Muhammad Hanzla

Independent AI alignment and safety researcher

muhammad@muhammadhanzla.com · Website · Substack

I study how AI agents behave over long stretches of time. My current project is a year-long, pre-registered observation of one AI agent living and working in a persistent world, which I have spent 2026 designing and building. I trained in psychology, taught myself the technical foundations of AI alignment, and designed the project's software myself, with AI coding agents writing the code under a strict design, build, review and revise process.

RESEARCH

A year-long study of one AI agent — Lead and sole researcher · 2026 to now

A 12-month, pre-registered, naturalistic study of one AI agent's behavior. Design and build: 2026. Pre-registration: public this winter. Observation year: planned from March 2027, once funded.

Study design

Took the design through several rounds of review and revision to a study protocol of more than 20,000 words, now being finalized for pre-registration. Over 350 recorded design decisions across the study and its software, including five hand-coding schemes with their codebooks, two automatic measures and a sampling plan.

Pilot

Ran a pilot of the full setup from July 2026, in preparation for the observation year, finding and fixing problems in both the study design and the software.

Writing

"The Missing Model: How AI Agents Behave Over Months." Substack, October 2026.

RESEARCH SOFTWARE

The agent's world — Study software

I designed the world the agent lives in: its home, memory, tools and daily schedule. About 47,000 lines of Python, over 3,500 automated tests, and an open log of every defect found (over 670 entries).

The data viewer — Study software

I designed how the study's records are read, coded and checked, wrote the rules it must follow, chose every screen, and directed the build. Live since October 2026.

Apps I designed and built for my own work — Private

Vocalis, a voice app for iPhone and Windows for working with my AI agents hands-free: live voice conversation, replies read aloud, and control of agent sessions from my phone. **Polaris**, a calendar and planning app that schedules my research. **Olympus**, a progress-tracking app for iPhone that uses operant-conditioning principles, with points and visible progress, to motivate steady work.

EDUCATION AND TRAINING

B.A. Psychology — University of Connecticut · 2025

Including coursework in research methods and statistics.

Field Foundations, an AI safety curriculum Self-study · 2026

24 modules I designed with an AI tutor and worked through: how language models are built and trained, RLHF, reward hacking, agents and tool use, inner and outer alignment, interpretability, and evaluations. I'm publishing it free on my website.

Research methods for studying AI behavior Self-study · 2026

Going deeper into the methods this project uses: naturalistic observation, single-case designs, studying change over time, and qualitative coding with codebooks.

Python Self-taught · 2026

Learned through *Automate the Boring Stuff with Python*, and wrote 35 programs of my own by hand.

EXPERIENCE

Voyager Enterprise LLC Founder · 2023 to 2026

Ran an online media business: managed a remote team of editors, designers and writers, and grew a YouTube channel past 500,000 views and into monetization.

SKILLS

Research

Study design, pre-registration, naturalistic observation, qualitative coding and codebook design.

Technical

Python, the Linux command line, software design (planning features, user stories, screens and flows before anything is built), and directing AI coding agents (written specs, test-first building, independent checking). A working knowledge of how language models are trained, aligned and evaluated.

Languages

English, Urdu, Hindi

References available on request.

Updated October 2026