

Yasir RAZA

✉ yasirabdali6@gmail.com | 🌐 github.com/Yasirrazaa | 🔗 linkedin.com/in/yasir-raza | 📍 Islamabad

☎ +92 314 5363379 | 🐍 pypi.org/user/yasirabdali6 | 🏆 kaggle.com/yasirabdaali | 🌐 yasirrazaa.github.io



SUMMARY: AI Engineer with 1.5+ years of remote experience on US-based teams, building production LLM systems: agent platforms, MCP tooling, voice AI, and model fine-tuning. Creator of **NotebookLLM** (10K+ PyPI downloads). I ship end-to-end — architecture, evaluation, deployment, and cost control.

TECHNICAL SKILLS

LLM & AGENTS: LangGraph, MCP (Model Context Protocol), function calling, RAG, agent memory, prompt engineering, red-teaming

MODELS & TRAINING: PyTorch, Transformers, LoRA/QLoRA, Unsloth, PEFT, GRPO (TRL), Whisper / Faster-Whisper, ElevenLabs, Gemini

INFRA & DEPLOYMENT: FastAPI, Docker, RunPod serverless, AWS, S3, CI/CD (GitHub Actions), CUDA graph optimization

LANGUAGES & DATA: Python, TypeScript, SQL, Bash; pandas, NumPy, Scrapy, Playwright

WORK EXPERIENCE

TEAMLIFT — UNITED STATES (REMOTE).....

Junior Data Scientist

Jan 2024 - Dec 2024

- Scraped **thousands of AI tools** from public tool directories and built an automated verification pipeline (HTTP status checks, screenshot capture, freshness validation) powering **Teamo**, Teamlift's recommendation engine.
- Preprocessed and normalized thousands of tool records into model-ready features, eliminating stale and duplicate listings from the dataset.

Data Science Intern

Jul 2023 - Dec 2023

- Built an automated pipeline collecting **thousands of job postings** from dice.com with deduplication, guaranteeing a clean non-redundant dataset.
- Trained **time-series models** to forecast skill demand extracted from job descriptions, informing product and market decisions.

TECHNICAL PROJECTS

Kausal — Production AI Agent Platform (private)

2026

- Extended the open-source ii-agent framework in **pure Python** (no LangGraph): custom **RAG** pipeline, layered memory, full **observability** (structured logging, OpenTelemetry tracing, metrics with self-hosted SigNoz), realtime **voice/video via LiveKit**, **LiteLLM** multi-model routing, agent skills, and S3 storage — guarded by a **7,275-test** suite.

NotebookLLM — Open-Source Python Package + MCP Server

2025 - 2026

- Published a **PyPI** package + MCP server (10K+ downloads) making Jupyter notebooks LLM-efficient: agent-optimized output cuts tokens up to **80%** across 8+ formats, with tiktoken token budgeting, streaming for 10 MB+ notebooks, and Claude Desktop / Cursor / VS Code integration.

Chatterbox-Optimized — Production Text-to-Speech Engine

2025

- Up to **4.5× faster inference** (RTF 0.59 → 0.13) via bucketed CUDA Graphs, bfloat16, and KV-cache tuning; zero-hallucination pipeline (parallel candidates + **Faster-Whisper** WER validation) deployed to **RunPod serverless**.

Archon 2.0 — Agent Security Platform

2026

- 8-layer defense pipeline, **222-probe corpus** across 10 attack packs, runtime defense proxy, MCP security scanning, and memory-poisoning targets — full **OWASP Agentic Top-10 (ASI01-ASI10)** coverage; **2,295 tests** at a 93% coverage gate. Predecessor ranked **13th/100+ overall** (49.7% avg; 85.1% defense / 14.4% attack) in Berkeley's AgentBeats Arena.

EDUCATION

B.Sc. Software Engineering, National University of Modern Languages (NUML), Islamabad

Aug 2026

CERTIFICATIONS

CERTIFICATES: Kaggle **2× Expert** (Datasets · Notebooks) · DataCamp DS/DA career tracks · 4 Google Cloud Gen AI badges · NUST LLM Bootcamp