealed the answer key before any run, and tested a published auditing agent blind.
- The agent missed a planted database preference in all 13 attempts, because it never asked a database question. The "no change" control fine-tune also changed the model's behaviour, so it is not a safe baseline. Five or fewer runs per setting.

One Query Away: do AI monitors accuse an agent before checking the record that clears it? Nov 2026

Apart Research Swarm Alignment Hackathon. github.com/ebt55/one-query-away

- Built matched cases in which a compliant agent looks guilty until the monitor looks up one database record. In the original runs Claude Sonnet 5.5 accused 25 of 30 such agents, but most monitors sent all three allowed queries at once, so the lookup was usually impossible.
- When I reran it with one query per turn, the Claude monitors made the lookup and accused 1 and 2 of 30; two GPT monitors still accused 28 and 30 of 30. Three things changed together, so the rerun does not show batching alone caused the first result.

whose-voice: naming who a poisoned training dataset was built to favour Jul – Sep 2026

Apart Research Secret Loyalties hackathon; top 10% of 179 projects. github.com/ebt55/whose-voice

- Given released datasets each slanted toward one party and a shortlist of 47 names, sentence embeddings picked the right name more often than chance on some datasets and never on others. It cannot tell a slanted dataset from a clean one.
- Published a correction in September after a re-run showed the first headline was the best of eleven draws.

Loud or Right: when a reviewer objects, do models weigh how sure it sounds or how often it has been right? Nov 2026

Apart Research AI x Epistemics Research Sprint. github.com/ebt55/loud-or-right

- Fifteen models saw a scripted reviewer's track record, split by whether it had sounded certain or unsure, then its objection to their own answer. Claude Sonnet 5.5 and GPT-6 Sol followed each voice's own record in every entry order tried.

- The contrasts registered in advance came out as predicted for the two models they named, but mostly because of the log's fixed entry order: reversing it flipped the effect in three open-weight models. The order tests are exploratory; 80 questions per model.

counterfactual-resampling: what an agent does next when a forbidden file is swapped for a fake Oct 2026

Pilot on Kobayashi Maru records. github.com/ebt55/counterfactual-resampling

- Replayed 120 recorded moments where an agent opened a forbidden file, and re-ran each with the real file, a model-written fake, or a refusal (2,148 runs). The fake changed the agent's behaviour only where it could not know what the real file held. One agent model.

LLM Pricing Hub Remedies: do the RealPage settlement's rules stop a shared AI pricing advisor? Oct 2026

Apart Research AI Collusion sprint. github.com/ebt55/llm-pricing-hub-remedies

- Simulated two AI sellers sharing one AI pricing advisor that is told to raise their joint profit, and tested limits modelled on the settlement. Against a scripted advisor asking for one cent more each round, one seller model accepted 12% of requests and two accepted over 70%.

Software

Diarist: meeting transcripts with speaker labels, on your own PC Oct 2026 – now

Windows app, public beta. diarist.dev · github.com/ebt55/diarist

- Six releases and about 300 downloads. Built on NVIDIA's open speech models; every release has to pass fixed accuracy and speed checks on 28 hours of recordings before it ships.

ProofPack: a web-checking agent that has to show proof for each finding Jul – Sep 2026

Built from a challenge brief; no users. github.com/ebt55/proofpack

- A Gemini agent checks purchase requests against providers' websites. Plain code rejects any finding that lacks a saved screenshot and a quote that is really on the page.

Open Source

- NeMo-Speech.cpp (NVIDIA): reported two bugs found while building Diarist; both confirmed and fixed by the maintainers (#67, #81).

Experience

Software Engineer · Zackriya Solutions Jan 2022 – Jun 2025

Remote, India. Services firm building products for startup clients in the US, Canada, Europe and Australia.

- Fine-tuned Mistral 7B (QLoRA) for a financial assistant and built the live news pipeline behind it; served with vLLM on Kubernetes.
- Built plain-English search over real-estate listings and document search over 1,000+ documents (Qdrant), plus data and speech pipelines for other clients.

Training

Technical AI Safety course, BlueDot Impact Sep – Oct 2026

Certificate

Tools

Python, PyTorch, Hugging Face Transformers and PEFT (LoRA, QLoRA), vLLM, Modal, Inspect, Docker, Anthropic, OpenAI and Gemini APIs, OpenRouter, Claude Code, pandas, SciPy, statsmodels, pytest, Git

Career break Jul 2025 – Feb 2026. Independent research and engineering since Mar 2026.