Last Updated on 17th December 2025
github.com/jaidhyani/resume

Jai Dhyani

jai.one | github.com/jaidhyani | jai@jai.one | it's pronounced "J"

I want the future to be full of human flourishing.
To that end I build things and bash them against reality to figure out why they don't work, and then I try to build something better.
Co-author on RE-Bench (METR's AI R&D time-horizon eval). Founder, Luthien Research.
Previously ML for hate speech at Facebook (the intersection of human values and AI is very hard).

LUTHIEN RESEARCH | Founder & Executive Director

2025 - Present | Technical AI Safety Nonprofit

- Founded nonprofit for runtime AI safety: control infrastructure for production agentic systems
- Building open-source LLM proxy (github.com/LuthienResearch/luthien-proxy) for Claude Code, Codex, etc.
- Streaming policy enforcement, real-time tool call intervention, LLM-as-judge evaluations
- Running user experience tests and iterating based on feedback

MODEL EVALUATION AND THREAT RESEARCH (METR) | Visiting Fellowship

2024 via ML Alignment and Theory Scholars (MATS) 6.0

- Co-author on ”RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts”
- Developed evaluations and tooling to detect AI research capabilities as part of Hjalmar Wijk's team

INDEPENDENT PROJECTS

2024: Delphi - Evals Technical Lead

- Led team building infrastructure for training and evaluating small, reproducible language models
- Rewrote training pipeline for improved reliability, speed, and configurability

2023 - 2024: Rational Animations Scriptwriter

- Scripts on AI alignment, safety, and existential risk; 2M+ views

TEMPLE CAPITAL | ML Engineering Consultant

2022

- Sped up quantitative investment analysis engine; added CI/CD pipeline
- Extended trading system to handle synthetic assets

BLUE ROSE RESEARCH | Senior ML Engineer

2021

- Accelerated TF-Probability variational inference pipelines for public opinion modeling
- Diagnosed and fixed slow database queries in analytical workflows

META (FACEBOOK) | ML Engineering for Hate Speech

2017 - 2020: Software Engineer - Machine Learning & Infrastructure

- Built Facebook's first proactive hate speech classifiers; integrated CV and multilingual NLP
- Built volume-adjusted thresholding system to optimize experts' time and maximize per-sample training value
- Discovered novel failure modes in training infrastructure; developed stratified offline datasets to prevent regression
- Trained and deployed a model that deleted the Declaration of Independence from Facebook on July 4th 2018, subsequently redoubling my respect for Goodhart's Law.

LENDINGROBOT | Senior Backend Engineer

2016 - 2017

AMAZON | Systems Engineer II, AWS Networking

2014 - 2015

DEMOCRATIC NATIONAL COMMITTEE | Linux Systems Engineering

2011 - 2012: Linux Systems Engineer | 2013: Lead Linux Systems Engineer

EDUCATION

University of Chicago | B.S. Computer Science | 2011