Last Updated on 16th December 2025
github.com/jaidhyani/resume

Jai Dhyani

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

ML engineer and AI safety researcher focused on evals and control infrastructure. Co-authored RE-Bench (METR's AI R&D evaluation for time-horizon estimates). Founder, Luthien Research. Formerly shipped production ML systems at Meta.

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 for Claude Code, Codex, and other Anthropic/OpenAI-compatible apps
- 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

- Optimized quantitative investment analysis engine; built CI/CD pipeline
- Expanded support for synthetic assets

BLUE ROSE RESEARCH | Senior ML Engineer

2021

- Optimized TF-Probability variational inference pipelines for public opinion modeling
- Optimized database queries to speed up time-consuming operations

META (FACEBOOK) | ML Engineering for Hate Speech

2017 - 2020: Software Engineer - Machine Learning & Infrastructure

- Built Facebook's first proactive hate speech classifiers, integrating computer vision and multi-lingual signals (English, French, Arabic, Portuguese)
- 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
- First-hand experience with high-profile AI failure: classifier removed the Declaration of Independence on July 4th, 2018

LENDINGROBOT | Senior Backend Engineer

2016 - 2017

AMAZON | Systems Engineer II, AWS Networking

2014 - 2015

DEMOCRATIC NATIONAL COMMITTEE | Lead Systems Engineer

2011 - 2013

EDUCATION

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