Google Summer of Code 2025 Proposal

Title

Optimizing Retrieval-Augmented Generation Pipelines for Accurate and Efficient Question-Answering Systems

Personal Information

- **Name:** Gopi Trinadh Maddikunta
- **University:** University of Houston
- **Program:** Master's in Engineering Data Science
- **Email:** trinadh7341@gmail.com
- **LinkedIn:** linkedin.com/in/gopitrinadhmaddikunta/

Synopsis

I'm passionate about improving Retrieval-Augmented Generation (RAG) systems to solve real-world problems effectively. In this project, I'll optimize embedding methods, streamline retrieval processes with FAISS, and fine-tune GPT-Neo 1.3B for precise and contextually relevant question-answering. I'll use rigorous evaluation tools, particularly Ragas, to ensure measurable improvement. My experience with NLP frameworks, such as HuggingFace and LangChain, combined with a structured, iterative approach, positions me well to deliver robust outcomes that benefit the broader open-source community.

Benefits to the Community

This project aims to significantly enhance open-source RAG-based systems by contributing an optimized pipeline for efficient, context-rich, and reliable question-answering. The improvements to embedding strategies, LLM fine-tuning, and retrieval mechanisms can be reused across a variety of projects and applications, from customer support bots to research assistants. The outcomes will be openly documented and shared for use and further enhancement by the community.

Deliverables

- A production-ready, modular Retrieval-Augmented Generation (RAG) pipeline.
- A fine-tuned GPT-Neo 1.3B model integrated with optimized retrieval and embedding modules.
- Evaluation report using Ragas with metrics like Faithfulness, Answer Relevance, Context Precision, and Recall.

- Complete documentation and usage guide.
- Code contributions merged into the organization's repository.

Timeline

Community Bonding Period (May 20 – June 16)

- Interact with mentors and the community
- Review existing codebase and literature
- Finalize design and setup the environment

Phase 1 (June 17 – July 15)

- Implement and evaluate advanced embedding models (HuggingFaceEmbeddings, SentenceTransformers)
- Integrate FAISS for fast and scalable retrieval
- Run retrieval benchmarking and establish baseline metrics

Phase 2 (July 16 – August 12)

- Fine-tune GPT-Neo 1.3B using context-rich retrieved documents
- Test the complete RAG pipeline end-to-end
- Midterm evaluation and refinements based on feedback

Phase 3 (August 13 – September 9)

- Run extensive evaluations using the Ragas library
- Refine retrieval and generation based on performance metrics
- Finalize documentation and submit project report

Technical Details

- **Languages:** Python
- **Libraries & Frameworks:** PyTorch, Hugging Face Transformers, LangChain
- **Retrieval:** FAISS
- **Embeddings:** HuggingFaceEmbeddings, SentenceTransformers
- **Evaluation:** Ragas
- **Development Tools:** Git, GitHub, Jupyter Notebook, Google Colab

Project Topics

Machine Learning, Natural Language Processing, Retrieval-Augmented Generation, Information Retrieval, Question-Answering, Embeddings, Open Source

Project Size

Large (350-hour)

Biography

I’m a graduate student in Engineering Data Science at the University of Houston, with a background in Computer Science and a growing focus on Natural Language Processing. I’ve worked on projects involving LangChain, FAISS, and Hugging Face Transformers, and am actively involved in developing RAG pipelines for robust question-answering. I’m committed to open-source contributions and excited to work with the community to build something meaningful and impactful.

Commitments and Availability

I will be available full-time during the GSoC period with no other internships, courses, or job obligations. I am fully committed to dedicating 30–35 hours per week to this project.

Contributions

- Built an evaluation pipeline for RAG using LangChain and Ragas
- Developed a RetrievalQA system using FAISS and HuggingFaceEmbeddings
- Actively contributing to a university-led project on benchmarking LLMs

---

Thank you for considering my proposal for GSoC 2025. I am excited to contribute to impactful open-source NLP tools and work alongside passionate mentors and developers.