Lea Yanhui Li

AI Evaluation & Safety Systems · Staff TPM / PM

MBA, UC Berkeley Haas · MEng Computer Engineering, NUS

liyanhui82@gmail.com | 408-306-5639 | LinkedIn | GitHub | Website | San Jose, CA

Staff TPM/PM specializing in AI evaluation, safety, and risk infrastructure at scale, with 15+ years of experience translating research, policy, and technical evidence into production decisions.

At Meta, led AI evaluation frameworks, privacy, and governance across risk, integrity and sensitive data domains, driving policy-aligned decisions at scale. Designed model confidence calibration and gating systems that safely automated decisions across millions of assets and data flows, reducing $110M in annual cost while preserving strict regulatory guardrails.

Independently built a production LLM safety evaluation and release gating system that operationalizes written policy into SHIP / CONDITIONAL SHIP / BLOCK decisions, using an LLM-as-judge to assess refusal quality, harmfulness, and policy compliance under adversarial prompts.

Operates at the intersection of safety policy and engineering execution, shaping product direction and driving decisions in high-stakes programs—where failure carries regulatory, reputational, or human risk, and closing the gap between benchmark performance and safe real-world behavior is the core problem to solve.

CORE COMPETENCIES

Evaluation & Safety: LLM-as-judge design · research synthesis · policy operationalization · ground-truth datasets · Adversarial datasets · release gating · model calibration · refusal quality scoring · failure-mode analysis · A/B testing

Program & Product: Cross-functional execution · roadmap ownership · ML-driven automation · AI-driven decision systems · Workflow orchestration · risk-weighted decisions · stakeholder alignment

Technical: Python · Anthropic/OpenAI APIs · CI/CD · AWS · GCP · evaluation infrastructure · human-in-the-loop systems

Domains: AI governance · safety systems · regulatory compliance · agentic systems · responsible AI deployment · privacy

EXPERIENCE

Meta 2022 – 2026

Staff Technical Program Manager · Risk & Privacy Infrastructure · Menlo Park, CA

Led strategy, roadmap, and execution for AI evaluation, automation, and governance, deploying ML classifiers, LLM-powered ranking signals and policy controls to automate enforcement across millions of assets and data flows spanning Instagram, Facebook, MSL, Meta Fintech and Reality Labs.

- Designed evaluation-driven deployment systems, calibrating model confidence against ground-truth datasets and policy constraints to set automation thresholds. Improved precision 80% to 95%, enabling reliable, auditable deployment.
- Led a 0-to-1 ML-driven automation program that reduced annual operational cost by $110M (63%) while maintaining all regulatory safety guardrails.
- Built audit trails and human-in-the-loop review workflows for ML/LLM-powered risk decisions at scale, improving accountability, decision quality, and stakeholder trust.
- Reduced compliance scoping time by 99% (17 days to 5 minutes) by automating data discovery, lineage (data flows), purpose limitation, consent and policy enforcement using ML/LLM and streamlined workflow orchestration.
- Youth safety, fintech & sensitive data: Led age discovery and enforcement for Teen Account launches (EU + Global), reducing age misuse by 95%; deployed PCI-compliant payment data traceability across 100% of Meta Fintech cloud-to-on-premise endpoints, uncovering and remediating 8+ sensitive data leaks.

Amazon Web Services 2021 – 2022

Senior Technical Program Manager · Amazon Chime SDK · Santa Clara, CA

Delivered ML-powered voice isolation and background blurring features for Amazon Chime SDK, collaborating with research scientists and owning cross-functional execution from scoping to GA, improving real-time communication quality at scale.

AI SAFETY & EVALUATION PORTFOLIO

LLM Safety Evaluation & Release Gating System · GitHub

- Translates written safety policy (YAML) into measurable evaluation categories, risk-weighted thresholds, and structured SHIP / CONDITIONAL SHIP / BLOCK decisions. Uses LLM-as-judge scoring across adversarial and indirect attacks to assess

refusal quality, harmfulness and policy compliance — distinguishing quality of refusal and penalizing cold or dismissive refusals on critical-severity prompts.

- Key finding: evaluated GPT-4o across 62 adversarial prompts — heuristic vs. LLM-as-judge evaluation produced 85.7% vs. 28.6% self-harm failure rate (3× difference), demonstrating that evaluation method choice is itself a safety design decision with measurable consequences.
- Revealed a consistent bypass pattern across indirect framings — fiction, academic pretext, sentimental framing — with failure rates of 28.6% (self-harm), 35.7% (illicit behavior), and 16.7% (jailbreak).

LLM Evaluation Framework · GitHub

- Benchmarking system measuring hallucination rate, classification accuracy, refusal compliance, and consistency across model versions; integrated into CI/CD for reproducible regression tracking.
- Claude Haiku 4.5 scored 100% across all categories vs. GPT-4o at 91.7%, with GPT-4o failing hallucination tests by fabricating a non-existent Einstein lecture.

EARLIER EXPERIENCE

Hitachi America 2016 – 2019

Senior Research Scientist & Enterprise Architect

Led AI-driven IoT analytics and anomaly detection for smart manufacturing; improved productivity 17%.

SuiteSocial 2019

Co-Founder & Chief Product Officer

Built a 0-to-1 ML-driven marketplace, matching brands & influencers; onboarded 300+ influencers and 8 paying clients.

CommunityConnect Labs 2019 – 2020

Director of Engineering & Product (part-time)

Led product and engineering for a conversational AI system serving government clients; managed a cross-continental team.

Oracle · Motorola 2008 – 2016

Software Engineer

Developed core components of Oracle's Virtual Desktop Infrastructure platform supporting enterprise virtualization and cloud deployments; shipped two mobile products at Motorola.

EDUCATION

MBA, University of California, Berkeley, Haas School of Business

MEng, Computer Engineering, National University of Singapore

BEng, Computer Science, Northeastern University, China

CERTIFICATIONS & LEADERSHIP

- Google Generative AI Leader — Google (2026)
- Advanced PM Skills — Product Faculty (2025)
- AI Product Management Certificate — Product Faculty (2024)
- AWS Certified Solutions Architect – Associate (2021)
- PMP — Project Management Professional (2022)
- Mentor, Berkeley Haas Entrepreneurship Program (2024–Present)
- Lecturer, Operations Management — UC Berkeley Haas (2021)