

# AREEF SYED

Philadelphia, PA 19104

☎ +1 215-452-8651 ✉ areefsyed96@gmail.com 🔗 linkedin.com/in/areefsyedfe 🌐 github.com/the-sniper 🌐 areefsyed.com

## Summary

AI engineer with a Master's in AI & Machine Learning and 6+ years of software engineering, specializing in agentic systems and LLM infrastructure: multi-agent orchestration and evaluation, Retrieval-Augmented Generation (RAG) over vector search, tool/function calling, and Model Context Protocol (MCP) / Agent-to-Agent (A2A) integrations. Ships agent products end to end, from retrieval pipelines and protocol-level transport to monitoring, evaluation, and production deployment. Authorized to work in the U.S., no sponsorship required.

## Technical Skills

**LLM & Agentic AI:** Agent orchestration, agent evaluation & drift detection, Retrieval-Augmented Generation (RAG), embeddings & vector search (pgvector, FAISS), tool/function calling, Model Context Protocol (MCP), Agent-to-Agent (A2A), LangChain, OpenAI API, Anthropic Claude, Google Gemini, hybrid retrieval & knowledge graphs, prompt engineering, token streaming over Server-Sent Events (SSE)

**Languages:** Python, TypeScript/JavaScript, SQL, C#, Bash

**ML & Data Science:** PyTorch, scikit-learn, pandas, NumPy, YOLOv8 & ResNet50 (computer vision), faster-whisper (ASR), Prophet (time-series forecasting), spaCy, NLTK, sentence-transformers; NLP, classification pipelines, reinforcement learning

**Backend & APIs:** FastAPI, Flask, Django, Celery, Node.js, Next.js, REST microservices, API gateway / backend-for-frontend, WebSocket, SSE, JWT & OAuth, Pydantic, Zod

**Data & Storage:** PostgreSQL/pgvector, MySQL, Supabase, Redis, Prisma, SQLAlchemy, S3/MinIO, IndexedDB

**Cloud & DevOps:** AWS (S3, ECR, ECS), GCP Cloud Run, Docker & Docker Compose, Fly.io, Vercel, Hugging Face Spaces, GitHub Actions, Continuous Integration/Continuous Deployment (CI/CD), Ansible, nginx

**Testing & Process:** Playwright, Vitest, pytest, structured JSON logging, Git, Jira, Agile/Scrum, code review & mentoring

## Experience

### Carmel Labs

Mar 2026 – Jul 2026

Forward Deployed Engineer

Python, TypeScript, LangChain, Agentic AI, MCP/A2A, Next.js

- Shipped 4 agent-testing products in 4 days (A/B response diffing, public agent probing, tool auditing, and behavioral drift detection), then delivered end-to-end validation for multi-agent AI pipelines in a coordinated 2-day backend and frontend release.
- Built AgentStatus, an outside-in AI agent monitoring and validation platform with drift detection, behavior checks, efficiency scoring, and anomaly alerting, onboarding agents through A2A Agent Cards and JWT-based cross-agent authentication over emerging MCP/A2A standards.
- Engineered protocol-level support for Model Context Protocol (MCP) Streamable HTTP transport, covering session management and dual JSON/Server-Sent Events (SSE) response handling, validated against production MCP servers.
- Delivered an AI knowledge assistant from first pipeline to interactive UI in under 3 weeks, pairing a knowledge-graph layer with hybrid retrieval over ingestion from 5 enterprise sources (Gmail, Zoom, Calendar, GitHub, webhooks).
- Built AgentDiff, a GitHub App running automated LLM-powered pull-request checks and AI code review as CI/CD status checks.
- Owned the monitoring, alerting, and billing layers: alert integrations, Time to First Byte (TTFB) SLA enforcement, usage-based subscription metering, and the public AOI Index benchmark for AI agents.
- Acted as technical face of the product, pitching at in-person industry events and running developer content, and built an internal GTM automation platform unifying enterprise data sources into one dashboard with activity logging and pipeline tracking.

### Collidascope VR

Sep 2025 – Feb 2026

Virtual Reality Developer Intern (Co-op)

Unity, C#, Jira, Webflow

- Redesigned the Developer Menu UI, improving usability by 40%; refactored legacy code to cut redundancy by 30% and server ping latency by 35%.
- Improved multiplayer synchronization and codebase maintainability, and ran the MVP playtest that validated core gameplay pre-launch.

### AuctionSoftware Inc

Feb 2019 – May 2024

Senior Frontend Developer

JavaScript, React.js, Next.js, Node.js, Python, FastAPI, SQL

- Managed end-to-end development of 80+ data-driven web applications, delivering React interfaces, Node.js and Python (FastAPI) backend services, REST APIs, and MySQL/PostgreSQL integrations across Agile sprints tracked in Jira.
- Lifted user activity by 35% by driving client platforms to a 90% Core Web Vitals score, tuning rendering, asset delivery, and front-end performance to Google's standards.
- Raised team productivity by 25% as team lead by establishing code-review, peer-review, and QA practices that reduced post-release defects, and by mentoring junior developers to faster onboarding.
- Led front-end development for Nellis Auctions, one of the largest secondary marketplaces for Amazon returns, coaching engineers on coding standards through structured code reviews.

