# KUNTAL PAL

**Riverside, CA | +1 9512028635 | kuntal.beehiveai@gmail.com | LinkedIn | Github | Blog**

AI Engineer specializing in LLM systems and agentic infrastructure, with experience building production-grade multi-agent workflows, retrieval pipelines, and real-time AI systems that drive measurable impact on automation and user engagement.

## EDUCATION

- **UNIVERSITY OF CALIFORNIA, RIVERSIDE** **Sep. 2017 - Jun. 2024**
  **Ph.D., Physics**

- **INDIAN INSTITUTE OF SCIENCE EDUCATION AND RESEARCH, KOLKATA** **Aug. 2012 - May 2017**
  **BS-MS Dual Degree, Physics**

## TECHNICAL SKILLS

- **Programming Languages**: Python (Advanced), SQL (Advanced), Go (Intermediate)
- **ML/AI:** PyTorch, TensorFlow, Scikit-learn, LLMs, RAG, Vector Databases, LangChain, LangGraph
- **Deployment & MLOps:** Docker, Kubernetes, GCP
- **Specializations:** Multi-Agent Systems, Multimodal Learning, LLM finetuning, Predictive Modeling

## WORK EXPERIENCE

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

- Engineered a production-grade agentic RAG system using LangChain and Postgres vector search, enabling natural language to SQL generation for real-time customer analytics.
- Developed an LLM-powered QA framework for data categorization, achieving 80% automation by validating and refining category labels, reducing manual effort, and improving classification 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 semantic search, integrating dynamic UI and conversational interfaces for personalized product recommendations.

**AI Research Scientist | Pocket FM** **Jan. 2024 - Jul. 2024**

- Delivered an LSTM-based sequential prediction model leveraging episode-level engagement data and metadata to forecast long-term content performance.
- Led development of a comment ranking system using semantic embeddings, LLM summaries, and engagement signals, resulting in +4.5% CTR and +4.0% increase in comment volume.

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

- Automated reliability assessment of medical research papers using LLM-based algorithms targeting key quality metrics based on predefined categories.
- Improved GPT-4 evaluation accuracy to 75% through prompt optimization and enhanced Mixtral performance to 69% using Quantized LoRA fine-tuning.

**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.
- Benchmarked multimodal representations against BERT-based text embeddings, outperforming text-only baselines by 5 percentage points.
- Engineered a multimodal embedding space aligning audio and text representations, enabling cross-modal tasks including speech recognition and translation.