

Jeremy Yao

Boston, MA • jeremyforwork121@gmail.com • [linkedin.com/in/jeremy121](https://www.linkedin.com/in/jeremy121) • github.com/qeoplkm

EDUCATION

Northeastern University, Boston, MA — M.S. in Computer Science, GPA 3.57

Sep 2024 – Dec 2026 (expected)

University of Massachusetts Boston, Boston, MA — B.S. in Computer Science

Sep 2019 – May 2023

EXPERIENCE

AI Engineer (Contract) | JTBC Inc. (jiangnanboston.com)

Jul 2025 – Present

- Built a bilingual AI concierge from the restaurant's menu and service policies (100+ dishes, 90+ validated policy records), with language routing, allergen safeguards, and staff escalation; served 1K conversations/month
- Chose the launch LLM from a 3-model bake-off, rejecting a cheaper model that answered 7/30 Chinese prompts in English and made an unsafe allergen claim; the chosen model had 0/60 hallucination or allergen-critical failures on the initial test set
- Developed a Python eval harness with 300 bilingual cases (60 authored, 240 generated from the knowledge base with price/allergen checks); shipped owner-approved rules that lifted spec pass rate 82.3% → 94.3% (McNemar $p < 0.001$), with 0 unsafe allergen answers
- Audited the LLM judge with 100 human labels before trusting it: it never passed a bad answer but over-flagged (1 of 8 hallucination flags held up), so pass rates count as conservative and judge-only hallucination counts were dropped; a draft judge matched the final at $\kappa = 0.82$ over 600 verdicts; cache-stable prefixes cut judge cost to ~1/8
- Red-teamed the concierge with a preregistered 70-case slice (50 attacks in 5 classes, 20 benign look-alike probes): 0/50 attacks succeeded with or without guards, but on 240 paired controls the guards broke 9 critical allergy answers; fixed in v2 (0 regressions); guards stay off until proven useful
- Hardened the Cloudflare Workers API with origin validation, IP hashing, rate limits, and daily usage caps

Open-Source Contributor | HiveMind (AgentFlow multi-agent runtime, 1.1K GitHub stars)

Aug 2026 – Present

- Authored 3 merged PRs (#92, #93, #94) for model fallback, reasoning persistence, and multimodal attachments across Python, Spring Boot, PostgreSQL, and Next.js
- Built model routing with bounded retries and ordered fallbacks (rate limits, 5xx, timeouts, and connection errors retry; invalid 4xx fail fast), tracing every attempt via structured logs and OpenTelemetry counters
- Separated SSE text, reasoning, and attachment events; reconciled live client drafts with persisted PostgreSQL messages to restore state after reconnects without passing stored reasoning into subsequent model context

Junior Software Engineer | Mole AI | Remote

Aug 2023 – Aug 2024

- Cut p95 latency on AI-backed endpoints from 500 ms to 200 ms by running independent AI-service calls concurrently instead of sequentially and streamlining Express middleware
- Built OAuth 2.0 Authorization Code login and JWT session management, securing sign-in and virtual-currency purchases across web and mobile

Co-Founder & Developer | US Auto Repair Guide (uscarrepairguide.com)

Jun 2023 – Present

- Built and now operate a nationwide directory of 200K+ auto-repair shops across 50 states, with Google Maps Places API data refreshes and bilingual search by city, ZIP code, and service type
- Designed indexed, paginated MySQL search on AWS (EC2, RDS, CloudFront), cutting page load 48% (2.5 s → 1.3 s); rebuilt v2 in Next.js/TypeScript with bounded LRU caching and CI data-integrity checks

PROJECTS

Restaurant Reservation Agent — prototype (Python, LangGraph, SQLite)

Jan 2026

- Gave up automatic confirmation rather than script the booking system's back office (Chowbus, no API): the LangGraph agent writes only to a request ledger staff confirm, keeping credentials out of the loop and the model away from real tables
- Kept authorization and contact data out of the model: server-session identity, guest approval of an argument-bound card before every write, and contact details from the account profile, so every “we'll email you” is backed
- Scored on final database state with a deterministic oracle, not an LLM judge, in a preregistered, hash-frozen run: 0 invariant violations across 272 conversations, 89/90 normal tasks

Break the Ice: Date Night (Swift, SwiftUI, Core Motion, XCTest)

Aug 2026 – Present

- Built and launched a SwiftUI app on the App Store with five pass-the-phone date-night games, on-device personalization, safety filtering, and no repeated prompts across sessions
- Implemented Core Motion effects, haptics, and local persistence; validated context-dependent behavior with an XCTest suite covering 60 user-context scenarios

FORESEE: Conversational Dietary Recall — Research Assistant, UMass Boston

Oct 2022 – Jul 2023

- Ran usability and privacy studies of a Google Home/Alexa dietary-recall assistant for older adults (SUS 80+)

SKILLS

AI & LLM: LLM evaluation (LLM-as-judge, human calibration), guardrails, prompt-injection testing, agent tool authorization, structured outputs, prompt caching, model routing/fallback, LangGraph, OpenAI API

Languages: Python, JavaScript/TypeScript, Java, Swift, C/C++, SQL

Backend: Spring Boot, Node.js/Express, REST APIs, SSE, OpenTelemetry, OAuth 2.0 / JWT, Docker

Frontend & Mobile: React, Next.js, Redux, HTML/CSS, SwiftUI, Xcode / XCTest

Data & Cloud: PostgreSQL/Alembic, MongoDB, MySQL, Redis, AWS (EC2, RDS, CloudFront), Cloudflare Workers