Lazaro Kawer

kawer.github.io · kawer7@gmail.com · github.com/kawer · linkedin.com/in/lko

ITESM Monterrey
B.S.E. Computer Science · GPA 4.0 · June 2020
Summa Cum Laude · Valedictorian

EXPERIENCE

Microsoft · Windows Fundamentals

Senior Software Engineer · Tech Lead, Cross-Vertical AI Evaluation *2021–Present*

AI EVALS FOR WINDOWS OPTIMIZER

- **Built the evaluation layer for a production frontier-model code-optimization system** that analyzes Windows perf, power, and memory bottlenecks. My framework evaluates whether model outputs are correct, grounded, actionable, and improving over time, on reproducible replay infrastructure varying one axis at a time (model, prompt, advice corpus, supporting data). Fix-acceptance climbed from 54% to 78%; same harness scales across Perf, Power, and Memory.
- **Found real failure modes in production AI outputs** and built the calibrated LLM-as-judge layer that cuts eval false positives ~35% and false negatives ~50% against a human-expert gold standard. The judges surface sycophantic agreement with prior production wins, hallucination patterns, prompt-template drift, and reward-hacking against the scorers themselves: alignment-relevant failure modes that look correct until acted on.
- **Designed evals and statistical methodology for production model behaviors.** Scorers detect source-grounding errors, hallucinated APIs, unsupported mechanism and magnitude claims, classification inconsistencies, and overconfident recommendations. Comparisons use paired tests, bootstrap CIs, McNemar, Cohen's kappa, and Simpson's-paradox-aware aggregation, supporting model migration across Claude Opus (Anthropic SDK) and Azure-hosted GPT/o-series.
- **Defined the bi-directional taxonomy and feedback loop for AI-generated Windows optimizations.** Authored a taxonomy across algorithmic, caching, concurrency, I/O, memory, and anti-pattern classes; models use it for grounding, and novel model-proposed patterns route through validation gates before being admitted back into the corpus, becoming a growing library of optimizations backed by shipped-fix evidence; the loop has surfaced optimization classes (rare or cross-component) that subject-matter experts had not catalogued. Architected the end-to-end loop, from telemetry-driven bug discovery through prompt assembly, model analysis, evaluation, code-change generation, automated PR, and regression testing, that was adopted as the team's technical direction.

WINDOWS GLITCHINESS METRIC

- **Designed and shipped Glitchiness, a new metric for perceived Windows responsiveness, validated with human-subjects studies.** Delay-signal end-time clustering with span-based duration and long-delay promotion, deployed at Windows scale with daily failure-rate CoV 4.8%. A 41-participant controlled perceptual study plus an n=2,234 in-the-wild user-feedback study show Glitchiness flags user-felt sluggishness ~60% more often than the prior signal I had built (itself the first attempt in the org to measure this). The methodology of operationalizing an inherently fuzzy human-judgment concept into a statistically-validated measurement is directly applicable to alignment concepts like helpfulness, honesty, and harmlessness.
- **Connected memory pressure to perceived OS responsiveness at fleet scale and surfaced a structural law that is reshaping memory-management strategy.** Memory-pressured 8 GB devices produce +33.7% more glitchiness than peers (Z = +28.3), reaching +231% for users with higher device interactivity. Followup analysis surfaced a clean, consistent knee in the failure-rate-vs-commit curve at every RAM tier, with knee location a simple multiplicative function of installed RAM, reproducing across CPU families and populations. I am leading the Glitchiness side of the resulting cross-org effort with memory and hardware experts; the finding drives executive-level memory-consumption strategy and informs next-generation Windows hardware planning.

FLEET TELEMETRY / DATA ENGINEERING

- **Rearchitected Windows performance telemetry infrastructure, ~4× device coverage** (hundreds of thousands to several million devices/day) after finding a measurement blind spot, multiplying the signal underneath every downstream Windows performance metric and decision.

Apple · Siri Search

AI/ML Engineer Intern *Summer 2019*

- Expanded Siri Knowledge Card entity-type coverage and built a Hive/Impala automation pipeline for query-log-driven inline card refresh. **Entity-type coverage +214%, positive abandonment rate +35%, engagement +7%** via A/B tests.

Google · Cloud (2017) and Search (2018)

Software Engineering Intern *Summers 2017, 2018*

- **Cloud:** built a multithreaded gRPC load generator and scheduler for next-generation network systems, **+31% network performance**. **Search:** built backend infrastructure for a food-ordering feedback system (Spanner schema migration, encrypted feedback ingestion APIs, real-time partner-rating dashboards).

SKILLS

ML / AI Evaluation: LLM-as-judge calibration, scalable oversight, replay infrastructure, scorer design, prompt engineering, sycophancy / hallucination / reward-hacking detection, frontier-model comparison (Claude, GPT-5, o-series), human-subjects studies. **Statistics:** paired hypothesis testing, bootstrap CIs, McNemar, Cohen's kappa & h, Simpson's-paradox-aware aggregation, calibration vs. human gold standard. **Systems:** Python, C#, .NET, SQL, Kusto/KQL, petabyte-scale analytical SQL (Spark/Hive-class), distributed data pipelines, fleet-scale telemetry.