Jasmine Li

jasminexinzeli@gmail.com | linkedin.com/jasminexli/ | github.com/jasmineee-li| jasminexli.com

EDUCATION

Cornell University, College of Arts & Sciences — Ithaca, NY

B.A. Computer Science & English — *Expected Dec 2026*

Awards: Constellation Visiting Fellowship (Fall 2026), CBTA Telluride Scholar (5-year full scholarship), Invisible College Fellow, Milstein Scholar in Technology & Humanity, Non-Trivial Scholar, Foresight Institute AI Node Grantee ($40k), Cosmos Institute Fellow

RESEARCH INTERESTS

AI evaluations and automated auditing; misalignment incident investigations; compute verification; institutions and standards for international AI coordination; China and AI governance.

EXPERIENCE

SaferAI — London, UK

Research Affiliate — *Aug 2026 - present*

- Leading secure evaluation & research transparency technical pilot.
- Seeding Asia-based forum to promote Sino-Western coordination on auditing best practices and standards.

Research Scientist — *Jun 2026 – Aug 2026*

- Developed safety evaluations and mitigation recommendations for 2 Chinese AI companies in context of EU AI Act GPAI CoP compliance. Led Chinese-language stakeholder engagement. Coauthored GLM-5.2 risk report.

Safe AI Forum — London, UK

Visiting Researcher — *July 2026 – Aug 2026*

- Governance research & direct fieldbuilding for compute verification in China.
- Scoped a competition to advance cryptography for ML (proof of useful work); supporting research collaborations between Eastern & Western researchers in hardware security.

ML Alignment & Theory Scholars 9.0/9.1 — Berkeley, CA

Research Fellow, co-mentored by Alex Turner (Google DeepMind) & Alex Cloud (Anthropic) — *Jan 2026 – May 2026*

- Developed *cooperation training*, a scalable posttraining mitigation to LLM evaluation gaming in safety evaluations.
- Research highlighted in blog post, *The Information* article. Winner of 2026 Corrigibility Prize.

Gray Swan AI — Pittsburgh, PA

ML Engineering Intern — *May 2025 – Dec 2025*

- Designed evals for agent reasonable care & legal liability, used in pre-deployment Grok-4 & GPT-5 testing.
- Led development of WARP, an automated adversarial framework for browser-use agent redteaming, achieving 80% attack success rate on Opus-4.7. For use in Gray Swan production; **AI4GOOD / AIWILD, ICML**.
- Led the first comprehensive study of model evaluation awareness in browser-use agent safety evaluations.

Center for AI Safety — San Francisco, CA

Research Collaborator — *Feb 2025 – Nov 2025*

- Methods development fo generalized honesty in LLMs. Iterated on various LoRRA, SFT, DPO, and GRPO approaches; best-performing method improved P(honesty) on the MASK benchmark by over 20%.
- Investigating ability & propensity of LLM agents to manipulate reward during online RL training. Helped set up tool calling and RL test runs with verl. (July - present).
- Worked on formalization for calibrated uncertainty in utility functions for inverse RL assistance games (Nov - Mar).

Peking University Alignment and Interaction Lab — Beijing, China

Visiting Researcher — *Jan 2024 – May 2024*

- Developed a benchmark framework and dataset, with 18 LLMs trained on 9 centuries of historical texts, to evaluate how alignment algorithms track evolving human values. **Spotlight at NeurIPS D&B '24**.
- Engineered a large-scale data pipeline to acquire and process terabytes of online text data; analyzed value embeddings revealing temporal trends in religiosity and liberalism.

Lionel Levine Lab — Ithaca, NY

Student Researcher — *Jan 2025 – May 2025*

- Co-developed EigenBench, a PageRank-based arena for assessing constitution-based dispositional alignment in LLMs. **ICLR Oral 2026**.

OTHER EXPERIENCE

Cornell AI Alignment Ithaca, NY
Founder, Advisor *Jan 2025 – present*

- Led two semesters of introductory AI safety fellowship, graduate-level technical safety reading groups, & AI
- Designed Cornell's first course on AI safety, CS 1998: Intro to AI Safety & Alignment.
- Scaled club to 50 active members and online community of 250 before handoff.

Babel Translation aistranslate.org Berkeley, CA
Founder *Sep 2026 – present*

- Translating AI safety writing into Chinese and Spanish, backed by Astralis Foundation.

Ellis Brooklyn, NY
Software Engineering Intern *Jun 2024 – Aug 2024*

- Led end-to-end development of internal job-finding database with advanced data viz and optimized data handling.
- Developed a data pipeline to integrate government data and external APIs; enhanced PostgreSQL performance via normalization, indexing, partitioning, and scheduled API requests.
- Directed product syncs and investor presentations, authored product specs, and co-designed wireframes.

SELECTED PUBLICATIONS

1. Casper, S., Stray, J., Gausen, A., Jones, C., Christian, B., **Li, J.**, Bengio, Y., Rand, D., et al. **AI Epistemic Risks: Emerging Mechanisms & Evidence.** *SSRN*, (2026). [SSRN]
2. Dixit, C., Davies, J., **Li, J.**, Bruckamp, B., Kengott, J., Papadatos, H. **GLM-5.2 Risk Evaluation Report.** *SaferAI*, (2026). [Report]
3. **Li, J.X.**, Chew, A., Lin, M., Jones, E., Zou, A. **WARP: Measuring and Mitigating Evaluation Awareness in Browser Agent Safety Benchmarks.** *ICML Agents in the Wild Workshop*, (2026).
4. **Li, J.X.**, Turner, A. **Eval Cooperativeness Mitigates Evaluation Gaming in LLMs.** *ICML AI4Good Workshop*, (2026).
5. Chang, J., Piff, L., Sana, S., **Li, J.X.**, et al. **EigenBench: A Comparative Behavioral Measure of Value Alignment.** *ICLR (Oral)*, (2026). arxiv.org/abs/2509.01938
6. Qiu, T., Zhang, Y., Huang, Z., **Li, J.X.**, et al. **ProgressGym: Alignment with a Millennium of Moral Progress.** *NeurIPS (Spotlight)*, (2024). arxiv.org/pdf/2406.20087
7. Chooi, J.Q.*, **Li, J.X.***, Lee, J.* **Measuring AI-Induced Disempowerment: A Framework and Proposed Metrics.** *ACL EvalEval Workshop*, (2026). *Equal contribution.
8. Niu, J., Liu, W., **Li, J.X.**, et al. **Machining water through laser cutting of nanoparticle-encased water pancakes.** *Nature Communications* 14, 3853 (2023). doi:10.1038/s41467-023-39574-3

SERVICE & INTERESTS

Speaking	Speaker, Harvard Kennedy School US-China AI Dialogue (Dec 2026) Speaker, Foresight SF Vision Weekend (Nov 2026) Speaker & Judge, Apart Global South AI Hackathon (Jun 2026)
Service	Expert Judge, Apart Verification Grand Challenge (forthcoming) Expert Technical Consultant, MATS
Mentorship	Mentor, Frontier AI Security Residency Mentor, Pathfinder
Peer Review	IASEAI 2027 ICML 2026 Workshops: AIWILD, AI4GOOD, TAIGR
Research Interests	Science of evaluation and auditing; compute verification; global AI governance institutions; US-China coordination.
Other Interests	Writing at The J-Space; progress and economics; backpacking and adventuring; hosting parties.