SETSCI conference proceedings

*SETSCI Conference Proceedings*
*6 (1), 81-85, 2023*

*Cognitive Models and Artificial Intelligence Conference*

*26-28 October 2023, Ankara, Türkiye*

https://doi.org/10.36287/setsci.6.1.029


BMYZ 2023

# FlexiGPT: Engaging with Documents

Abdalrhman Alquaary*, Numan Çelebi [2]

[1]*Information Systems Engineering, Sakarya University, Sakarya, Turkiye*
[2] *Information Systems Engineering, Sakarya University, Sakarya, Turkiye*
*(apoalquaary@gmail.com) *Email of the corresponding author*

***Abstract –*** Leveraging the robust potential inherent in large language models, their profound and pervasive impact has transcended various domains in recent years, ushering in their widespread integration across diverse sectors. In resonance with this predominant trend, the current study introduces an innovative application that endows users with the capacity to actively engage in conversations with their digital files. This program integrates state-of-the-art large language models with the techniques of Retrieval Augmentation, thereby crafting an immersive and sophisticated experience that not only amplifies but fundamentally elevates user engagement to new heights of interactivity and responsiveness. Functioning as a pivotal nexus, Hugging Face, a renowned platform for machine learning models, assumes the role of the primary repository and catalyst for these transformative language models. Through the medium of this application, users can have interactive engagement, perfectly aligned with the continually evolving tapestry of linguistic technology and digital interaction. Significantly, users possess the freedom to choose from an extensive array of open-source large language models available on the Hugging Face platform, thereby, they also retain the option to seamlessly update to newer models as they become available, ensuring continuous access to the latest large language models and maintaining the applicability of the application in line with evolving user needs. Importantly, the operational viability of the program is extended to local execution, contingent upon the availability of sufficient hardware resources.

*Keywords* – LLM, Retrieval Augmentation, Hugging Face, Langchain, Local LLMs

## I. Introduction

The field of Natural Language Processing (NLP) has undergone a remarkable transformation in recent years, driven by the proliferation of Large Language Models (LLMs). These models, endowed with the ability to comprehend and generate human language, have transcended disciplinary boundaries, ushering in a new era of possibilities across various sectors. The impact of NLP advancements is palpable, extending from document analysis and summarization to chatbots and code analysis. It is within this dynamic landscape that we situate our project.

Central to our exploration is the LangChain framework, a remarkable open-source tool that simplifies the development of applications powered by LLMs [1]. Launched in October 2022 by Harrison Chase during his tenure at Robust Intelligence, LangChain offers developers a versatile solution. It streamlines the integration of LLMs into applications, effectively democratizing the capabilities of these models. Its core objective is to empower developers with the tools needed to create applications that harness the power of LLMs and connect seamlessly with data sources.

At the heart of our endeavor is the application we have developed, which leverages the LangChain framework to facilitate human-file interaction. This application represents a practical realization of the transformative potential of NLP advancements. Users can now engage in dynamic, interactive conversations with their digital files, transforming these repositories into responsive knowledge hubs. Notably, our application provides users with the flexibility to select LLM models and embedding models from the Hugging Face platform, aligning their interactions with specific domains and preferences [2] [3].

In essence, this project introduces a user-centric and developer-friendly approach to harnessing the power of large language models (LLMs) in conjunction with retrieval augmentation for dynamic interactions with digital files. Our application streamlines the complex process of integrating LLMs into file-based interactions, offering developers a practical toolkit and end-users an intuitive platform. By democratizing access to LLMs and retrieval augmentation techniques, we empower both developers and users to engage in seamless, conversational exchanges with their files, thus bridging the gap between sophisticated NLP capabilities and practical utility. This contribution not only underscores the transformative potential of LLMs but also aligns with the ongoing evolution of NLP in enhancing human-file interaction. The application is available in github [4].

## II. Related Works

The field of natural language processing has witnessed remarkable advance-ments in recent years, driven by the development of increasingly sophisticat-ed Language Model Models (LLMs). This progress can be traced back to the idea "Attention is All You Need" [5], which introduced the Transformer archi-tecture. The Transformer marked a paradigm shift in the way we approach natural language understanding tasks, primarily through the incorporation of self-attention mechanisms that allowed models to consider global context de-pendencies efficiently.

