LLM-Based QnA for restaurants

Sarthak Doshi
MS Data Science, HDSI
sadoshi@ucsd.edu

Sarvesh Khire
MS Data Science, HDSI
skhire@ucsd.edu

Vivek Sharma
MS Data Science, HDSI
v7sharma@ucsd.edu

Rijul Sherathia
MS Data Science, HDSI
rsherathia@ucsd.edu

1 Project overview

In this project, we aim to develop an advanced recommendation system that leverages the capabilities of pretrained Large Language Models (LLMs), such as Llama 2, to offer suggestions for restaurants. The system will be designed to tailor recommendations based on individual user preferences, considering factors like cuisine, occasion, guest size, and location. By fine-tuning the LLMs and employing techniques such as Retrieval Augmented Generations (RAG), we intend to ensure that the suggestions align closely with user expectations. Additionally, the project will involve analyzing restaurant reviews and relevant data to enhance the accuracy and contextuality of the recommendations. This project not only aims to demonstrate the practical application of LLMs in creating nuanced recommender systems but also plans to evaluate and compare the performance of these models, enhancing user experience.

2 Introduction

In the rapidly evolving world of technology, personalized recommendation systems have become a cornerstone in enhancing user experiences. Recognizing this trend, our project aims to create an novel restaurant recommendation system. By harnessing the power of pretrained Large Language Models (LLMs), specifically Llama 2, this system aims to transform how individuals discover restaurants that align perfectly with their preferences.

We have made use of Llama 2, a state-of-the-art LLM known for its ability to understand and process natural language with lesser parameters as compared to its counterparts. Complementing Llama 2 is the Retrieval Augmented Generation (RAG) technique, which plays a pivotal role in enhancing the recommendation accuracy. By integrating these technologies, the system is designed to offer suggestions that are relevant to the prompt, taking into account various user-specified criteria such as cuisine type, occasion, the size of the guest group, and location preferences.

Central to our recommendation engine is the extensive dataset from Yelp, specifically focusing on restaurant reviews and associated data. This rich repository of real-world information provides a foundation for our LLMs to understand diverse culinary landscapes and consumer opinions. Analyzing this data allows our system to offer suggestions that are not only contextually relevant but also up-to-date with the latest dining trends and popular choices.

As a deliverable of this project we will be demonstrating the practical application and effectiveness of our recommendation system. This user-friendly interface will serve as a testament to the project's success in merging advanced AI with real-world applications. Furthermore, a significant goal of this project is to evaluate and compare the performance of Llama 2 and other LLMs in the context of recommendation systems.

3 Literature review

Recent studies show how recommendation systems are changing, mainly thanks to Large Language Models (LLMs). "Recommender Systems in the Era of Large Language Models (LLMs) [5]" highlights how LLMs are changing the game in recommendation systems, making it possible to give very personalized and context-aware suggestions. Overall, these studies show that LLMs, along with prompt engineering and fine-tuning, are playing a big part in improving recommendation systems, making them better for users and enhancing the quality of recommendations. BERT4Rec[1], a bidirectional model tries to overcomes the limitations of unidirectional architectures in encoding users' historical interactions for recommendation systems. It employs the Cloze task to train on masked items in a sequence.

In "A Survey on Retrieval-Augmented Text Generation," Li et al. (2022) [3] explores the integration of external data retrieval into text generation, a novel approach that enhances the relevance and accuracy of generated content. This survey contrasts various models and methods, highlighting their applicability in fields like conversational AI and automated content creation. The paper also identifies current challenges and future prospects, suggesting a trajectory for advancing retrieval mechanisms and data complexity in text generation systems.

The paper [4] presents a methodology that merges retrieval-based and generative NLP approaches. This is highly relevant for recommender systems as it opens up possibilities for leveraging extensive knowledge bases and databases in generating recommendations. The retrieval-augmented generation (RAG) model proposed in the paper enhances the ability of language models to retrieve and utilize external information. This aligns well with the needs of a recommender system that must effectively access and process information from various sources.

