GABRIEL HUANG

Research Scientist — Agent Safety, Post-Training & Evaluation

gbxhuang@gmail.com | 438 924 4894 | Montréal, QC — Canadian PR | gabrielhuang.github.io | gabrielhuang
gbxhuang | _X1TI-MAAAAJ | FR native • EN bilingual • ZH fluent (everyday)

CAPABILITIES

Agent-security red-teaming & threat modeling. Design attack libraries, threat compositions, and adaptive red-teaming strategies for production-style LLM agents.

- **DoomArena** (COLM 2025): developed prompt-injection attacks on BrowserGym-based agents; ran experiments quantifying the utility-security tradeoff; showed SOTA LLMs of the time were easily compromised under realistic injection scenarios; evaluated LlamaGuard- and LLM-as-a-Judge-style defenses — as architect, also designed the plug-in *AttackGateway* abstraction enabling attacks across heterogeneous benchmarks.
- **6th place** at Microsoft LLMail-Inject international prompt-injection challenge (2025).
- Co-designed a 3-stage adaptive attack cascade (direct injection → semantic deception → model-specific adaptation) driving rigorous firewall evaluation. *Evidenced by: Prompt-Injection Firewall paper.*
- Co-designed and empirically evaluated a self-replicating multi-agent worm in a 30-VM isolated corporate network (heterogeneous Linux/Windows, 20+ CVE- and CWE-class vulnerabilities); worm autonomously compromised ∼10 machines within 48 hours via an 8-phase attack lifecycle and 3-tier hierarchical memory system enabling coherent state across dozens of steps. *Evidenced by: LLM Worm (to be submitted, Nature).*

Evaluation design & evaluator-quality management. Design end-to-end evaluation pipelines for LLM agents: rubric design, annotator workflow, LLM-as-a-Judge calibration, and joint reporting of utility vs. safety trade-offs.

- Co-designed **GREADTH** — the binary six-criteria human-rating rubric used on conversational agents (Grounded, Responsive, Accurate, Disciplined, Transparent, Helpful), derived from HHH to make inter-rater agreement tractable. *Evidenced by: TapeAgents (co-authored, core contributor).*
- Calibrated GPT-4o as an LLM-as-a-Judge against lightweight guardrail classifiers (LlamaGuard) inside the agent-environment loop. *Evidenced by: DoomArena, COLM 2025 (co-authored, joint primary contribution).*
- Jointly reported Benign Utility / Utility-Under-Attack / Attack Success Rate across 4 agent-security benchmarks; systematically critiqued benchmark metric flaws. *Evidenced by: Prompt-Injection Firewall paper (co-authored, under review COLM 2026).*

Synthetic data generation & annotator tooling. Build the pipelines and tools that turn non-technical annotators into high-throughput data contributors, and turn agent execution traces into training data.

- Built the **dialogue-authoring tool** that bootstrapped TapeAgents' early models: intuitive UI, fast iteration cycles, measurable reduction in non-technical-linguist error rates — critical when zero-shot agent capabilities weren't good enough to self-bootstrap.
- Co-designed synthetic dialogue generation over realistic simulated users, company APIs, and policies; used for both training and rigorous evaluation.

Post-training, distillation & agent systems. SFT and full RLHF loops on grounded tool-calling agents; framework and abstraction design that reduces research iteration cost.

- Co-developed teacher-student distillation pipeline: expensive multi-node Teacher agent → Llama-3.1-8B Student matching GPT-4o utility on tool-calling tasks at *orders-of-magnitude lower inference cost* (∼300× at our operating point). *Evidenced by: TapeAgents.*
- Ran full RLHF loops combining human labelers and LLM-as-a-Judge on conversational enterprise agents.

PROFESSIONAL EXPERIENCE

Senior Research Scientist

ServiceNow Research

Oct 2022 – Present | Montréal, QC

- Built red-teaming frameworks, evaluation pipelines, annotator tooling, and teacher-student distillation for production-style LLM agents — from attack design through post-training and deployment-readiness evaluation.
- Joint primary contributor on **DoomArena** (COLM 2025) and **TapeAgents**; co-author on an indirect prompt-injection firewall paper (under review, COLM 2026).
- Supervise five graduate researchers (McGill, UdeM, Guelph) through to co-authored publications on agent security, evaluation, and test-time adaptation.

PhD Research Intern — Multimodal Pretraining

Google Research

Sept 2019 – Jun 2020 | Mountain View & Montréal

Led multimodal pretraining research for dense video captioning on YouTube-scale instructional video.