In recent years, there has been a significant and ongoing enhancement in the performance of Large Language Models (LLMs), primarily attributed to the expansion of both high-quality data and model parameters. This evolution has heralded a genuine revolution in the domain of large-scale language model-ing. Notably, with the introduction of ChatGPT by OpenAI, the global interest in LLMs has surged, driven by their remarkable capacity to surpass the capa-bilities of traditional language models. Despite the formidable cost associated with training large language models, several forward-thinking organizations have undertaken the challenge, contributing to the open-source ecosystem by making these models publicly available. Leading the forefront of open-source LLMs is LLama-2, a prominent model accessible through the Hugging Face platform. This model exists in three iterations, distinguished by their parame-ter sizes: 7 billion, 13 billion, and an astounding 70 billion parameters. These publicly accessible models have ignited a plethora of innovative applications that harness the potential of LLMs, fundamentally reshaping the landscape of natural language understanding. The ubiquitous integration of LLMs into var-ious applications underscores their transformative impact, solidifying their position as a foundational element in contemporary natural language pro-cessing and AI-driven systems. One notable application focuses on the evalu-ation of large language model (LLM)-based chat assistants. In a previous study, researchers delved into the utilization and inherent constraints of em-ploying Large Language Models (LLMs) as judges. They presented innovative strategies to mitigate potential biases in LLM judgments and introduced two benchmarking mechanisms to gauge the alignment between LLM judgments and human preferences. The findings of this investigation highlighted the fea-sibility and interpretability of utilizing LLMs as evaluative proxies for ap-proximating human preferences effectively [6].

The foundation of Retrieval Augmented Generation (RAG) lies in the evolu-tion of text embeddings, a critical component in understanding and generating natural language text. The journey of text embeddings began with the advent of word vectors, which aimed to represent words as continuous, semantically meaningful vectors in a high-dimensional space. Word2Vec and GloVe were pioneering models in this regard, enabling the capture of word semantics and relationships [7], [8]. However, the field of embeddings rapidly evolved, spurred by the introduction of models like BERT (Bidirectional Encoder Rep-resentations from Transformers), GPT (Generative Pre-trained Transformers) and BGE (Bidirectional Generative Embeddings). BGE has achieved state-of-the-art results on several benchmarks, such as the TREC-COVID challenge. These state-of-the-art models revolutionized the way we represent text, offer-ing contextual embeddings that could capture nuanced word meanings and their relationships within sentences and documents. This transition from tra-ditional word vectors to context-aware embeddings marked a significant leap in natural language understanding. Harnessing these embeddings within the context of RAG has opened new frontiers in text generation and information retrieval, allowing systems to seamlessly blend the power of language models with the depth of semantic understanding.

## III. Materials and Method

In this study, we aim to create a system that allows users to interact with vari-ous document types, including PDFs, text files, and other document formats. The methodology consists of several interconnected steps designed to facili-tate this interaction (Fig. 1).

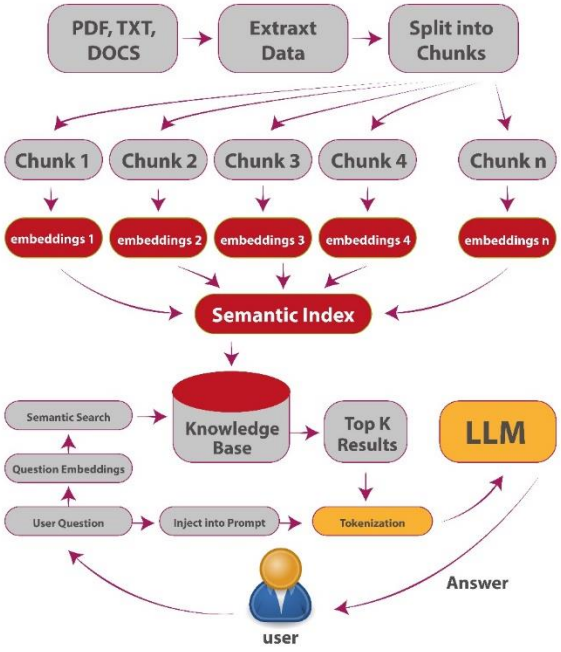


Fig. 1. The Process of the program

**Data Collection and Preparation:**

The initial step involves data collection and preparation. We assume that all data files are located in a single directory, which is by default named 'docs' and resides in the same folder – path – as the program. The program is responsible for extracting data from this directory, ensuring that the data is readily available for further analysis.