A conversational LLM to provide a natural language capaility to RAG model is chosen. "Llama 2: Open Foundation and Fine-Tuned Chat Models,"[7] introduces Llama 2, a series of pretrained and fine-tuned large language models (LLMs) with parameters ranging from 7 billion to 70 billion. These models, particularly Llama 2-Chat, are tailored for dialogue applications and demonstrate superior performance over existing open-source chat models in benchmarks for helpfulness and safety. The paper details the fine-tuning and safety improvement methods used, and emphasizes the contribution of these models to responsible LLM development.

4 Data Collection & Preprocessing

In the pursuit of a comprehensive understanding of the Yelp ecosystem, the integration of business and review data has proven instrumental. This strategic merger enhances our analytical capabilities by establishing nuanced connections between individual reviews and specific businesses, fostering a more holistic perspective. Given the substantial scale of the dataset, a deliberate approach was undertaken to select a limited subset of businesses and their corresponding reviews for testing purposes. This curated subset comprises essential attributes such as business name, restaurant rating, city, state, category, hours of operation, review text, and review rating, ensuring a representative and meaningful sample for analysis.

To streamline the data processing and facilitate efficient analysis, a consolidated text column was generated, incorporating pertinent details from the selected subset. This unified text format serves as the primary document for our investigation, encapsulating crucial information including business name, restaurant rating, city, state, category, hours of operation, review text, and review rating. The amalgamation of these attributes into a singular text column enables a cohesive and streamlined approach to data analysis.

In our refined dataset, we focused on businesses that garnered more than three reviews, resulting in a subset of 3,490 businesses and 63,774 corresponding reviews. This deliberate selection ensures that the dataset remains robust and representative, fostering a balanced exploration of the Yelp ecosystem. The culmination of these efforts not only provides a comprehensive view of the integrated data but also lays the foundation for meaningful insights and findings in our research endeavor.

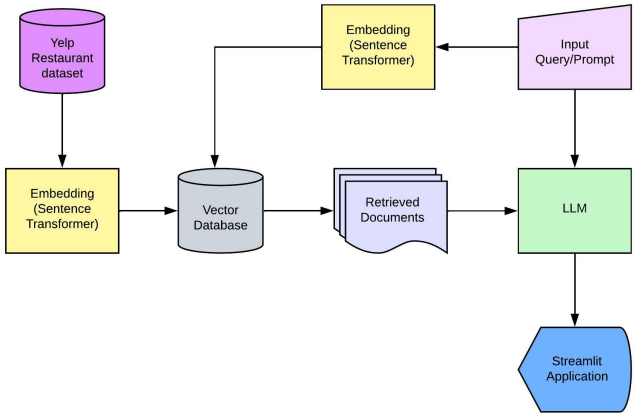

Figure 1: System Architecture

5 Proposed Methodology

In developing a restaurant recommendation system, we have outlined a multi-step methodology leveraging large language models (LLM) and natural language processing (NLP) techniques. Our approach involves the integration of business and review data from the Yelp Dataset, followed by the transformation of this information into embeddings using the AllMiniLM-l6v2 Sentence Transformer.

The dataset that we have used is, the Yelp Dataset, with 6,990,280 reviews and 150,346 businesses, is a valuable resource for research in restaurant recommendation systems, sentiment analysis, and data analysis. Structured as JSON documents, it includes Business Data (names, locations, ratings), crucial for a recommendation system, and Reviews Data revealing user experiences, with user details, review text, star ratings, and additional attributes. This dataset is frequently utilized in diverse data-driven projects due to its comprehensive insights into businesses and user interactions on the Yelp platform, making it particularly relevant for endeavors focused on restaurant recommendations and sentiment analysis.

Since the data is in a tabular format, we have first created a document for each row using the values in the columns. To capture the semantic nuances and relationships within the dataset, we employ embeddings of size 384 on the documents, created through the AllMiniLM-l6v2 Sentence Transformer. This transformation allows us to represent each document, comprising business and review details, as a dense vector in a high-dimensional space. This step is crucial in preserving the inherent meaning and context of the textual information.

In the process of obtaining embeddings, we leverage Pinecone VectorDB for efficient storage and management. Pinecone's scalable solution facilitates efficient storage and retrieval of these dense vectors, contributing to the speed and effectiveness of our recommendation system. By leveraging Pinecone's vector database capabilities, we enhance the seamless identification of businesses with comparable characteristics based on these embeddings. This optimized vector storage enables swift similarity searches, a crucial aspect in dynamically retrieving businesses aligned with user inputs through the Retrieval-Augmented Generation (RAG) Pipeline. Pinecone's advantage lies in its ability to streamline the vector search process, ensuring that our system can rapidly identify and recommend relevant businesses, thereby enriching the overall user experience.

