JULIAN MERCER

AI ENGINEER

San Francisco, CA | (415) 555-0148 | julian.mercer@email.com | linkedin.com/in/julianmercer

PROFESSIONAL SUMMARY

AI Engineer with 6+ years building large language model features, retrieval-augmented generation systems, and evaluation tooling for production software at Notion and Grammarly. Turns prototypes into reliable, observable services with clear quality gates and safe fallbacks.

CORE SKILLS

Large language models | Retrieval-augmented generation | Prompt engineering | Vector search | LLM evaluation | Python | PyTorch | FastAPI | Kubernetes | Fine-tuning | Model serving | Observability | Cost optimization | Safety guardrails

PROFESSIONAL EXPERIENCE

Senior AI Engineer Mar 2024 - Present
Notion, San Francisco, CA

- Built a retrieval-augmented question-answering feature over workspace content, reaching 41% weekly adoption among eligible teams in six months.
- Designed hybrid vector and keyword retrieval with reranking, raising answer groundedness scores from 71% to 89% on a 1,200-question test set.
- Created an automated evaluation harness with rubric graders and human spot checks, catching 14 regressions before release.
- Cut inference spend per active user 33% by routing simple requests to lightweight models and caching repeated context.

AI Engineer Jun 2022 - Feb 2024

- Reduced median response latency from 4.1 to 2.3 seconds through prompt caching, streaming, and smaller routing models.
- Prototyped an AI writing assistant on a hosted language model and moved it to general availability after a staged rollout to 5% of users.
- Wrote prompt templates with versioned test cases, so quality changes could be compared before each release.
- Added fallback responses and timeout handling for model outages, keeping the assistant usable during two provider incidents.

Machine Learning Engineer Jun 2020 - May 2022
Grammarly, San Francisco, CA

- Fine-tuned transformer models for tone suggestions, improving accepted-suggestion rate 7.5% across 12 million weekly writers.
- Built data labeling workflows with guideline audits, lifting inter-annotator agreement from 0.68 to 0.82 for training sets.
- Packaged models behind a gRPC service with canary releases, keeping availability above 99.9% through four major launches.
- Added drift monitors on input length and language mix, flagging two silent quality drops within a day of occurrence.
- Profiled inference code and batched requests, trimming serving cost for the tone model by about 18% at the same latency.

EDUCATION

B.S. in Computer Science, Georgia Institute of Technology, Atlanta, GA

CERTIFICATIONS

AWS Certified Machine Learning - Specialty | DeepLearning.AI Generative AI with LLMs