Tariq Rashid Senior AI Engineer

San Francisco, CA · tariq.rashid@gmail.com · +1 415-555-0152 · linkedin.com/in/tariqrashid

PROFILE SUMMARY

- **Senior AI Engineer** with **7 years** building **retrieval and LLM systems** for **consumer search and developer platforms**, specializing in **RAG infrastructure, citation grounding, and LLM evaluation at scale**.
- Hands-on across primary language (**Python**), model APIs (**OpenAI**, **Anthropic**, **Cohere**), agent frameworks (**LangGraph**), vector databases (**Pinecone** and **Qdrant**), evaluation (**LangSmith** and **Braintrust**), and serving (**vLLM** on **Kubernetes**), with strong fundamentals in **information retrieval and distributed systems**.
- Deep expertise in **hybrid retrieval**, **reranking**, **LLM-as-judge evaluation**, and **agent cost guardrails**, using **golden-set regression** and **retrieval-quality SLOs** to ship **answers that cite the right source the first time**.
- Owns the retrieval and grounding services end to end with **product, infrastructure, and the LLM platform team**, from index design and offline eval to **on-call, incident review, and the quarterly roadmap**.
- Emerging tech lead who mentors 3 mid-level engineers, chairs the **weekly eval-review forum**, and wrote **4 RFCs on agent patterns** the platform team adopted across three product surfaces.

TECHNICAL SKILLS

Languages:	Python, Go (services), SQL, TypeScript, Bash
LLM Tooling:	OpenAI, Anthropic, Cohere Rerank, LangGraph, LangChain, tool calling, structured outputs
Retrieval & RAG:	Pinecone, Qdrant, hybrid BM25 and dense retrieval, reranking, chunking, query rewriting, citation grounding
Evaluation:	LangSmith, Braintrust, golden sets, LLM-as-judge, retrieval metrics (recall@k, nDCG), regression gates
Serving & Inference:	vLLM, Kubernetes, gRPC, caching, batching, latency and cost budgets
Agents & Guardrails:	planner-executor and ReAct patterns, tool-use policies, cost guardrails, fallback routing
Observability:	OpenTelemetry, Datadog, retrieval-quality SLOs, incident review, on-call rotation

EDUCATION

University of Michigan — **B.S. in Computer Science** — Ann Arbor, MI · Sep 2014 - May 2018

WORK EXPERIENCE

Perplexity — **Senior AI Engineer** — San Francisco, CA · Jul 2022 - Present

- Own the **RAG infrastructure** across 10 production indexes and **320M chunks** in **Pinecone** and **Qdrant**, with full responsibility for retrieval-quality SLOs, on-call, and the quarterly roadmap.
- Designed the **citation-grounding verifier** on **Cohere Rerank** plus an in-house grounding model, lifting grounding accuracy by **38%** and cutting unsupported-claim incidents by 54%.
- Moved retrieval to **hybrid BM25 and dense** search with **LangGraph** query rewriting, lifting **recall@10 by 22%** on the public-search eval set without adding latency.
- Built the golden-set automation in **LangSmith** and **Braintrust**: **1,200 prompts** across 12 categories with LLM-as-judge regression checks on every release.
- Wrote **4 RFCs on agent patterns** (planner-executor, ReAct, tool-calling cost guardrails) adopted by the LLM platform team and three partner product surfaces.

Pinecone — **AI Engineer** — New York, NY · Jul 2018 - Jun 2022

- Built the customer-facing **RAG starter kits** in **Python** and **LangChain**, used by **650+ enterprise customers** in early-access deployments.
- Shipped **4 retrieval-quality benchmarks** covering hybrid retrieval, namespace isolation, and metadata filtering, which became the reference numbers in the public docs.
- Ran the solutions on-call for the top 20 accounts, turning recurring support cases into **11 product fixes** and a chunking guide that cut related tickets by **40%**.