**Data Chunking:**

Once the data is extracted, it is transformed into manageable chunks. These chunks are created based on predefined criteria, which may include word count or paragraph boundaries. The purpose of this step is to break down the documents into smaller, digestible portions, making them easier to process.

**Embedding Models:**

A critical aspect of our methodology is the flexibility in choosing embedding models. Users have the option to select an embedding model from the Hugging Face Transformers library. The program is designed to dynamically download and apply the selected model to the data. It is important to note that the choice of the embedding model significantly influences subsequent stages of the process.

**Semantic Index and Knowledge Base:**

Following the embedding of data, a semantic index and knowledge base are created. These components serve as the

backbone of our system, allowing for efficient retrieval of relevant information. The semantic index is formed based on the embeddings, while the knowledge base stores critical information from the documents. This step plays a pivotal role in retrieving the most related chunks when we apply search similarity with the taken question from users.

**User Query Processing:**

In parallel with the above processes, the user interacts with the system by entering questions or queries. The user's question is embedded using the same selected embedding model. A semantic search is conducted to identify the most relevant document chunks in response to the user's query. Users also have the flexibility to specify how many relevant topics they want to retrieve, giving them control over the depth of information they receive.

**Engineered Prompt Generation:**

To further enhance the quality of responses from the large language model, the user's question is injected into an engineered prompt. This engineered prompt is carefully crafted to optimize the model's understanding of the user's query, thereby improving the relevance and accuracy of the generated answers.

**Model Input and Output:**

In the final step, we employ a technique called QLoRA to minimize the hardware requirements necessary to run Large Language Models (LLMs) [9], [10]. This innovative approach quantizes the weights of the model into 4 or 8 bits, reducing the computational demands significantly while preserving model performance.

Both the related document chunks and the engineered prompt are tokenized, and these tokenized inputs are then fed into the selected large language model with quantized weights. The model processes this input and generates an output, which is presented as the answer to the user's query, completing the user interaction loop. This efficient utilization of QLoRA empowers users to engage in an iterative process of querying the model within a while loop, providing a seamless and responsive experience while conserving hardware resources. Users have the flexibility to exit the program at their discretion, all while benefiting from the hardware-efficient capabilities of QLoRA.

Our methodology combines data extraction, chunking, embeddings, semantic indexing, user query processing, prompt engineering, and model interaction to enable users to seamlessly interact with document content using state-of-the-art language models and retrieval augmentation techniques.

## IV. RESULTS

In the context of this study, the application that integrates state-of-the-art large language models with Retrieval Augmentation techniques was subject to comprehensive experimentation. The aim was to assess the effectiveness and user experience of this innovative approach to digital interaction.

The experimental results indicated several noteworthy findings:

Enhanced User Engagement: In the early stages of testing the console-based application, it became evident that interaction with digital content experi-enced a significant enhancement. Conversations with digital files within the console interface exhibited a notable increase in immersion and sophistica-tion, resulting in fundamental improvements in interactivity and responsive-ness. These initial observations underscore the untapped potential of large language models in transforming user experiences across various domains through a console-based interface. This highlights the capacity of this ap-proach to pioneer novel and groundbreaking digital interactions, even in the absence of specific user participants. In the following instances, the LLM lacked knowledge of "BERTopic." However, providing the LLM with the PDF article titled "BERTopic: Neural topic modeling with a class-based TF-IDF procedure" enabled the program to answer the question, "What is BERTopic?" This highlights our program's ability to access and provide relevant infor-mation from digital documents, even in cases where prior knowledge is lim-ited (Fig.2) [12]. We can run the program using the command, selecting the LLM model 'meta-llama/Llama-2-7b-chat-hf' and the embedding model 'sen-tence-transformers/all-MiniLM-L6-v2':

> python flexiGPT.py --llm_model meta-llama/Llama-2-7b-chat-hf --embedding_model sentence-transformers/all-MiniLM-L6-v2 (1)

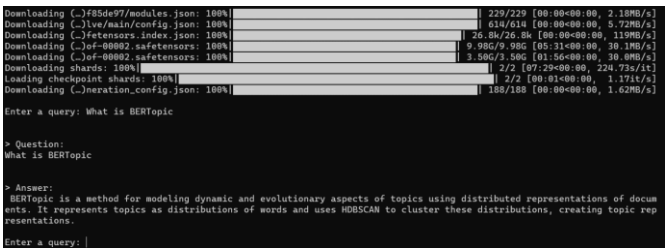


Fig. 2. Example of FlexiGPT output

In the second example, when we inquired about how BERTopic functions, the program provided a comprehensive and accurate response (Fig. 3).