Our experimentation involved utilizing the llm llama 2 with 7B parameters. This substantial Large Language Model (LLM) significantly contributes to the contextual understanding of user preferences and queries. By incorporating llm llama 2, our system gains the ability to provide more nuanced and accurate recommendations by comprehending user intent and context. The core of our methodology is the creation of a Retrieval-Augmented Generation (RAG) Pipeline. This innovative pipeline seamlessly integrates vector search capabilities using Pinecone and the language generation prowess of Llama2. The vector search efficiently retrieves businesses with embeddings akin to the user's

preferences, while LLM fine-tunes these recommendations by incorporating contextual understanding and user-specific nuances.

In practical terms, our user-centric system allows users to input preferences, queries, or contextual information. The RAG Pipeline dynamically utilizes the vector search to identify businesses aligned with these inputs. Subsequently, the Llama2 model refines these recommendations, ensuring the final output is not only vector-similar but also contextually relevant.

In conclusion, our proposed methodology intricately weaves together state-of-the-art NLP techniques, vector embeddings, and powerful language models to create a robust and context-aware restaurant recommendation system. By applying these steps to the Yelp Dataset, we aim to demonstrate the effectiveness of our approach in providing users with highly tailored and satisfying restaurant recommendations.

6 Experimentations and Results

In order to create a recommendation system which could respond to user queries in natural language, we started out with a simple language model, Llama2 and prompted it with recommendation queries. Though this method result resulted in

Base Model Evaluation: Initially, we employed a base version of the LLM (llama2) without any fine-tuning. The performance evaluation focused on the model's ability to provide coherent and relevant responses to a variety of queries using the Yelp dataset. However, this approach resulted in responses that lacked in specificity and relevance, indicating the necessity for a more structured and contextual approach.

Vector Database Creation and Indexing: To enhance the model's response quality, we created a vector database encompassing all columns from the Yelp dataset, merging them into a unified structure. Each record was converted into a vector representation, and a vector index was created for efficient querying. This method showed improvement in the relevance of the recommendations, though it fell short in terms of factual accuracy and specificity.

This document was further restructured into more coherent, natural language documents. This involved transforming the raw data into well-structured sentences that encapsulate key information about each restaurant (e.g., cuisine type, opening hours, location). The aim was to create a dataset that mirrors natural language queries more closely. Creating indices with this resulted in factually accurate recommendation.

Model Exploration: In addition to evaluating the baseline Llama2 model, our experimentation extended to exploring the performance of various Large Language Models (LLMs) on the Yelp dataset. This comprehensive assessment involved the following models: GPT 3.5 Baseline, GPT 3.5 with Retrieval-Augmented Generation (RAG), Llama2 13-billion parameter Baseline, Llama2 13-billion parameter with RAG, Llama2 7-billion parameter with RAG, GPT 4 Baseline, and GPT 4 with RAG.

The GPT 3.5 model, a widely recognized language model, served as the baseline for comparison. The inclusion of a Retrieval-Augmented Generation (RAG) variant aimed to investigate the impact of incorporating external data retrieval on recommendation quality. Similarly, the Llama2 models with 13-billion and 7-billion parameters represented variations in model complexity and capacity for contextual understanding.

GPT 4, a more advanced iteration of the GPT series, was introduced as another baseline, with the GPT 4 RAG model exploring the potential improvements achieved through retrieval-augmented techniques. Each model brought distinct characteristics to the experimentation, with parameter variations influencing their contextual understanding and recommendation capabilities.

The evaluation of these diverse LLMs provided valuable insights into their relative strengths and weaknesses, guiding our quest for an optimized recommendation system. Through iterative testing and refinement, we aimed to identify the most effective combination of model architecture and techniques for generating contextually relevant and accurate restaurant recommendations.

7 Analysis of Results and Methods

BLEU Score: BLEU is a metric designed to evaluate the quality of machine-generated text, particularly in the context of machine translation. It assesses the overlap of n-grams (sequences of n words) between the generated response and one or more reference responses.

