

STEVEN LEON

Queens, NY 11385 | 347-612-7281 | Steven_leon@live.com
github.com/stevenleon30 | linkedin.com/in/leonsteven

PROFESSIONAL SUMMARY

Quality engineer with 10+ years in QA and hands-on experience building and evaluating LLM-powered products end-to-end. Designs and ships Promptfoo evaluation pipelines with CI-gated regression testing, RAG evaluation, adversarial/prompt-injection testing, and multi-agent workflow validation. Builds full-stack AI products from architecture through deployment (Supabase, GitHub Actions, Railway, Lovable, Claude and OpenAI APIs), translating domain requirements into testable model behaviors and measurable release gates.

TECHNICAL SKILLS

LLM Platforms & APIs: Anthropic Claude API, OpenAI API, LangChain.js, model parameter tuning (temperature, top-p, cost/latency optimization)

Evaluation & Quality: Promptfoo (YAML configs, rubric scoring, red-team evals), RAGAS, gold set construction, regression testing, release gating, JSON schema validation, non-deterministic & multi-agent workflow testing, prompt injection / adversarial testing, over-refusal analysis

Automation & CI/CD: Playwright (Page Object Model, UI & API), GitHub Actions (CI pipelines, pass-rate gates), BrowserStack

Languages & Tools: JavaScript, TypeScript, HTML/CSS, JSON, YAML, VS Code, GitHub Copilot, Supabase, Railway, Lovable, Jira

Workflows: Agentic pipeline design, RAG evaluation, prompt engineering, system prompt optimization, data-driven feedback loops, full-stack AI product delivery (architecture through deployment)

EDUCATION & TRAINING

AI Test Automation of Large Language Models — 2025

Engenious University | Intensive 8-Week Master's-Level Program

- Designed and executed 150+ structured LLM regression tests spanning safety, ambiguity handling, hallucination detection, and structured output enforcement across multiple model providers.
- Built Promptfoo evaluation pipelines with rubric-based scoring and release thresholds ($\geq 90\%$ compliance pass rate), functioning as release gates before production deployment.
- Constructed gold sets and YAML-configured test suites to measure model accuracy and consistency, enabling data-driven release decisions.
- Conducted 50+ adversarial prompt simulations to detect injection vulnerabilities and unsafe outputs, reducing false negatives in safety-critical scenarios.
- Tuned model parameters (temperature, top-p) to reduce output variance and improve determinism by $\sim 25\%$, balancing accuracy against cost and latency.
- Implemented structured JSON validation checks ($\geq 99\%$ schema integrity) for downstream system reliability, catching malformed outputs before they reached production.
- Integrated automated evaluation tests into GitHub Actions workflows for pre-release validation and continuous regression detection.

Certificate, Quality Assurance & Software Testing Principles

Per Scholas, New York, NY

Bachelor of Arts, Liberal Arts & General Studies

Adelphi University, Garden City, NY

AI & LLM PROJECTS

DropLinx (droplinx.fyi) — Founder & Lead Builder: URL-to-multi-format content repurposing SaaS built end-to-end on Lovable/React and Supabase, with a multi-provider AI gateway (Claude, OpenAI, Kimi/Moonshot) orchestrating per-format content generation across batch jobs. Shipped an automated Promptfoo eval suite, GitHub Actions CI workflow, and a graded technical documentation package (README with pass/fail status labels) as part of an investor due-diligence review.

Nova Assistant Eval Suite (ai-assistant-eval): Promptfoo evaluation project covering five test categories (accuracy, safety, tone, edge cases, structured output) with a GitHub Actions CI pipeline enforcing an 85% pass-rate release gate.

Multi-Agent Trip Planner Eval: Evaluation scaffold for testing multi-agent workflow coordination, handoff accuracy, and failure modes across chained agent calls.

FDCPA Compliance Eval Suite (Rifa AI): Built Promptfoo test harness evaluating conversational AI compliance against federal debt collection regulations. Tested both GPT-4o-mini and Claude models, measuring false positive and false negative rates across regulatory scenarios.

Cafe RAG Eval: Promptfoo + TypeScript evaluation pipeline testing retrieval-augmented generation accuracy against curated ground truth datasets. Demonstrated gold set construction and scoring rubric design.

MockBook QA Suite: Full Playwright/TypeScript test suite with 43 passing tests (API, SQL, E2E), GitHub Actions CI, and a custom test harness for a sportsbook testing sandbox.

PROFESSIONAL EXPERIENCE

Quality Technician

Boyce Technologies, Oct 2025 – Present | New York, NY

- Inspect fabricated components against engineering specifications, verifying dimensional accuracy and design intent using precision measurement tools.
- Perform field quality inspections on MTA subway installations to confirm systems meet safety and performance criteria before release.
- Conduct root-cause analysis on defects, identify trends, and recommend process improvements to reduce rework across production batches.
- Maintain detailed QA reports and nonconformance records for traceability and regulatory compliance.

Quality Assurance Analyst

Creative Circle, Mar 2025 – Jun 2025 | New York, NY

- Executed functional, usability, and cross-browser test cases across devices, documenting and tracking bugs in GitHub with clear reproduction steps and severity levels.
- Re-tested resolved issues to prevent regressions and maintain stability after deployments on Vercel.
- Validated branded marketing materials, including email campaigns, landing pages, and social assets for content and visual accuracy.

Senior Quality Assurance Analyst

Vox Media Inc (The Verge), Oct 2016 – Dec 2023 | New York, NY

- Led QA for The Verge across 7 years, owning end-to-end testing of website functionality, page performance, image rendering, and cross-device usability on the Vercel platform.
- Reviewed pull requests and feature logic within GitHub repositories, collaborating with cross-functional engineering teams to enforce testing protocols before release.
- Identified performance bottlenecks and collaborated with developers to troubleshoot and optimize site speed, directly improving user experience at scale.
- Built and maintained test matrices and documented test cases, findings, and best practices for continuous improvement across release cycles.
- Collaborated with project managers and tech leads to define project scope, prioritize issues, and align sprint goals within Agile/Scrum workflows.
- Supported the successful launch of numerous product releases by coordinating QA efforts, including test planning, execution, and stakeholder reporting.

Quality Assurance Coordinator

Thrillist Media Group, Mar 2015 – Feb 2016 | New York, NY

- Maintained test matrices and performed exploratory and edge-case testing to catch issues outside defined test scopes.
- Collaborated with designers to validate builds against mockups and participated in daily Scrum meetings to coordinate deployments.
- Utilized BrowserStack for cross-browser testing across Chrome, Firefox, and Safari on macOS and Windows.