Alejandro Cuadron

Website	alexcuadron.github.io	**Email**	alex.cl.2000@gmail.com
Scholar	Google Scholar	**GitHub**	github.com/AlexCuadron
LinkedIn	linkedin.com/in/acuadron	**X**	@Alex_Cuadron

Experience

Amazon AGI | Applied Scientist II — Feb 2026 – Present

- Built evals exposing agent limits, showing coding agents solve 95% of tasks alone but 51% chained, published as BOBA-Bench.
- Built RSI pipelines whose tasks lifted Qwen3.6-35B-A3B above Opus 4.8 on Vending-Bench, an economic agent benchmark.
- Ran a 200-agent factory that turned 41K repos into 7,000+ 20-step coding chains that teach models not to regress prior work.
- Built an agentic audit that scanned 378 RL environments (55K+ tasks) for reward hacks and task correctness.
- Led Amazon AGI's first asynchronous, long-context agentic RL post-training of a 300B+ MoE model on hundreds of GPUs.

Amazon AGI | Applied Scientist I — Jul 2025 – Jan 2026

- Found errors in τ^2-bench (used by OpenAI, Anthropic, Google, Meta) and built τ^2-Bench-Verified, whose fixes are in τ^3-bench.
- Contributed RL environments and thousands of SFT tool-use trajectories to Nova 2 Lite, matching Sonnet 4.5 on τ^2-Bench.
- Built a framework that verifies agents' state-changing actions and cleans stale context, raising task success up to 28% (SABER).

UC Berkeley, Sky Computing Lab | Research Scholar with Ion Stoica and Joey Gonzalez — Jan 2024 – Jun 2025

- Showed in an extensive study of o1-style models as coding agents that they overthink instead of acting (paper, 180+ citations), findings adopted by Meta, Microsoft, CMU, Berkeley and others and covered by The Decoder and IEEE Spectrum.
- Co-created learned token hashing for long-context agents (32× fewer tokens) and led its Llama training (HashAttention, ICML).
- Developed error-guaranteed sparse attention and LLM caching for agent inference, published at ICLR as vAttention and vCache.
- Built a coding split of right and wrong solutions that strong LLM judges barely tell apart, published as JudgeBench at ICLR.

Publications

The Danger of Overthinking: the Reasoning-Action Dilemma in Agentic Tasks	189 citations	arXiv 2025
JudgeBench: A Benchmark for Evaluating LLM-Based Judges	465 citations	ICLR 2025
vCache: Verified Semantic Prompt Caching	49 citations	ICLR 2026
HashAttention: Semantic Sparsity for Faster Inference	37 citations	ICML 2025
SABER: Safeguarding Mutating Steps in LLM Agents	31 citations	ICLR 2026 W
vAttention: Verified Sparse Attention	14 citations	ICLR 2026
Amazon Nova 2: Multimodal Reasoning and Generation Models	14 citations	Tech. report
BOBA: Why Coding Agents Cannot Be Left Alone		Under review

Projects

Open Kernel Challenge (Meta × Paradigm) | 5th of 2,000+ teams — 2026

- Built an RSI framework that writes and optimizes GPU kernels and placed 5th with no prior CUDA experience.
- Turned it into a GPU-kernel training environment that teaches models to write fast kernels out of the box.

OpenHands | Open-source coding agent, 90k+ GitHub stars — 2024 – 2025

- Built the multi-agent setting and o1 support (native function calling, reasoning effort).

Education

University of California, Berkeley | Graduate studies, Computer Science — 2024 – 2025

ETH Zürich | M.Sc. Computer Science, Secure and Reliable Systems — 2022 – 2024

Universidad Politécnica de Madrid | B.Eng. Computer Engineering, top 5% of class — 2018 – 2022

Honors & Awards

- Received 789 citations on Google Scholar in the last two years and was interviewed by IEEE Spectrum (2025).
- Invited reviewer for NeurIPS 2026, ICLR 2026 and ICLR 2027.
- Member of IEEE-Eta Kappa Nu (HKN) and founding president of Spain's first HKN chapter.
- Excellence scholarships from Fundación Mutua Madrileña and the Government of Madrid.