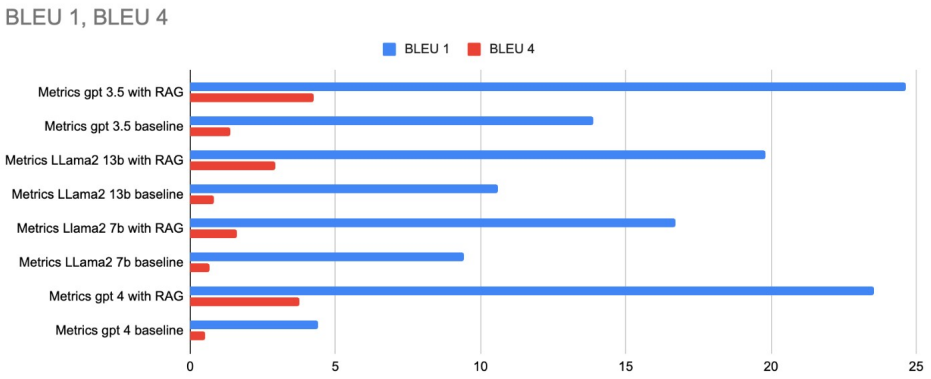

Figure 2: Bleu Score Comparison

BLEU typically considers unigrams (BLEU-1) and optionally bigrams, trigrams, and so on (BLEU-2, BLEU-3, etc.). BLEU-1 focuses on the accuracy of individual restaurant names, while higher n-grams like BLEU-2 can capture the coherence of paired restaurant output. The score ranges from 0 to 1, where 1 indicates a perfect match between the generated text and the reference text. Precision, recall, and F1 score are calculated based on the count of overlapping n-grams, ensuring that the generated restaurant suggestions align with the expected references.

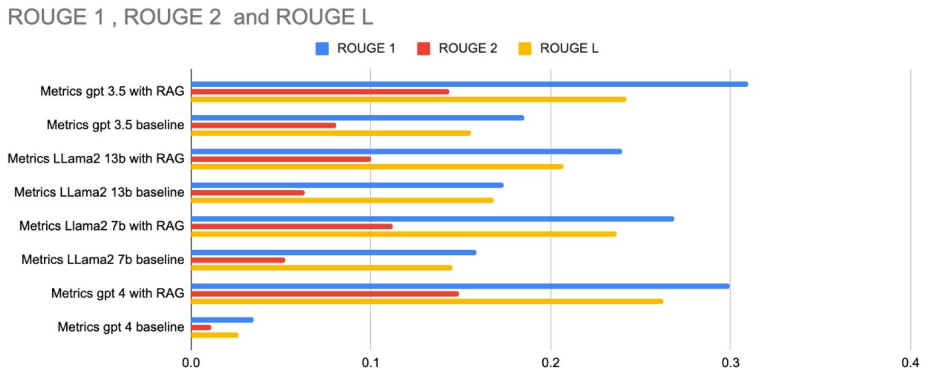

Figure 3: ROGUE Score Comparison

Similarly, ROUGE metrics, including ROUGE-1, ROUGE-2, and ROUGE-L, are employed to assess the quality of generated summaries or responses by measuring the overlap of unigrams, bigrams, and the longest common subsequence, respectively. ROUGE-1 (Unigram Overlap) measures how well individual restaurant names in the generated response align with those in the reference response. ROUGE-2 (Bigram Overlap) evaluates the coherence of paired restaurant answer by considering consecutive word sequences. ROUGE-L (Longest Common Subsequence) captures the recall of content, ensuring that the generated response includes essential information from the reference, even if not in the same order.

Figure 4 presents a comprehensive comparison of various Large Language Models (LLMs) in the context of a restaurant recommendation system, evaluating their performance using BLEU and ROUGE metrics. Notably, the GPT 3.5 with RAG stands out as the top-performing model, achieving

the highest BLEU-1 and BLEU-4 scores, indicating superior accuracy in generating restaurant names and coherent suggestions. Its robust performance extends to the ROUGE metrics, where it outshines competitors with substantial margins, reflecting its proficiency in capturing unigram and bigram overlaps, as well as the longest common subsequence.

