Satyanarayana Reddy Kovvuri

Business Analyst

kovvuri.satyanarayanareddy@cognizant.com
+91 8977409021
https://www.linkedin.com/in/kovvuri-satyanarayana-reddy-439284170

Summary

With 2+ years of experience in natural language processing and machine learning, I have gained expertise in developing large-scale models for text classification, summarization, and question-answering tasks. I have hands-on experience with T5, BART, GPTNeo, and DeBERTa models and have employed techniques such as transfer learning and few-shot learning improve model performance. My background includes working on projects like Myntra product classification, interview classification, and other research oriented work. I am passionate about constantly learning and applying new techniques to enhance the capabilities of language models.

Areas of expertise

- Natural Language Processing
- Large Language Models
- Machine Learning

Work experience

Cognizant
July 2021 – Present

Programmer Analyst
Utilized and fine-tuned the cutting-edge T5 language model to enhance product categorization for an e-commerce organization.

Employed saliency mapping as a visualization technique to comprehend language model predictions.

Designed and executed extractive summarization methods using GPT2.

Applied advanced prompt engineering methodologies such as Chain of Thought, Multi-agent Debate, and Reflexion to achieve 80% accuracy on various benchmarks, including AGIEval.

Developed and integrated a comprehensive backend system for an Interview Bot, leveraging GPT-4, Redis-search, MongoDB, and Azure Functions for efficient candidate evaluations through few-shot learning.

Gained expertise in fine-tuning language models on multi-GPU environments utilizing state-of-the-art packages such as Accelerate, Deepspeed, FSDP, Peft Lora, and Quantization.

Incorporated Retriever Augmented Generation for specific use cases, employing language models and vector databases like Faiss.

Cognizant
Mar 2021 – Jun 2021

Intern/Programmer Analyst Trainee
During internship at Cognizant, I gained hands-on experience in Datascience and Cloud technologies (Azure), having completed relevant Udemy courses such as Python, Machine learning, SQL, Azure cloud, Docker, Kubernetes.

I successfully applied my knowledge to real-world project to detect accidents in real time, demonstrating strong problem-solving skills and adaptability. This experience enhanced my understanding of industry best practices and allowed me to excel in my exams.

Education

M.Tech, BITS Pilani, WILP, India
2023 – 2025

Currently Pursing master's in data science (WILP).

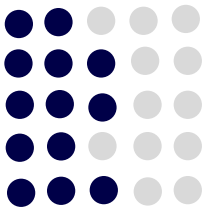
B.Tech, Gayatri Vidya Parishad college of Engineering, Vizag, India
2017 – 2021

Completed bachelor's in computer science and engineering with 7.69 CGPA

Language

Python
C
SQL

Skills



Project experience

Cognizant, India
Research

Myntra Product Classification.

- Trained T5 Language model on product classification dataset and applied saliency mapping for visualization; evaluated using Bleu score and Perplexity achieving 96% accuracy.
- Generated additional data using GPTNeo 1.3B model to make dataset uniform through few-shot learning.
- Implemented Information Bottleneck regularizer (InfoBERT) on T5 Language model.
- Utilized Sentence-T5 Language model to obtain nearest neighbors within the product classification dataset.

Cognizant, India
Internal

Interview Bot.

- Conducted 20 hours of interview labeling in Label Studio.
- Trained Flan-T5-xl and Long-T5-global on interview dataset for candidate acceptance/rejection classification; using Accelerate and performed inference using Deepspeed stage-3. achieved 75% accuracy.
- Applied PMI (Pointwise Mutual Information) for interview summaries and utilized Flan-T5 and Long-T5 for token reduction.
- Created sentence embeddings for speaker tags in interviews and applied KDS clustering for summary generation.
- Implemented few-shot learning with GPT-JT-6B and Flan-T5-xxl to generate appropriate interview questions. (First iteration)
- Extracted QA pairs and summaries from raw interviews using Flan-T5, GPT-JT, GPT-neox-20B, and GPT-3 with various prompts.
- Created QA pairs and summaries using Davinci model and trained separate Curie models for QA pairs, summaries, and bullet points.
- Worked on GNG (Growing Neural Gas) clustering algorithm to obtain 10,000 clusters; optimized silhouette score using spectral clustering to remove unnecessary questions for interview bot.
- Trained T5-3B and T0-3B models for candidate acceptance/rejection classification based on extracted bullet points.
- Obtained Ada-embeddings of 10,000 QA pairs and implemented Redis Search for few-shot interview applications.
- Integrated Azure Functions for open access to the interview bot in UI and resolved access issues with OpenAI models.
- Implemented GenC bot using chatgpt and GPT-4 APIs and GPT-4-based classification task for interview bot with chatlog history.
- Utilized MongoDB in Azure for storing chatlogs and tested various prompts for interview performance.
- Implemented an HR Bot using GPT-4 for demo purposes, with minor modifications to the technical interview code.

Cognizant, India
Research

Additional Reseach Tasks:

- MCA Question-Answer Classification: Trained Flan-T5, DeBERTa, ERNIE, and SVM Classifier on top of T0-3B sentence embeddings for question-answer correctness classification. Achieved up to 81.6% accuracy with DeBERTa model.
- Applied BottleSum Extractive Summarization on CNN Dataset.
- Trained T5, BART, and GPTNeo Language models on text summarization tasks using extractive summaries produced by BottleSum; evaluated using BLEU, ROUGE, and BERT scores got 0.78 score.
- Explored Chain of Thought, Multi-agent-debate, and Reflexion techniques on benchmark datasets such as JEEBench and AGIEval achieving 35% and 80% accuracy respectively.
- Studied GPU usage of various language models with different context lengths using advanced finetuning techniques like Peft-Lora.
- Worked on Starcoder-8K implementation, focusing on Peft-LORA and 3D Parallelism, experimented with Bloom-3D parallelism, and attempted Qlora quantization for larger context lengths on multiple GPUs