Omar Ahmed ElHawary

Applied AI Engineer | Retrieval, Evaluation & Knowledge Systems

Cairo, Egypt • +20 115 037 7005 • omaraelhawary@gmail.com
omarelhawary.com • linkedin.com/in/omaraelhawary • github.com/omaraelhawary • profiles.wordpress.org/omarelhawary

SUMMARY

I work on the part of an AI product that decides whether an answer is correct: the curated knowledge it draws on, the retrieval that grounds it, the guardrails on what it may assert, and the measurement that proves any of it holds. Currently ship that on a Claude-powered advisor for membership-site operators, along with the Python and TypeScript tooling the team curates and evaluates content in. Came to it through years of support escalations on a global SaaS product, the hardest of them at Tier 3, which is a long apprenticeship in telling a plausible answer from a true one. Write publicly on evals, retrieval, and context engineering.

APPLIED AI WORK

RETRIEVAL, GROUNDING & GUARDRAILS

- **Citation integrity.** The advisor was emitting bracket tokens that looked like citations but resolved to nothing, landing on exactly the data-grounded answers the product exists to give. Root cause was a system prompt that scoped bracket citation too broadly, so the model improvised brackets from context section labels, while the renderer stripped only numbered forms. Narrowed the instruction, replaced the renderer with a dedicated strip layer, and added a grounding eval that fails the build if a descriptive token reaches an answer.
- **Refusing to assert a false zero.** Integrations that report members but no transactions made "no revenue data" and "revenue is zero" render identically, so the advisor stated a zero it had no basis for. Declared which sources carry a billing ledger, added a revenue-coverage layer through the profile and rollup paths, and made the response layer mark coverage unavailable with an explicit instruction not to report, infer, or estimate.
- **Sensitive-vertical hard filter.** Wired the sensitive-vertical policy into the retrieval path so global content in excluded verticals is dropped at read time rather than merely deprioritized by ranking, reusing the existing predicates rather than introducing a second definition of "sensitive."

EVALUATION & MEASUREMENT

- **Found that "cited" wasn't measuring citation.** While instrumenting retrieval, traced that the citation metric counted entries appearing in the top-k handed to the model — the answer text was never parsed for citation markers, so "cited" had silently meant "was retrieved," never "the model used it." Built a retrieval-event log recording every entry retrieval considered per answer, not just those returned.
- **Per-entry performance signal.** Built the nightly rollup closing the feedback loop from message-level thumbs down to individual knowledge-base entries: citation counts, last-cited, feedback-bearing citations, and helpful rate over all-time and 30-day windows, so content review runs on a signal instead of an opinion.
- **Coverage and gap reporting.** A reporting mode over the knowledge base that grids category against question type, ranks the thinnest cells, overlays drafts in progress, and emits machine-readable output with a tunable threshold — so authoring targets the gaps rather than the easy topics.

KNOWLEDGE CURATION & CONTEXT

- **Curation and eval tooling.** Build the interfaces the team curates and evaluates content in — Python on the backend, TypeScript and SvelteKit on the front. Includes the edit-before-approve path for knowledge-base candidates, which routes edited content back through the same sensitive-vertical, sanitization, and embedding gate as new content, so a human edit can't bypass the checks.
- **The corpus itself.** Authored 214 operator playbook entries, the large majority of the curated knowledge base the advisor answers from, and defined the editorial standard and review process the wider team writes against.
- **Conversation context.** Onboarding's follow-up dropped users into an empty thread because the generated insight was cached on the user record and never written into a conversation — the context was never sent. Made the completion endpoint seed a real conversation carrying it.

PERSONAL

- **File-based agent memory** — a structured memory layer with rules for what persists and what is excluded, so an agent resumes from the previous session instead of being re-briefed. In daily use for months.
- **Agent permission boundaries** — audited every safety rule in a working AI setup, separated prose instruction from enforced control, and published the autonomy ladder it produced.

EXPERIENCE

Caseproof — MemberPress Sep 2022 – Present · Remote (USA)

Applied AI Engineer · Jul 2026 – Present

- Own answer quality on a production Claude-powered advisor, end to end — the work described above.

Tier 3 Support Engineer · May 2024 – Jul 2026

- Resolved escalations that got past Tier 1 and Tier 2 across memberships, subscriptions, and payments (Stripe, PayPal, Authorize.net) on a global SaaS product.
- Diagnosed production failures in high-traffic WordPress environments: PHP fatals, SQL deadlocks, webhook and IPN delivery failures, REST integration breaks.
- Shipped custom PHP and JavaScript solutions for logic outside core product scope, built on hooks and REST endpoints with no core modifications.

WordPress Support Engineer · Sep 2022 – Apr 2024

- Supported WordPress-based SaaS implementations at scale. Earned 15 public five-star Trustpilot reviews naming me personally over the Caseproof tenure.

SwiftX — Self-Employed Mar 2023 – Present · Cairo, Egypt

WordPress Technical Consultant / Lead Developer · parallel to Caseproof

- Scope, architect, and deliver custom plugins, integrations, and automation for WordPress and SaaS clients; mentor junior developers.

Leap13 / LeapWorx Jun 2020 – Aug 2022 · Ras Al Khaimah, UAE

WordPress Developer

- Built and maintained features for Premium Addons for Elementor using PHP, JavaScript, and the Elementor and Gutenberg APIs. Hybrid role across development, technical support, and technical content.

Freelance WordPress Developer Jan 2015 – Aug 2019 · Remote, Egypt

- Delivered 30+ WordPress sites for international clients, from requirements through launch and maintenance, including custom plugins and third-party API integrations.

OPEN SOURCE

- **Members plugin** (300,000+ active installs) — contributed the Administrator Rescue Link, a time-limited magic link that restores admin capabilities without database access, and selective role reset that restores WordPress roles without destroying roles other plugins created.
- **WordPress.org** — Admin Filters for MemberPress and Gift Reporter for MemberPress, published and maintained under my account.
- **GitHub** — Forward-Only Access, Staging Safe Mode, bulk PDF invoice generation, and per-member content rules for MemberPress.

SKILLS

Applied AI: RAG and retrieval grounding, eval design and regression coverage, retrieval observability, knowledge curation, context engineering, agent design, MCP, prompt scoping

Tools: Claude, Cursor, Claude Code, scheduled agent loops, Git/GitHub

Languages: Python, PHP, JavaScript, TypeScript, SQL

WordPress: Plugin development, REST API, Gutenberg, hooks/filters, WP-CLI, SVN publishing

Payments: Stripe, PayPal, Authorize.net, recurring billing, webhook debugging

EDUCATION & LANGUAGES

B.Sc. Computer Science — Cairo Higher Institute for Engineering and Computer Science
Arabic (native) • English (professional proficiency)