Model	BLEU 1	BLEU 4	ROUGE 1	ROUGE 2	ROUGE L
GPT 3.5 RAG	**24.65863833**	**4.24979153**	**0.3094765853**	0.1438718299	0.242068924
GPT 3.5 Baseline	13.86434735	1.386262233	0.1853730246	0.08104452264	0.1559532137
LLama 2 13b RAG	19.80933278	2.943654645	0.2396679444	0.09989794115	0.2070630066
LLama 2 13b Baseline	10.59146373	0.8351355419	0.1735664171	0.06340614639	0.1684596101
Llama 2 7b RAG	16.72312978	1.629905527	0.2686351222	0.1123995418	0.2368219017
LLama 2 7b Baseline	9.445291584	0.682678094	0.1584050643	0.0522108409	0.1451827432
GPT 4 RAG	23.54007133	3.749093371	0.2992184467	**0.149151744**	**0.2627706251**
GPT 4 Baseline	4.400076871	0.519422978	0.03478809336	0.01158398347	0.02622704803

Figure 4: LLM Metric Score Comparison

Contrastingly, the baseline GPT 4 model exhibits suboptimal performance, particularly evident in its significantly lower BLEU and ROUGE scores. This suggests that the introduction of the RAG technique plays a crucial role in enhancing recommendation quality, as seen in the notable improvement observed in GPT 4 with RAG.

In the Llama2 series, the 13-billion parameter model with RAG demonstrates commendable performance, surpassing its baseline counterpart across all metrics. The significance of RAG is further emphasized by the consistent improvements observed in Llama2 7-billion parameter models with RAG compared to their baseline versions.

The findings underscore the pivotal role of fine-tuning techniques, such as RAG, in refining LLMs for recommendation systems. As the GPT 3.5 with RAG emerges as the frontrunner, this research advocates for the adoption of advanced LLMs with retrieval-augmented generation for optimal performance in personalized restaurant recommendations.

8 Conclusion

In conclusion, our paper presents a comprehensive exploration of an advanced restaurant recommendation system, leveraging Large Language Models (LLMs) and innovative natural language processing (NLP) techniques. The initial experimentation with the Llama2 language model revealed the need for refinement in specificity and relevance. To address this, we employed a multi-faceted approach involving Vector Database Creation, Indexing, and Document Restructuring. These steps, combined with the integration of a comprehensive Yelp dataset, paved the way for more accurate and contextually relevant restaurant recommendations.

Furthermore, our study extended its focus to assess the performance of various LLMs, including GPT 3.5, Llama2 with different parameter sizes, and GPT 4. The inclusion of Retrieval-Augmented Generation (RAG) techniques in certain models underscored the significance of external data retrieval in enhancing recommendation quality. Notably, GPT 3.5 with RAG emerged as a frontrunner, exhibiting superior performance across BLEU and ROUGE metrics, showcasing the potential of combining advanced language models with retrieval-augmented approaches.

The experimentation and evaluation process highlighted the iterative nature of system refinement, emphasizing the importance of context-aware document embeddings and efficient vector storage facilitated by Pinecone VectorDB. The proposed methodology intricately weaves together state-of-the-art NLP techniques, vector embeddings, and powerful LLMs, providing a robust foundation for a user-centric and effective restaurant recommendation system.

References

[1] Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., and Jiang, P. (2019). BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. Proceedings of the 28th ACM International Conference on Information and Knowledge Management.

[2] Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T., Wang, G. (2023). Text Classification via Large Language Models. ArXiv, abs/2305.08377. https://doi.org/10.1007/978-981-19-5221-063

[3] Li, Huayang, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. "A survey on retrieval-augmented text generation." arXiv preprint arXiv:2202.01110 (2022).

[4] Lewis, Patrick, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel and Douwe Kiela. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” ArXiv abs/2005.11401 (2020): n. pag.

[5] Fan, W., Zhao, Z., Li, J., Liu, Y., Mei, X., Wang, Y., Tang, J., Li, Q. (2023). Recommender Systems in the Era of Large Language Models (LLMs). ArXiv, abs/2307.02046.

[6] R. M. Gomathi, P. Ajitha, G. H. S. Krishna and I. H. Pranay, "Restaurant Recommendation System for User Preference and Services Based on Rating and Amenities," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), Chennai, India, 2019, pp. 1-6, doi: 10.1109/ICCIDS.2019.8862048.

[7] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, Thomas Scialom 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models