Vishnu Pulipaka

Dallas, Texas | vishnupulipaka22@gmail.com | vishnup22.github.io | linkedin.com/in/vishnup22

Research Interests

- **Multilingual and low-resource NLP:** how language models learn, represent, and transfer knowledge across languages when training data is limited, with a focus on small language models and evaluation for Hindi, Telugu, Tamil, Kannada, and Malayalam. Grounded in my work training dedicated and joint causal language models for Dravidian languages and building BabyLM-scale Hindi and Telugu corpora through machine translation.
- **Embodied multilingual intelligence (future direction):** extending multilingual models toward multimodal grounding, vision-language-action systems, and multilingual human-robot interaction, so that robots and edge devices can understand and act on instructions across languages.
- **Language variation and evaluation:** how dialects, regional forms, and scripts affect tokenization, representations, and evaluation, including disparities hidden by aggregate benchmark scores. Currently studied through dialect variation in Telugu and Hindi and native versus romanized text.
- **Efficient models for constrained settings:** training and serving language models under limited compute, multi-GPU training, ONNX, CUDA, and TensorRT, with the longer-term aim of running multilingual models on edge hardware.

Education

Montclair State University — *Graduated January 2026*
M.S. in Data Science, ***GPA: 3.87/4.0*** — United States

Sreenidhi Institute of Science and Technology — *Graduated 2024*
B.Tech in Electronics and Communication Engineering — India

Research

- **Can Translation Bridge the Data Gap for Low-Resource Languages? A Case Study of Hindi and Telugu**
  Translated the 100M-word BabyLM corpus into Hindi and Telugu using IndicTrans2; trained and evaluated monolingual and bilingual GPT-Wee and GPT-BERT models to study whether machine-translated data can supply training data for small language models under data scarcity. *BabyLM Workshop, 2026.* [Code] [Data]
- **Evaluating Dedicated Monolingual and Joint Multilingual Causal Models for Dravidian Languages**
  Trained five GPT-2-style causal language models from scratch for Telugu, Tamil, Kannada, and Malayalam (four monolingual with dedicated 32K tokenizers, one joint multilingual with a shared 64K tokenizer); compared tokenizer efficiency, bits-per-byte, and downstream sentiment and NER performance to study the trade-offs between shared and dedicated models. *arXiv:2608.07727, 2026.* [Paper] [Code]
- **Reproducible Fairness Regression Testing for Large Language Model Updates**
  Developed a regression-testing framework that compares LLM releases on matched prompts; evaluated 19 model variants from 4 providers on 7 bias benchmarks (12,789 scored responses), tracking sentiment, toxicity, and stereotype-related drift across release transitions. *Under Review, 2026.* [Code]
- **Script and Morphology Effects on Telugu and Hindi Tokenization Are Tokenizer-Dependent**
  Compared 8 tokenizers on formal, informal, native-script, and romanized Telugu and Hindi text, plus 144 morphological forms, to measure how structure affect token fertility. *Under Review, 2026.*

Experience

Technical Intern I **at WorldLink, United States** — *September 2026 – Present*
- Building an enterprise Network Intelligence Platform.

AI/ML Intern **at VTS360, United States** — *August 2026 – September 2026*
- Built and evaluated RAG pipelines using Python, PyTorch, and Hugging Face Transformers for LLM-powered applications and domain-specific knowledge retrieval.
- Developed agentic AI systems and LLM workflows for enterprise use cases using Claude Code, LangChain, and cloud-based infrastructure.
- Supported MLOps workflows using Azure DevOps, MLflow, and Prometheus for experiment tracking, CI/CD, model evaluation, and production monitoring.
- Contributed to data engineering and backend integrations using scikit-learn, NumPy, Pandas, and REST services while communicating technical architecture and system behavior to stakeholders.

Data Engineer (Volunteer) **at Saayam For All, United States** — *March 2026 – August 2026*
- Built and hardened AWS Lambda workflows with input validation, exception handling, and reliable PostgreSQL connectivity.
- Migrated CSV-based data workflows to PostgreSQL-backed APIs and supported automated data validation for application dashboards.
- Improved backend and data engineering workflows with a focus on reliability, maintainability, and cleaner application data flows.

Software Engineer **at 4SS Software Solutions, India** — *June 2022 – July 2024*
- Built scalable microservices and REST APIs in Python, Java, and JavaScript on AWS, adding Redis caching that reduced latency by 25% under growing service load.
- Designed Kafka-based asynchronous workflows to decouple services and improve event-driven processing, reducing interservice processing time by 22%.
- Built AI-assisted product features using LLMs to automate language-based workflows and improve internal data processes.

- Optimized SQL-backed data layers and refactored core services, reducing post-release defects by 15%, while building CI/CD infrastructure with Jenkins that reduced deployment time by 20%.
- Collaborated across frontend and backend development using JavaScript, React, Docker, SQL, and cloud services to ship and maintain production applications.

Technical Skills

- **Languages:** Python, C++, SQL
- **Language Modeling & NLP:** PyTorch, Hugging Face Transformers, Hugging Face Accelerate, SentencePiece tokenizer training, IndicTrans2, training causal language models from scratch
- **Evaluation & Statistics:** perplexity, bits-per-byte, multilingual benchmarks (MultiBLiMP, SIB-200, WikiANN), paired bootstrap confidence intervals, Wilcoxon signed-rank tests, Benjamini–Hochberg and Holm correction
- **Training & Inference Systems:** multi-GPU training (NVIDIA A100), mixed precision (bf16), CUDA, TensorRT, ONNX Runtime
- **AI Systems & Infrastructure:** RAG, agentic AI, MCP, LangChain, OpenAI Agents SDK, FastAPI, Docker, Kubernetes, AWS, Azure, Prometheus
- **Data & MLOps:** Pandas, PySpark, PostgreSQL, MLflow