KUNTAL PAL

Riverside, CA • +1 951-202-8635 • kuntal.beehiveai@gmail.com
kuntalpal.dev • linkedin.com/in/kuntal-pal • github.com/kpal002

SUMMARY

AI Engineer with a PhD in physics and 2+ years building LLM systems at scale. Designed and shipped multi-agent pipelines, RAG architectures, and applied ML models across varied domains. Brings research rigour to engineering problems: comfortable owning a system end-to-end, debugging critical failures, and translating ambiguous product requirements into reliable production systems.

TECHNICAL SKILLS

- **Languages & Data:** Python (Advanced), SQL (Advanced), Go (Intermediate)
- **Agentic AI:** LangGraph, LangChain, Claude Code, Google Antigravity, RAG, Multi-Agent Orchestration, ReAct, Prompt Engineering, MCP
- **ML & Data Science:** PyTorch, TensorFlow, scikit-learn, LSTMs, sktime, statsmodels, ARIMA/Prophet
- **MLOps & Cloud:** Docker, Kubernetes, AWS, GCP, Vertex AI, Vector Databases, REST, gRPC
- **Specializations:** Predictive Modeling, Multi-Agent Systems, LLM Fine-tuning, Multimodal Learning, Anomaly Detection

WORK EXPERIENCE

AI Engineer | *Beehive AI, Inc.* Jul 2024 – Present

- Engineered a production agentic RAG system (LangChain + Postgres pgvector) enabling natural-language-to-SQL generation for real-time customer analytics over customer datasets – reducing analyst query time and democratizing data access across non-technical teams.
- Developed an LLM-powered QA framework for data categorization, automating 80% of label validation and classification – reducing manual effort while improving accuracy.
- Built a scalable virtual persona simulation framework using LLMs, contextual retrieval, and memory systems to model user behavior and drive data-informed product decisions.
- Architected a multi-agent commerce system using Vertex AI, LangGraph, and contextual retrieval with tool use and conversational interface, delivering personalized customer experiences end to end.

Research Scientist | *Pocket FM* Jan 2024 – Jul 2024

- Delivered an LSTM-based time-series forecasting model using episode-level engagement data and metadata to predict long-term content performance, owning the full model lifecycle from feature engineering and hyperparameter tuning through production deployment and monitoring.
- Led development of a content ranking model using semantic embeddings and engagement signals, driving business impact – +4.5% CTR and +4.0% increase in comment volume.

AI Fellow | *PI School, Rome, Italy* Dec 2023 – Feb 2024

- Built an end-to-end pipeline for automated quality assessment of long-form medical research papers, extracting and embedding structured and unstructured content into a vector database and retrieving evidence to ground model-based scoring across predefined categories.
- Improved classification accuracy to 75% via systematic prompt optimization and to 69% via Quantized LoRA fine-tuning, reducing manual review effort across hundreds of medical documents.

Research Scientist Intern | *Deepgram, Inc.* Jun 2023 – Sep 2023

- Trained a self-supervised audio embedding model on 35K+ hours of audio data, achieving 86% accuracy across downstream tasks including language detection and domain classification.
- Engineered a multimodal embedding space aligning audio and text representations, enabling cross-modal tasks; outperformed BERT text-only baselines by 5 percentage points.
- Operated distributed training pipelines across large-scale audio datasets, optimizing data throughput and infrastructure efficiency for high-volume learning workloads.

PROJECTS

σ-RAG: Significance-Threshold Retrieval — PyPI | Github

- Identified a gap in standard RAG: top-k retrieval cannot distinguish answerable from unanswerable queries. Modeled cross-document embedding similarity as a Gaussian noise floor and built a retrieval gate at a configurable σ threshold.
- Achieves 100% out-of-domain query suppression at 2σ with no accuracy loss, reducing chunks passed to the LLM from 3 to 1.8; deployed in production at Beehive AI.

Finley - Intelligent Financial Advisor Agent — Github | Demo

- Designed a production-grade financial analysis agent using a 5-node LangGraph state machine orchestrating ARIMA forecasting, Markowitz portfolio optimization, VaR/CVaR risk modeling, and Isolation Forest anomaly detection, synthesized into a structured advisory report.
- Built a typed state contract between nodes with a validation gate that re-routes on confidence < 0.7, catching hallucinated tickers and reversed signals; intelligent routing keeps first-run analysis at 15–30s while follow-ups answer from prior context in <1s.

Agentic Forecaster for sktime — Github | Demo

- Programmed a drop-in sktime forecaster using a ReAct agent loop to automatically select, fit, and score time-series models (ARIMA, Prophet, ExponentialSmoothing, Theta, TBATS) from plain-English descriptions; constrained to 6 auditable tools with full transcript logging.

EDUCATION

Ph.D., Physics | *University of California, Riverside* Sep 2017 – Jun 2024

- Dissertation on particle physics involving statistical modeling, hypothesis testing, and sensitivity analysis on large-scale collider datasets to detect deviations from Standard Model predictions and identify new physics signals.

BS-MS Dual Degree, Physics | *IISER, Kolkata* Aug 2012 – May 2017