- Processed **terabyte-scale** video corpora (HowTo-100M, YouTube-8M, WikiHow, Recipe-1M, YouCook II): joint video + ASR feature extraction pipelines.
- Hand-tuned a **multimodal transformer architecture** and adapted self-supervised pretraining objectives to video+text, then fine-tuned on captioning targets.
- **Beat state-of-the-art** on dense video captioning at publication time. *AACL-IJCNLP 2020 (first author).*

PhD Research Intern — Few-Shot & Self-Supervised Learning

ServiceNow Research / Element AI

2021 – 2022 | Montréal, QC

- Self-supervised and few-shot object detection (project lead). ⇒ *IEEE TPAMI 2022.*
- Co-developed **Geo-Bench**, a benchmark for Earth-observation foundation models. ⇒ *NeurIPS 2023 (Datasets & Benchmarks).*

Chief Data Scientist & Co-founder

Patterns Technologies

2017 – 2018 | Montréal, QC

- Co-founded an early-stage ML startup; shipped customer-facing object-detection APIs on GCP for enterprise clients.
- Built NLP digest-generation pipelines from on-site user reviews.

Research Intern — Computer Vision & Multimodal Perception

Dassault Systèmes

2013 – 2016 | Paris, France

Multimodal perception from Microsoft Kinect RGB-D sensors (human activity recognition); content-recommendation algorithms; implemented and parallelized supervised / unsupervised ML algorithms in C++ and CUDA.

EDUCATION

Ph.D. in Machine Learning
Mila, Université de Montréal — Advisor: Simon Lacoste-Julien — 2017 – 2022 | Montréal, QC

M.Sc. Mathématiques, Vision & Apprentissage (MVA)
École Normale Supérieure, Paris — 2015 – 2016 | Paris, France

Visiting Student — supervisor: Brian Kernighan
Princeton University — Spring 2015 | Princeton, USA

M.Sc. in Applied Mathematics & Statistics
CentraleSupélec — ranked 1st of 500 graduates — 2013 – 2016 | Paris, France

SELECTED PUBLICATIONS

Full list: Google Scholar.

- **An Adaptive, Self-Replicating LLM-Powered Computer Worm.** Guan, Blanchard, Jia, Foerster, **Huang**, Papernot. *Submitted to Nature.* (Core contributor — agent architecture and harness engineering; ServiceNow Research × University of Toronto.)
- **DoomArena: A Framework for Testing AI Agents Against Evolving Security Threats.** Boisvert, Bansal, Evuru, **Huang**, Puri, Bose, Fazel, Cappart, Stanley, Lacoste et al. *COLM 2025.* (Joint primary contribution.)
- **Indirect Prompt Injections: Are Firewalls All You Need, or Stronger Benchmarks?** Bhagwatkar, Kasa, Puri, **Huang**, Rish, Taylor, Dvijotham, Lacoste. *Under review, COLM 2026.* (Co-author.)
- **TapeAgents: A Holistic Framework for Agent Development and Optimization.** Bahdanau, Gontier, **Huang**, Kamalloo, Pardinas, Piché, Scholak, Shliazhko, Tremblay, Ghanem et al. *arXiv 2412.08445, 2024.* (Core contributor.)
- **Geo-Bench: Toward Foundation Models for Earth Monitoring.** Lacoste, Lehmann, Sherwin, Guzman, Sherif, Ellsworth, Benabou, Gomes, Maldague, Lucas, Dao, Steinegger, **Huang**, Lütjens. *NeurIPS 2023 (Datasets & Benchmarks).*
- **A Survey of Self-Supervised and Few-Shot Object Detection. Huang**, Laradji, Vazquez, Lacoste-Julien, Rodriguez. *IEEE TPAMI, 2022.*
- **Multimodal Pretraining for Dense Video Captioning. Huang**, Pang, Zhu, Rivera, Soricut. *AACL-IJCNLP 2020.* (First author; Google Research.)

AWARDS, SERVICE, SKILLS

Awards. 6th — Microsoft LLMail-Inject Prompt Injection Challenge (2025, international). Winner — McGill AI Hackathon (2017). Winner — Dassault Aviation UAV Challenge (2015).

Reviewing. ICLR (2018–now), NeurIPS (2019–now), ICML (2021–now), ECCV (2016–now).

Teaching. Probabilistic Graphical Models TA, Université de Montréal (2018–2021).

Technical. Python, PyTorch, HuggingFace stack; distributed LLM training (DeepSpeed, Accelerate, LLaMA-Factory); filtered SFT / reward-weighted regression / RLHF loops (human + LLM-as-a-Judge feedback); LLM-as-a-Judge and rubric-based human eval; synthetic data pipelines; agent frameworks (TapeAgents, BrowserGym); red-teaming / prompt-injection tooling; penetration testing workflows (recon, foothold, privesc, lateral movement); corporate network simulation (KVM/QEMU, heterogeneous multi-VM environments); multi-agent coordination architectures.