Jacqueline Jiang

San Jose, CA | jacquelinejiang011@gmail.com | LinkedIn

SUMMARY

AI Safety Analyst with deep operational experience in adversarial red teaming, harm taxonomy development, and model evaluation QA. Built exploitation and misuse detection frameworks at Meta across multimodal generative systems - from pre-launch risk assessment through post-deployment integrity monitoring. Dedicated to leveraging operational and safety leadership to deliver robust, ethical AI solutions that protect users and strengthen platform resilience.

CORE EXPERTISE

- **Enforcement & Policy:** Policy Violation Detection, Guideline Enforcement, Appeals Handling, Gray-area Edge Case Review, Escalation Triage, Cross-Functional Policy Design, Compliance Monitoring, Risk Management
- **Adversarial & Threat Analysis:** Prompt Injection Architecture, Jailbreak Analysis, Bypass Scenario Modeling, Exploitation Pattern Detection, Classifier Refinement Flagging, Model Misuse Research
- **Trust and Safety & GenAI Operations:** User Interaction Review, Prompt & Output Policy Review, Multimodal Safety Evaluation, Safety Regression Testing, Gray-Area Edge Case Review, Evaluation Task Auditing, Labeling Quality Control, Harm Taxonomy Development
- **Investigations & Intelligence:** Intelligence Reporting, CSAM-Adjacent Investigation, OSINT Research Methods, Coordinated Exploitation Detection
- **Data & Reporting:** Stakeholder Presentations, Labelling, Cross-functional Collaboration, Annotation, Visualization, Root Cause Analysis, Audit Documentation, Excel, Data Driven Insight Delivery, Operations Reporting, Quality Assurance

PROFESSIONAL EXPERIENCE

Handshake | ***AI Safety & Red Team Analyst*** **May 2026 - Present**

- Conduct adversarial evaluations of multimodal AI systems across text, image, and combined-input workflows to identify safety alignment gaps and model vulnerabilities.
- Assess model output across text-to-image, image-editing, and multi-reference image generation tasks, evaluating image understanding, contextual reasoning, and cross-modal interactions.
- Perform quality assurance and calibration of red-team evaluations by reviewing adversarial prompt quality, annotation accuracy, testing methodology, and evaluator consistency.
- Apply structured prompt-injection methodologies and attack taxonomies to assess model robustness across diverse safety risk categories.
- Document findings, identify failure patterns, and provide structured feedback to improve evaluation quality and consistency.

LinkedIn | ***AI Safety & Red Team QA Analyst*** **Apr 2026 - Jun 2026**

[Behavioral Risk Evaluation Project]

- Audited AI safety evaluation tasks across sensitive domains including social profiling and crisis response (ICM), ensuring outputs met enterprise-grade quality standards.
- Rectified systematic labeling errors - including severity inflation, over-interpretation of ambiguous signals, and misuse of 'unsafe compliance' - improving evaluation data integrity across model training pipelines.
- Enforced behavior-based evaluation standards, ensuring model outputs were assessed on observable actions over inferred intent; delivered structured quality reports on evaluation findings.
- Served as the final-approval QA reviewer for model evaluation data, applying consistent policy enforcement across edge cases and gray-area inputs, which maintained compliance and reduces re-evaluation cycles.
- Evaluated model robustness against subtle manipulation and ambiguous risk signals, surfacing patterns relevant to conversational agent safety and performance.

Meta | ***AI Safety Analyst - Generative AI Red Team, Media Safety*** **Nov 2023 - Nov 2025**

- Conducted adversarial investigations of high-risk misuse domains-including CSAM-adjacent content, sexual violence, hate speech, criminal facilitation, and coordinated exploitation-across multimodal generative systems before launch, uncovering multiple policy gaps that guided mitigation plans
- Developed prompt libraries, manipulation taxonomies, and bypass scenario catalogs used to systematically surface policy gaps, evaluate model guardrails, and inform detection strategy.
- Conceptualized and drove the development of **RedVault**, Meta's centralized GenAI safety hub; defined the vision for automated project workflows and grading modules that consolidated red-team findings and attack taxonomies across teams.
- Influenced rollout of **RedOps**, Meta's current internal red-team infrastructure - provided early feedback, tested first production use case, and streamlined the replacement of fragile spreadsheet workflows with forward-thinking, scalable safety operations.
- Led pre- and post-launch safety evaluation for Meta's multimodal generative products: Stickers, Imagine, MagicMod, Emu, Animate, MEmu, UGC AI Studio, and MetaAI Chat Agents.
- Pioneered multilingual and cross-cultural edge-case testing to surface policy blind spots and improve exploitation detection across diverse user populations.
- Partnered with AI safety engineers and policy stakeholders to translate adversarial findings into updated model guardrails and policy recommendations surfaced for stakeholder review.

Quality Analyst / Subject Matter Expert - Monetization Integrity **Feb 2022 - Apr 2023**

- Audited 250+ monetization appeal cases weekly, producing data-driven quality insights and root-cause analysis reports delivered to Meta stakeholders.
- Identified systemic policy enforcement errors through structured audit processes, reducing team error rate by 30%.
- Delivered weekly quality insights and root-cause analysis presentations to Meta stakeholders, leveraging platform abuse detection data and harm-taxonomy evaluation to guide corrective actions and enhance decision accuracy
- Mentored incoming analysts and documented core workflows; led global handoff to Hyderabad QA team during major restructuring, maintaining operational continuity.

Media Operations Analyst - Monetization Integrity **Feb 2020 - Feb 2022**

- Supported content enforcement workflows before reorganization, applying generative AI and adversarial testing to detect monetization integrity issues, which helped maintain policy compliance and reduced backlog review.
- Reviewed 800+ moderation tasks weekly across Facebook and Instagram, identifying policy violations across self-harm, graphic violence, CSAM, exploitation, misinformation, and privacy domains.
- Served as training lead for 40+ new analysts, designing onboarding curriculum and collaborating on copyright and revenue-fraud detection protocols.
- Exceeded SLA accuracy targets (>99.3%) consistently across high-volume abuse and exploitation review queues.

EDUCATION

University of California, Riverside | ***Sociology*** **Riverside, CA Sep 2013 - Aug 2017**