Jalend Bantupalli

Research Engineer — LLM Agents, GenAI, Evaluation, RLHF
Linkedin ⋄ +1(858) 518-3637 ⋄ jalend.bantupalli@gmail.com ⋄ Portfolio ⋄ Github

PROFILE

Research Engineer with experience shipping agentic LLM systems for multimodal understanding, speech interfaces, and personalized retrieval. Skilled in reinforcement learning, dense IR, and scalable ML infrastructure for real-time, user-aligned reasoning. Passionate about building AI agents that bridge foundational research and production, with deep experience in streaming inference, and embodied interaction across voice, vision, and motion modalities.

SKILLS

Programming Languages	Python, C++17, C, Java, Go, JavaScript
ML Frameworks	PyTorch, TensorFlow, JAX, Hugging Face, Transformers, Whisper
Infra & Tooling	TF-Serving, vLLM, Kubernetes, Redis, Docker, GCP (Vertex AI), Git, REST APIs, LangChain, LlamaIndex, FFmpeg, Weaviate
Core Competencies	LLM Agents, RLHF, model evaluation, tokenization, retrieval, ML debugging, multimodal generation, Voice assistants

EXPERIENCE

MyoLab.AI — Member of Technical Staff, NLP
New York City, New York — Jun 2024 – Present

- Shipped two agentic **LLM systems for wearable, multimodal understanding and voice interfaces**; designed streaming RAG pipelines using vLLM, OpenAI and Redis for 150+ QPS at 99.9% uptime.
- Led evaluation and performance tracking of LLM output quality across grounding, retrieval, and user interaction fidelity, incorporating **reinforcement learning feedback** loops for continual improvement.
- Actively researching **embodied GenAI systems** combining motion data and LLMs; designed tokenization and training pipeline for multimodal agent behavior.

Shang Data Lab, UCSD — Graduate Researcher
San Diego, California — Sep 2024 – May 2025

- Co-authored ACL 2025 paper on long-term personalized dialogue modeling; introduced ImplexConv, a large-scale multi-session dataset, and TaciTree, a hierarchical IR framework enabling **implicit reasoning via persona- and memory-aware retrieval.**
- Built personalized dense retrieval pipelines and evaluation suite to benchmark grounding fidelity in multi-session dialogue, optimizing for semantic relevance, latent persona modeling, and long-term user alignment.

Google — Software Engineer, Cloud Networking
Bengaluru, India — Jul 2022 – Aug 2023

- Built a black hole detection system for L3 miss packets in Gemini-based switches (˜600K devices); wrote 5K+ LOC in C/C++ to capture and debug network packet loss in production.
- Led system integration across protocol and hardware teams, reducing root-cause diagnosis time and improving observability.

Microsoft — Software Engineer Intern
Hyderabad, India — May 2021 – Jul 2021

- Built centralized configuration REST system adopted by 7 feature teams; improved deployment velocity by **40%**.

PUBLICATIONS & OPEN SOURCE

- Li, Bantupalli et al. *Toward Multi-Session Personalized Conversation.* **ACL 2025 Submission**.
- Vahini, Bantupalli et al. *Decoding Demographic Un-Fairness from Indian Names.* **SocInfo '22**.
- Agentic LLM System for Data Analysis – Open-source pipeline for LLM + motion data modeling.

PROJECTS

Sketch Tuning for Code Translation *UC San Diego, Sep 2024 – Dec 2024*

- Proposed "Sketch Tuning," a method for program synthesis using LLMs with structured intermediate representations to improve code translation.
- Evaluated performance on multilingual code tasks and demonstrated improvement over autoregressive baselines.

IMPART: Image Matching and Pairing using RA-CLIP Vanilla *UC San Diego, Apr 2024 – Jun 2024*

- Built retrieval-augmented vision-language agent with multi-head attention; improved image-text alignment for downstream pairing tasks.
- Applied to video-grounded QA and retrieval-based navigation scenarios.
- Code: github.com/Jalend15/IMPART

EDUCATION

University of California San Diego 2023 - 2025
MS in Computer Science and Engineering **4.0/4.0**
Relevant Coursework: Reinforcement Learning, Probabilistic Reasoning, NLP, Optimization, Networked Systems, Advanced Vision

Indian Institute of Technology, Kharagpur 2018 - 2022
B.Tech in Electronics and Electrical Communication Engineering **9.45/10**
Minor in Computer Science **9.56/10**
Relevant Coursework: Machine Learning, Deep Learning, Information Retrieval, Algorithms, Operating Systems, NLP