POORNA PRANEESHA D

(+91) 8985 650250 ◇ poorna7d@gmail.com ◇ Linkedin ◇ GitHub ◇

Senior AI Engineer at Hivepath with 6+ years in industrial roles. Building AI solutions to enhance customer experiences.
Focus Areas: *Deep Learning, Generative AI, PEFT (LoRA, QLoRA, RLHF), RAG, AgenticRAG.*

EDUCATION

B.Tech. Chemical Engineering - Sri Venkateswara University, India *June 2013 - May 2017*
Courses: Data Structures, Algorithm, Calculus, Python Programming, MATLAB
CGPA: 7.29/10 (Distinction)

Online. Data Structures Specilization - University of Colorado Boulder, Online CourseWork *2024-present*
Courses: Data Structures, Dynamic Programming, Linear Programming Algorithm, Greedy Algorithm, Linear Algebra.
Grade: 100%

CERTIFICATIONS / COURSE WORK

Microsoft Azure AI Fundamentals Link
Generative AI with Large Language Models Link
Supervised Machine Learning Link
Approximation Algorithms and Linear Programming Link
Databricks - Generative AI Solution Development Link
Advanced Learning Algorithm Link
Dynamic Programming, Greedy Algorithms Link
Algorithms for Searching, Trees and Graphs Link

WORK EXPERIENCE

Senior AI Engineer(GenAI) Sep 2021 - Present
Hivepath *Bangalore, India*

Working towards enabling intelligence in autonomous AI Agents for business.

Agentic RAG Applications:

- Architected and deployed a multi-agent Retrieval-Augmented Generation (RAG) platform integrating semantic routing, functional tooling, and domain-specific retrieval strategies, enabling dynamic query handling.
- Engineered a robust ingestion pipeline for structured and unstructured data, leveraging tools like Unstructured.io, LlamaParser, and PyPDF for parsing and metadata enrichment. Selected and implemented effective chunking strategies to optimize computational costs and reduce response latency.
- Developed multi-step query optimization techniques, including query translation, decomposition, and semantic expansion, to enhance coverage and improve ranking stability across multiple retrievers.
- Deployed an observability stack with Langfuse for real-time query tracing, retrieval quality assessment, answer relevance tracking, and continuous performance tuning.
- Implemented a multi-layered guardrail system to ensure application safety, reliability, and ethical operation. This system included Input Rails to filter harmful or sensitive user queries, Retrieval Rails to protect against sensitive information retrieval from the knowledge database, and Execution Rails to mitigate risks associated with external tool and plugin usage.
- Optimized runtime efficiency through semantic compression pipelines, redundancy elimination, and selective context injection, reducing LLM token usage by ∼30% while preserving content fidelity.

AI Model Development:

- Involved in the entire spectrum of AI model development, including conceptualization, deployment, instruction tuning, RLHF, and parameter-efficient fine-tuning.
- Leveraged API interaction, prompt engineering, and Retrieval-Augmented Generation (RAG) to define the landscape of autonomous AI agents.
- Constructed the End-to-End pipeline, integrating multiple Language Model (LLM) models (GPT-4, Llama 2).

Continuous Improvement and Responsible AI Implementation:

- Contributed to the gathering, construction, and annotation of domain-specific datasets. Enhanced the training of LLMs for a wider range of tasks and applications.
- Adapted and integrated a small LLM to enhance the ecosystem of responses.Actively measured and benchmarked model and application performance, analyzing accuracy and bias.
- Established a framework for questions and responses, conducting LLM evaluations based on question coverage and response quality. Implemented continuous improvements in AI models and maintained model evaluation systems.

AI Engineer Sep 2020 - Aug 2021
ShopConnect *Bangalore,India*

- Implemented CNN based architecture models(ResNet, VGGNet) for object classification problems.
- Implemented YOLO algorithm for applying in retail domain - Customer Behavior Analysis, Foot Analytics, Inventory stock analysis.
- Presented recommendations on retail analytics based on computer vision technologies - Monitoring store traffic (at various customer touch points)..

