

Muhammad Ali Nasir

+92-325-9308643 | muhammadalinasil00786@gmail.com | linkedin.com/in/al1-nasir | github.com/al1-nasir | alinasil.me

EDUCATION

Pakistan Institute of Engineering and Applied Sciences (PIEAS)

Islamabad, Pakistan

B.S. Computer and Information Sciences | CGPA: 3.44 / 4.0

June 2026

- HEC National Skills Competency Test (NSCT) | **96.7th Percentile** nationwide

EXPERIENCE

ML Research Engineer

Dec 2025 – Mar 2026

S2Cool | GIKI Research Collaboration (via PIEAS) | [Demo]

Remote, Pakistan

- Designed an AI-driven platform combining ML-based solar irradiance prediction with thermodynamic cooling optimization for active university research. Processed **485,000+** weather records across 4 Pakistani cities.
- Benchmarked 10 ML models achieving **94.5% accuracy** with Extra Trees; delivered a real-time dashboard with live/historical analysis modes and multi-city comparison.

PROJECTS

Decision Frame | ML Pricing & Demand Intelligence | FYP | [Demo]

May 2026

Chronos-Bolt, DoubleML, LightGBM, Sentence-Transformers, FastAPI, Shopify

- Architected an end-to-end pricing intelligence platform combining **Chronos-Bolt 30-day demand forecasting** with **DoubleML** and **LightGBM** causal elasticity estimation and bounded profit optimization to recommend business-constrained, profit-maximizing prices for e-commerce retailers.
- Engineered cold-start elasticity transfer via Sentence-Transformer embeddings + KMeans clustering. Designed a Shopify-ready adapter architecture extendable to WooCommerce and Daraz.

LocalForge | Self-Hosted AI Control Plane | [GitHub]

Apr 2026

FastAPI, llama-cpp-python, Qdrant, Unsloth, scikit-learn, SQLite WAL

- Built a self-hosted LLM platform with OpenAI-compatible routing. Developed a multi-signal ML router (**<5 ms, ~90% accuracy**) and LoRA fine-tuning pipeline (**2x faster, 60% less VRAM**).
- Engineered a VRAM-aware model lifecycle manager (UNLOADED → LOADING → HOT) with real-time SSE streaming and RAG-based knowledge retrieval.

Slack AI Agent | Human-in-the-Loop Automation | [GitHub]

Mar 2026

LangGraph, FastAPI, Slack Bolt, SqliteSaver, SQLite

- Developed a production Slack AI agent on LangGraph with **AsyncSlackRequestHandler** and a FastAPI backend, implemented full multi-turn conversation state persistence via SqliteSaver across sessions.
- Designed human-in-the-loop approval workflows using the `interrupt.before` pattern, pausing agent execution before irreversible actions and resuming seamlessly after human review.

Research Council | GraphRAG Multi-Agent LLM System | [GitHub]

Feb 2026

LangGraph, langgraph-bigtool, Neo4j, ChromaDB, Groq, FastAPI

- Structured a **4-agent deliberation system** (Evidence, Skeptic, Connector, Methodology) reasoning in parallel over a Neo4j scientific paper graph, producing confidence-scored answers with citation nodes.
- Applied dynamic tool loading to serve 2-4 tools per query instead of 50+, cutting context from **25,000 to ~2,000 tokens** and reducing per-query cost to **\$0.002** across PubMed and arXiv.

CodeGraph CLI | Open-Source Code Intelligence | [GitHub] | `pip install codegraph-cli`

Feb 2026

CrewAI, tree-sitter, LanceDB, SQLite, FastAPI

- **Published pip package** parsing codebases into dependency graphs via tree-sitter across **40+ languages**, exposing semantic code search over 5 embedding models, multi-hop impact analysis, and an interactive browser-based code explorer.
- Architected a 4-agent CrewAI system with 11 purpose-built tools, 6 pluggable LLM providers, and per-file rollback before every modification.

RESEARCH & PUBLICATIONS

LightUHope: Lightweight Transformer for Multiclass Urdu Hope Speech Classification

2026

First Author | Under Review, Springer

- Proposed a **3.2M-parameter** transformer achieving macro **F1 = 0.92** on a 4-class Urdu task, reducing parameters by **97%** vs. mBERT. Built a **24,124-sample** dataset. LIME/SHAP identifies future tense (71%) as the primary optimism driver.

TECHNICAL SKILLS

Agentic AI: LangChain, LangGraph, CrewAI, LlamaIndex, MCP, Qdrant, ChromaDB, LanceDB

ML / DL: PyTorch, scikit-learn, PEFT, Unsloth, LoRA/QLoRA, Transformers (BERT, XLM-R, DistilBERT)

NLP: Semantic Search, RAG, GraphRAG, TF-IDF, LIME, SHAP

Backend: FastAPI, Django REST Framework, Docker, Neo4j, PostgreSQL

MLOps: llama.cpp, SSE Streaming, REST API Design, CI/CD, Docker-based Microservices