

Lazaro Kawer

kawer.dev | kawer7@gmail.com | github.com/kawer | linkedin.com/in/lko

EXPERIENCE

Microsoft - Windows Fundamentals | 2021-Present

Senior Research Engineer | AI Evaluation Lead, 2025-Present

Senior SWE 2024-Present | SWE II 2022-2024 | SWE 2021-2022

Technical lead building production AI, agentic workflows, predictive ML, and fleet-scale decision systems.

AI OPTIMIZATION AND AGENT EVALUATION

Technical lead for a production AI engineering platform spanning performance, power, and memory. Our platform has delivered hundreds of confirmed bug fixes and optimizations across Microsoft and generated tens of millions of dollars in yearly savings for partner OEMs.

- **Lead cross-vertical AI evaluation; designed and built the Python evaluation pipeline.** Integrated calibrated, source-grounded judges, release gates and self-service tooling. Developer fix acceptance rose from 54% to 78% after the evaluation framework was introduced.
- **Launched a self-service agentic evaluation harness for prompt, context and tool-use comparisons, cutting prompt iteration from 2+ weeks to 1-2 days.** Adopted by all vertical owners across performance, memory, disk and power; graded feedback drives controlled iteration, regression checks and human-reviewed release.
- **Drove production-model selection for bug detection and code optimization; the model change increased developer fix acceptance by 18%.** Compared 6 frontier-model configurations using source-grounded paired non-inferiority gates and cross-family bias controls; identified the only candidate passing all predeclared quality gates.
- **Architect and technical lead for a generalized evaluation platform for agent harnesses and skills.** Implemented a grader-led revision and re-evaluation loop for skills and harnesses, improving accuracy and execution speed across internal performance-engineering workflows.
- **Doubled addressable fix scope by building a self-improving reverse-engineering harness** for third-party binaries and OS components without source access.

WINDOWS AI PRODUCTS AND ON-DEVICE ML

- **Built and internally deployed an on-device SLM across ~3K Windows devices.** Predicted 71% of targeted performance and memory failures at least 20 seconds ahead and 92% at least 5 seconds ahead; scheduler interventions mitigated ~16% of targeted failures before onset. Internal builds only.
- **Owned a distributed evaluation platform for 8 AI-powered Windows features** across ~11K internal devices and hundreds of lab machines, catching 3 critical regressions before retail release.

FLEET-SCALE DATA AND DECISION SYSTEMS

- **Created Glitchiness, an OS-wide perceived-responsiveness metric guiding Windows performance priorities and reported at CVP level.** Built its perceptual algorithm and sampled-telemetry pipeline; validated the metric through controlled and field studies.
- **Built forecasting and regression-attribution systems used by leadership** for fleet-wide quality tracking, population-wide regression attribution, and sales forecasting in partnership with sales and marketing. Device-population forecasts backtested at $R^2=0.90$ across active silicon.
- **Built the performance decision platform senior leadership uses for release decisions and next-generation hardware and component planning.** Replaced weeks of status reviews with automated rankings and a Next.js UI for regressions, improvements, and release blockers; normalized internal and partner feeds, cutting platform-caused failures from ~2/day to effectively none.

Apple - Siri Search | Summer 2019

AI/ML Engineer Intern

- Built recommender and data systems that increased Siri Knowledge Card entity coverage 214%, positive abandonment 35%, and engagement 7%; delivered a Hive/Impala refresh pipeline and validated outcomes through production A/B tests.

Google - Search and Cloud | Summers 2017, 2018

Software Engineering Intern

- Built Spanner-backed feedback infrastructure and real-time quality dashboards for Search; built a multithreaded gRPC load generator and scheduler for Cloud, increasing network performance 31%.

SKILLS

ML and evaluation: TensorFlow, on-device transformers, forecasting, production A/B testing.

Systems: Python (primary), Linux/Unix/macOS, C++, C#/ .NET, TypeScript, SQL, APIs, distributed systems, Kusto/ADX, Cosmos/Scope, Azure.

Agent systems: MCP integration, prompt/context workflows, GitHub Copilot CLI, Codex, immutable lineage, fail-closed recovery.

EDUCATION

ITESM Monterrey | B.S.E. Computer Science | GPA 4.0 | June 2020