**Client Engagement, AuctionSoftware: DeBeers Diamonds** (media-heavy auction platform) **Mar 2024**

- Built the React front end and real-time data services (bidding APIs over REST microservices, an API gateway, and a backend-for-frontend layer), applying distributed-systems patterns to scale server-paginated search across a 10k+ item catalog.

**Client Engagement, AuctionSoftware: World Security Services (DP World)** (live auction platform) **Nov 2023**

- Integrated WebRTC live video with real-time data streaming and bidding-trend analytics through a reusable, scalable platform architecture later adopted by multiple Fortune 500 companies.

## Projects

---

**Klyro (Open Source):** embeddable RAG copilot, [github.com/the-sniper/klyro](https://github.com/the-sniper/klyro) **Jan 2026**

- Open-sourced an embeddable AI copilot built on a from-scratch RAG pipeline over PostgreSQL/pgvector: document ingestion (PDF, DOCX, URL), ~300-token chunking, top-15 cosine-similarity retrieval with keyword-boost re-ranking, and strictly grounded answers with source citations; 1,100+ npm downloads for @klyro/widget in week one.
- Extended the copilot into a tool-calling agent with 5 live tools (GitHub projects and repo details, URL fetch, Calendly event types and open slots) plus conversational query rewriting for follow-ups, and enforced multi-tenant vector isolation via user-filtered similarity search to prevent cross-tenant leakage; covered retrieval and tool logic with 26 unit tests.

**AirLog:** voice-first user testing platform, [airlog.live](https://airlog.live) **Nov 2025**

- Built a speech-to-insight ML pipeline: self-hosted faster-whisper ASR on Docker/Fly.io behind a retrying proxy, LLM classification of tester feedback into 5 categories, and session-level AI summarization persisted to Supabase/PostgreSQL across 107 API routes.
- Eliminated feedback loss from dropped connections and OS-killed recordings with a client-first durable capture queue (IndexedDB persistence before upload, up to 5 retries, session recovery), and shipped stakeholder analytics behind three-tier JWT multi-tenant access control.

**Forecast My Park:** time-series visitor forecasting service **Jul 2025**

- Built a forecasting service producing 1 to 365-day daily crowd predictions for 30 parks, training per-park Prophet models with 10+ regressors (weather, holidays, school calendars, gas prices) and custom seasonalities on a synthetic 38.7k-record dataset, validated by time-series cross-validation on MAE, MAPE, and RMSE.
- Served 19 trained models through a 9-endpoint FastAPI service with a 24-hour model cache and on-demand retraining, behind a Next.js map and chart dashboard (Leaflet, Recharts), with an AWS ECR/ECS container deployment pipeline behind an nginx rate-limited proxy.

**GesturePro:** real-time ASL sign-to-text PWA **Jun 2025**

- Led a 4-person team (35 of 56 commits) building a webcam ASL translator: YOLOv8 inference over 23 sign classes feeding a sentence builder that turns noisy per-frame detections into coherent text via 3-frame stability checks, a confidence floor, and word debouncing.
- Shipped dual authentication (Google OAuth plus FastAPI JWT) behind unified middleware across 16 endpoints on a FastAPI, SQLAlchemy, and PostgreSQL stack; prototyped an alternate MediaPipe keypoint plus LSTM/TCN track on a WLASL pipeline (468 videos to keypoint sequences).

**Fact Checker (Open Source):** multi-stage LLM claim verification **Feb 2025**

- Co-built a real-time claim verification engine (Next.js front end, FastAPI backend) orchestrating pluggable multi-stage LLM solver pipelines, with 25 registered solvers behind a decorator registry, over multi-source web evidence retrieval.
- Sustained throughput inside provider rate limits using async OpenAI batching with exponential backoff, and deployed the stack via Docker to GCP Cloud Run and Hugging Face Spaces.

## Research

---

**Curiosity-Driven Exploration in Sparse Reward Environments** **Mar 2025**

- Co-authored research on a curiosity-driven Deep Q-Network solving Montezuma's Revenge: 4.18M intrinsic reward in 2M steps against a zero-reward baseline, reaching a 2.093 average reward metric.

## Education

---

**Drexel University** **Mar 2026**  
*Master of Science in AI & Machine Learning, GPA 3.7* *Philadelphia, PA*

- Coursework: Deep Learning, Machine Learning, Natural Language Processing, Computer Vision, Reinforcement Learning

**Bharath Institute of Higher Education and Research** **May 2018**  
*Bachelor of Technology, Computer Science, CGPA 7.92/10* *India*