Intern May 2019 - June 2019
Crafter *Hyederabad,India*

- As the sole developer on the project, assumed full responsibility for leading the engineering strategy and making technology decisions independently, ensuring efficient and effective project execution...
- Proactively seeking feedback and input from the executive team and reflecting in the product Frontend..

Software Engineer June 2017 - April 2019
Bluecom *Bangalore,India*

- Engineered an AI-powered e-commerce tool for seamless multi-channel Product Information, Inventory and POM .
- Integrated multiple third-party APIs, for product, inventory, order data from Shopify, Woocommerce, and Bigcommerce. Optimize backend systems and database architecture for scalability to handle increased data volumes.
- Optimized and automated business logic for the core marketing experiments, including A/B testing..

TECHNICAL SKILLS

AI Platforms	Huggingface, langchain, OpenAI,Langfuse
Observability	Evalutaion platforms : Arize, Langufuse Programming languages/ packages
Python, R, Reactjs, Nextjs, Java, Scala, Haskell, Javascript	
Deep learning	Pytorch, Tensorflow, ONNX, CUDA, torchvision, Hugging Face, Langchain, Pi
Database	Oracle database, MySQL, GitHub, Bitbucket, CircleCI, Heroku, Mandrill,Dock

WHITE PAPERS

Autonomous Multi-Agent System Using LLM Link
AI Compute, a Technical Deep Dive Link
Data Loss Protection Link
Computer Vision Application Link
AI Inference Link
MultiChannel Ecommerce Link
Messaging App Link
Data Tech Stack Link
Transformers Variants Link
Chat Bot Using Deep Learning Link

AI PROJECTS

AI Agents - LangChain GitHub Link

- Implemented end-to-end Retrieval-Augmented Generation (RAG) using Langchain and Openai APIs.

Key Skills: Python, Langchain, RAG Systems, VectorDB, Inferences.

Transformer Architecture - GPT GitHub Link

- **Tokenization:** Implemented BPE with regex preprocessing to optimize input text representation. BPE reduces vocabulary size by 40% through iterative merging of frequent byte pairs, while regex normalization enhances segmentation and addresses rare words. This integration results in a 30% reduction in out-of-vocabulary errors.
- **Multi-head Attention(MHA):** Integrated MHA enables the model to attend to various parts of the input sequence simultaneously, dividing embeddings into query(q), key(k), value(v), resulting in more coherent and relevant text generation.

Key Skills: Python, Transformer Architecture, Multi-Head Attention, PyTorch, NLP, Model Training, Tokenomics.

Byte Pair Encoding(BPE) - GPT GitHub Link

- **BPE:** Adapted BPE with regex preprocessing to optimize input text representation. BPE reduces vocabulary size by 40% through iterative merging of frequent byte pairs, while regex normalization enhances segmentation and addresses rare words.
- This integration results in a 30% reduction in out-of-vocabulary errors and boosts encoding efficiency by 25%.

Key Skills: Tokenomics, Regex, PyTorch, Python

Tiny Machine Learning - TinyML GitHub Link

- Gained expertise in model quantization techniques, reducing precision and optimizing models for low-power, memory-constrained devices, alongside transfer learning for adapting pre-trained models to acoustic, image, and sensor data.
- Developed TinyML applications in image recognition, audio processing, and gesture detection using Arduino hardware, focusing on the efficient deployment of machine learning models to edge devices.

Key Skills: Python, Quantaisation, TensorFlow, TensorFlow Lite, TensorBoard, Google Colab, and Jupyter Notebook

- Leveraged the OpenAI API to dynamically generate blog content, utilizing prompts and context for coherent and contextually relevant posts, emphasizing iterative refinement and fine-tuning for optimal output.Developed a blog editor and grouped them into segments, like ”Frequent Shoppers” or ”Tech Enthusiasts,” based on behaviours and preferences.

ACHIEVEMENTS

Best Employee Award Certificate	Received the Best Employee Award in June 2022.
WorldQuant Certificate	Outstanding Performance Award (**3rd Position**).