CALISSA MAN

linkedin.com/in/calissa-man | calissaman@gmail.com | +65 9776 9978

PROFESSIONAL SUMMARY

AI safety evaluation specialist with 4+ years of experience developing and operationalising safety guardrails, content policies and LLM-driven content moderation for multimodal products and social media platforms. Deep experience in building user safety systems from 0 to 1 and using LLMs for both policy drafting and enforcement across **multimodal and generative AI products (video, image and virtual assistants). Eligible for H1B1 visa, which is exempt from H1-B fees and restrictions.**

EXPERIENCE

ByteDance Seed Lab | Safety Policy Research Analyst **Aug 2025 – Feb 2026**

- **Designed and deployed evaluation suites and LLM-driven policy enforcement systems across multimodal and generative video products,** informing global launches and product delivery for video models, including Seedance.
- **Developed safety evaluation frameworks and product-level guardrails based on red-teaming, safety research and model evaluations** across video, image, coding and chatbot AI applications, covering both P0 and edge cases.
- **Built and scaled safety review and feature assessment systems before launch from 0 to 1, improving precision by 8x** and pioneering safer prompt responses in models released.
- Partnered cross-functionally with product, legal and engineering to **monitor AI safety research trends** and iterate **new evaluations/guardrails** for future model launches.

Singapore's Online Safety Commission (IMDA) | Assistant Manager, Policy Planning **Jun 2024 – Jun 2025**

- **Evaluated severity and prevalence of user safety risks across 50+ AI video, image and avatar-based applications,** informing regulatory escalations and national standards for AI consumer applications.
- **Built policy frameworks and enforcement guidelines for AI consumer applications and AI-enabled social media,** as a subject matter expert on LLM-driven policy development/enforcement and new use cases.
- **Led triage and analysis of high-severity AI-generated mental health incidents,** including risks affecting minors, coordinating across policy, legal, and operational teams from mitigation to content review.

(US startup incubator) Plug & Play Ventures | Partner Success Associate **Nov 2023 – May 2024**

- **Advised early-stage consumer AI startups** on safety, deployment and governance, by building trust and safety policies and enforcement processes from scratch while building actionable guardrails in line with user acquisition and regulatory constraints.
- **Led partnerships with Fortune 500 companies and governments** on AI safety investment programs worth **USD 1.5M**.

Vriens & Partners | Government Affairs & Public Policy Analyst **Jun 2022 – Oct 2023**

- **Advised YouTube, Google, Meta, and TikTok on AI-related user safety concerns**, including harassment, mental health risks, synthetic media misuse, and coordinated abuse.
- **Reviewed enforcement policies and escalation frameworks**, translating legal and policy requirements into clear operational decision criteria for escalation, review and appeals.

TikTok | Country Policy Intern (Trust & Safety, Product Policy) **Nov 2021 – May 2022**

- **Improved policy enforcement accuracy by 20%** by setting localized abuse and safety frameworks across five P0/P1 markets during elections and conflict periods, with LLM-driven policy enforcement.

RESEARCH & PUBLICATIONS

Co-author, AgentChangeBench: A Multi-Dimensional Evaluation Framework for Goal-Shift Robustness

- **Accepted to 2 NEURIPS 2025 workshops** on multi-turn LLM interactions (LAW 2025, MTI-LLM 2025). Developed a multi-turn evaluation benchmark analyzing how models respond to changing user intent, informing understanding of model behaviour, maintaining user safety and misuse adaptation.

World Economic Forum – Global Shapers | Vice President, Trust & Safety Lead Oct 2021 – May 2024

- Led the first national research study on **mental health harms from AI tools and platforms,** advised Meta, the Singaporean government and World Economic Forum on mental health risk mitigation strategies.

EDUCATION

National University of Singapore (NUS) - *Bachelor of Arts (Political Science) with First Class Honours (Highest Distinction)*

- Cumulative Average Point: 4.93/5.00 (Top graduate of the Arts Faculty)

SKILLS

- **Languages:** English, Mandarin, Bahasa Indonesia.
- **Technical skills:** Python, Microsoft Office, Google Data Analytics.