JULIAN MERCER

PROMPT ENGINEER

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

PROFESSIONAL SUMMARY

Prompt Engineer with 7+ years in applied NLP, including 3 designing production LLM prompts, evaluations, and retrieval workflows for customer-facing assistants. Measures every change against labeled test sets for accuracy, safety, latency, and cost. Works with product and engineering to turn prototypes into reliable features.

CORE SKILLS

Prompt design and versioning | LLM evaluation harnesses | Retrieval-augmented generation | Few-shot and chain-of-thought prompting | Red teaming and safety testing | Python | OpenAI and Anthropic APIs | Embeddings and vector search | Cost and latency optimization | Annotation and labeling | Output validation | LangChain | Experiment tracking | NLP fundamentals

PROFESSIONAL EXPERIENCE

Prompt Engineer Feb 2023 - Present
Intuit, Mountain View, CA

- Design and version prompts for a tax and accounting assistant that serves several hundred thousand questions a month.
- Built a 1,800-example evaluation set and an automated grading harness, raising answer accuracy from 74 to 89 percent over five prompt revisions.
- Lowered average inference cost per conversation by 38 percent by shortening system prompts, caching retrieved context, and routing simple questions to a smaller model.
- Reduced policy-violating responses on red-team tests from 6 percent to under 1 by adding refusal rules, output validators, and few-shot examples.
- Tuned a retrieval-augmented pipeline with chunking and re-ranking changes, improving answer grounding on help-article questions by 17 points.
- Coach eight product managers and support leads on prompt writing and run a monthly review of failure cases.

Machine Learning Engineer, NLP Jun 2021 - Jan 2023

- Trained and shipped text classification models that route customer chats to the right support queue, reaching 92 percent top-1 accuracy.
- Built annotation guidelines and led a team of six labelers, reaching inter-annotator agreement of 0.82 on the intent taxonomy.
- Prototyped a GPT-based summarization tool for support agents, cutting average after-call notes time by 45 seconds in a pilot of 40 agents.
- Deployed models behind a Python service on AWS with monitoring for drift and latency.

Computational Linguist Aug 2019 - May 2021
Nuance Communications, Burlington, MA

- Wrote grammars and dialogue flows for healthcare voice assistants, reducing no-match errors on common requests by 20 percent.
- Analyzed call transcripts to find recurring intent confusions and delivered fixes to the speech tuning team each sprint.
- Created test sets of 3,000 utterances covering accents and noisy audio, used for each release regression.
- Documented conversation design guidelines adopted across three client projects.

EDUCATION

Bachelor of Science in Cognitive Science, University of California, San Diego, La Jolla, CA

CERTIFICATIONS

Generative AI with Large Language Models, DeepLearning.AI and AWS | Machine Learning Specialization, DeepLearning.AI and Stanford Online