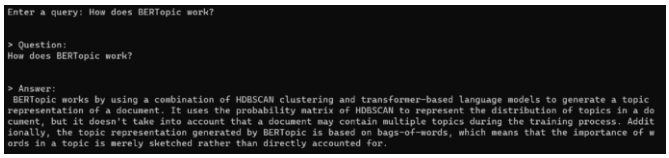


Fig. 3. Example 2 of FlexiGPT output

However, if we pose a question unrelated to the provided data, the program will respond by indicating that it lacks knowledge on the topic (Fig. 4).


Fig. 4. Example 3 of FlexiGPT output

Diverse Model Selection for LLM Models: In the context of this application, users had the flexibility to select from a

diverse range of open-source Large Language Models available on the Hugging Face platform. Notably, the default Large Language Model employed within the application was the highly capable Llama-2-7b-chat-hf. However, users also had the option to explore and choose from various other LLMs, each with its unique set of capabilities, enabling them to finely tailor their interactions to specific tasks and preferences (Table 1) [11]. This experimentation illuminated the critical role of LLM selection, showcasing that the choice of model significantly influenced the quality of conversational interactions and the depth of content understanding [2].

Tabel 1. Best Large Language Models in HF

| Model Name | Average | Arc | HellaSwag | MMLU | TruthfulQA |
|---|---|---|---|---|---|
| uni-tianyan/Uni-TianYan | 73.81 | 72.1 | 87.4 | 69.91 | 65.81 |
| Riiid/sheep-duck-llama-2 | 73.69 | 72.35 | 87.78 | 70.82 | 63.8 |
| fangloveskari/ORCA_LLaMA_70B_QLoRA | 73.4 | 72.27 | 87.74 | 70.23 | 63.37 |
| budecosystem/genz-70b | 73.21 | 71.42 | 87.99 | 70.78 | 62.66 |
| garage-bAInd/Platypus2-70B-instruct | 73.13 | 71.84 | 87.94 | 70.48 | 62.26 |

Diverse Model Selection for Embedding Models: In addition to LLMs, users could also choose from a variety of embedding models, with the default being the BAAI/bge-base-en embedding model (Table 2). These embedding models played a pivotal role in enhancing the application's contextual understanding of digital files. The experimental results highlighted that the choice of embedding model significantly enriched the application's capacity to furnish users with comprehensive and contextually relevant information during their interactions. This dual-layered diversity in model selection, encompassing both LLMs and embedding models, offered users a powerful toolkit to tailor their digital interactions according to their unique requirements and preferences [3].

This dual-layered diversity in model selection not only enhanced user engagement but also showcased the adaptability of the application across different domains and use cases. It allowed users to harness the strengths of both LLMs and embedding models to optimize their digital interactions.

Continuous Model Updates: The ability to seamlessly update to newer models as they became available was a notable feature. This ensured that users had access to the latest advancements in large language models, keeping the application relevant and in line with evolving user requirements.

Tabel 2. Best Embedding Models in HF

| Embedding Model Name | Size | Embedding Dimensions | Sequence Length |
|---|---|---|---|
| bge-large-en | 1.34 | 1024 | 512 |
| bge-base-en | 0.44 | 768 | 512 |
| gte-large | 0.67 | 1024 | 512 |
| gte-base | 0.22 | 768 | 512 |
| e5-large-v2 | 1.34 | 1024 | 512 |

Local Execution: The operational viability of the program, with the caveat of sufficient hardware resources, extended to local execution. This local execution option provides users with additional flexibility and control over their interactions.

