

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, building the structured dataset that powers **Teamo**, Teamlift's recommendation engine.
- Built an automated **verification pipeline** (HTTP status checks, screenshot capture, freshness validation) that kept listings accurate and trustworthy — eliminating stale and duplicate entries at scale.
- Preprocessed and normalized thousands of tool records into model-ready features for the recommendation engine.
- Worked directly with US-based stakeholders remotely, shipping dataset improvements on tight iteration cycles.

Data Science Intern

Jul 2023 - Dec 2023

- Built an automated pipeline collecting **thousands of job postings** from dice.com with **deduplication** — hardened with retries, rate limits, and robust parsers.
- Extracted skill-demand signals from job descriptions at scale; trained **time-series models** to forecast demand, 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.	
Archon 1.0 — AgentBeats Security Arena (Competition)	2026
Scenario-agnostic attacker/defender agents (A2A protocol + Google ADK): ranked 13th/100+ overall (49.7%), 6th Defense (Blue, 85.1%), 18th Attack (14.4%) on the private on the private leaderboard — generalizing to unseen scenarios by design; hardened by 286 tests.	
Cerberus — Stealth Browser Infrastructure for AI Agents	2026
Self-hostable control plane pairing a stealth Chromium build with token-efficient orchestration: a 60-tool MCP bridge, multi-agent isolation gateway with worker leases, and 86 automated verification checks across twelve verified build phases.	
Forge — The API Compiler for the Agentic Era	2024 - 2026
One OpenAPI 3.x spec in, production artifacts out: SDKs in 10 languages, CLIs, MCP servers, Agent Skills, Terraform providers — 17 output types with TypeScript SDKs compiling strict on Stripe (3K files) and Cloudflare (14K files) real specs.	

EDUCATION

B.Sc. Software Engineering , National University of Modern Languages (NUML), Islamabad	Aug 2026
---	-----------------

CERTIFICATIONS

Datasets Expert · Notebooks Expert (13 medals; Datasets rank 731/11,277, best #109), Kaggle	2024 - 2026
Associate Data Scientist in Python · Data Analyst with Python (Career Tracks), DataCamp	2025 - 2026
Generative AI Fundamentals · Intro to LLMs · Intro to Generative AI · Intro to Responsible AI, Google Cloud Skill Badges	2023
Python · Intro to Programming (Professional Certificates), Kaggle Learn	2025
Advanced LLM Bootcamp, NUST & Sky Electric	2024