The comprehensive analysis of the experimental results presented in this study serves as a compelling testament to the transformative power of harnessing large language models in conjunction with Retrieval Augmentation techniques to craft a profoundly enriching user experience within the realm of digital interaction. The insights gleaned from this research strongly imply that these innovative applications possess not only the inherent potential but also the practical capability to redefine the very paradigms by which users engage with their digital files. By seamlessly blending the prowess of state-of-the-art language models and sophisticated retrieval augmentation strategies, this groundbreaking approach achieves a multifaceted augmentation of user interactions, encompassing heightened interactivity, unparalleled adaptability tailored to specific user needs, and an unobstructed conduit to the forefront of linguistic technology, where the latest advancements in large language models stand ready to amplify the digital experience. This study, thus, positions such applications at the forefront of digital innovation, poised to usher in an era where user engagement transcends prior boundaries, seamlessly integrating cutting-edge language technologies to create an immersive and responsive digital landscape that aligns seamlessly with evolving user expectations and preferences.

## V. DISCUSSION

The development and deployment of our application, designed to facilitate dynamic interactions with digital files using Large Language Models (LLMs) and Retrieval Augmentation, have provided us with valuable insights. In this discussion chapter, we delve into key factors that influence the quality of responses generated by our application. It is evident that the quality of responses is intimately tied to several critical elements, with the choice of LLMs, particularly their size and associated hardware requirements, standing out as a prominent factor.

**Model Size and Response Quality:**

One of the fundamental factors that significantly impacts the effectiveness of our application is the choice of LLMs. Notably, the size of the LLM emerges as a decisive factor in determining the quality of responses. In the realm of LLMs, it's generally observed that larger models, characterized by a greater number of parameters, tend to produce higher-quality responses. The increased capacity of these larger models allows them to capture a broader spectrum of language nuances, leading to more nuanced, contextually aware, and coherent interactions with digital files.

Take, for example, the contrast between the Llama-2-7b model and its more massive counterpart, the Llama-2-70b model. While both are powerful in their own right, the Llama-2-70b model, with its expansive parameter count, consistently outperforms the Llama-2-7b model in terms of response quality. Users should be aware that opting for larger models often demands more substantial hardware resources. Thus, the choice of model size should align with both their quality expectations and their available computing infrastructure.

**Language, Tokens, and Datasets:**

In addition to model size, the linguistic dimension continues to play a vital role in determining the quality of responses. LLMs are inherently dependent on the languages they were trained

on. The tokenization process and the datasets used for training contribute to the model's linguistic proficiency. For instance, LLMs trained on specific languages may exhibit limitations in comprehending or generating content in languages not adequately represented in their training data. Furthermore, the choice of tokens and datasets can influence how well the model understands and responds to specific queries. Models trained on a diverse range of tokens and datasets tend to exhibit greater language versatility and may offer more accurate responses, especially in multilingual contexts.

**Future Studies**

Future studies for our project encompass two key areas of development. Firstly, the creation of multilingual Large Language Models (LLMs) tailored to a broader range of languages holds significant promise. Expanding language coverage will ensure that users from diverse linguistic backgrounds can engage with their digital files effectively. Secondly, the integration of a Graphical User Interface (GUI) is on the horizon to enhance accessibility and user-friendliness. A GUI interface will simplify the user experience, making it accessible to a broader audience and offering a more interactive and visually engaging platform for dynamic conversations with digital files. These advancements underscore our commitment to continuous improvement and the ever-evolving nature of Natural Language Processing (NLP) in reshaping human-file interaction.

REFERENCES


1. Web Access: LangChain. https://python.langchain.com/docs/get_started/introduction.
2. Web Access: LLM Leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.
3. Web Access: Embedding Models LeaderBoard. https://huggingface.co/spaces/mteb/leaderboard.
4. Web Access: FlexiGPT. https://github.com/apoalquaary/FlexiGPT.
5. Vaswani, A., Shazeer, N., Parmar, N., and at all. 2017. Attention Is All You Need. NIPS. arXiv:1706.03762v5.
6. Zheng, L., Chiang, W.L., Sheng, Y., et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv:2306.05685v2.
7. Brochier, R., Guille, A., Velcin, J. 2019. Global Vectors for Node Representations. arXiv:1902.11004v1.
8. Mikolov, T., Chen, K., Corrado, G., et al., 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781v3.
9. Hu, E., Shen, Y., Wallis, P., et all. 2021. ORA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS. arXiv:2106.09685v2.
10. Dettmers, T., Pagnoni, A., Holtzman, A., et al. 2023. QLORA: Efficient Finetuning of Quantized LLMs. arXiv:2305.14314v1.
11. Touvron, H., Martin, L., Stone, K., et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288v2.
12. Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794v1.