JeopardyLLM: Adventures in Retrieval-Augmented Generation

Job Bangayan, Jerry Gonzalez, Sean Huang, Vineeta Kumar, Tianyu Wang

Abstract

This project explores retrieval-augmented generation technologies by using questions found on the popular game show Jeopardy. Our data includes 470,000 Jeopardy clues from 39 seasons. Using this, we deployed a framework by which old clues are embedded and stored in a vector database, then contexts are created from novel questions and along with a prompt, fed to a FLAN-T5 large language model. This was successfully deployed as JeopardyLLM, an online interactive website where users can engage with and discover answers to questions based solely on knowledge gained from old Jeopardy clues. When measured on newer championship-level Jeopardy questions, accuracy was found to be 31.7%

1. Introduction

The trivia hobbyist is one that enjoys both the pursuit of knowledge and the spirit of competition. Within the trivia world, no question-and-answer format is more pervasive and enduring than Jeopardy's. Since it aired episodes in 1984, the show has curated a devoted following with many considering it daily appointment viewing. With over 470,000 clues over close to 30 years, the show has amassed a large database of questions that theoretically do not get repeated (unless there is an unforeseen event such as television and film writer's strike of 2023). An episode of Jeopardy consists of general knowledge questions in the form of 'answer' clues while contestants must identify the person, place, object, or idea/concept that the clue describes. Categories are often on the sillier side and the questions are formatted very peculiarly (the question is an answer, and the answer is a question).

Likewise, the clues are written in a variety of different formats, often placing emphasis on context clues and associations between key phrases. Understanding these contexts is an essential skill for a Jeopardy contestant. Another is having experience with previous questions. By watching enough episodes, contestants can start to develop a cadence and awareness of what question topics are frequently asked. Some studies explore trends while analyzing the questions themselves. (Hamilton 2020, Salo 2019).

However, practice and studying are still essential elements for winning. Answering Jeopardy-style questions using algorithms became especially popular when IBM Watson defeated co-game show legends Ken Jennings and Brad Rutter in 2013 via a simple stochastic model (Tesauro, 2013). Some newer language transformer models, such as BARD and ChatGPT, perform this task with the accuracy of a human expert. Moreover, they demonstrate higher accuracy than Watson in doing so. (O'Leary, 2023) Yet many of these are trained on general knowledge sources found online. A novel attempt would be to answer questions from the most recent season using information and context gleaned from solely previous Jeopardy questions.

2. Our Mission

Our project JeopardyLLM, aims to develop a sophisticated model that leverages advanced natural language processing (NLP) techniques to answer questions based on a vast dataset of over 470,000 clues spanning nearly 30 seasons of the renowned TV show, Jeopardy! By harnessing the power of contextual embeddings and retrieval tasks, the JeopardyLLM model seeks to offer an unparalleled experience for trivia enthusiasts and knowledge seekers alike.

Information retrieval lies at the heart of this project, underscoring its importance within the realm of natural language processing. Information retrieval is the process of accessing and retrieving relevant information from a large collection of data or documents. It involves techniques and algorithms designed to efficiently search for and retrieve specific information based on user queries or search terms. As a core principle in NLP, information retrieval plays a pivotal role in enabling question-answering systems to effectively respond to user queries.

In our solution, **JeopardyLLM**, we present a website powered by state-of-the-art NLP technology, ensuring precise and accurate responses to a wide array of trivia questions. While JeopardyLLM is category agnostic, capable of handling questions from various topics and categories found in Jeopardy, it is expected to excel in certain categories over others. Our primary focus is to deliver a seamless and enjoyable experience for users, achieved through a simple and intuitive interface for submitting questions. Moreover, we prioritize real-time responsiveness, ensuring users receive prompt answers and enabling a dynamic engagement with the platform. With JeopardyLLM, we aim to create a space where trivia enthusiasts can freely explore questions and enhance their knowledge.

2.1 Market Opportunity

The market opportunity for a language model trained on Jeopardy questions is substantial, given the immense popularity of the game show and the widespread interest in trivia and similar games. With Jeopardy averaging 9.2 million total viewers during the 2021-2022 season, there exists a vast audience of trivia enthusiasts, prospective game show contestants, and curious individuals worldwide. This project presents a unique opportunity to cater to these segments by providing a platform where users can test their knowledge and explore trivia questions in a Jeopardy-style format. Assumptions underlying this opportunity include the belief that users, while not necessarily expecting 100% accuracy, are inherently curious and seek to expand their knowledge beyond traditional trivia boundaries. Furthermore, it is assumed that users will structure questions in a way that is answerable, facilitating meaningful interaction with the platform.

2.2 Competitive Landscape

In the landscape of language models and question-answering solutions, major players and vendors include ChatGPT and other language model frameworks that can be fine-tuned to answer questions. These models, often accessible through platforms like Hugging Face, leverage advanced natural language processing techniques to understand and respond to user queries. Additionally, cloud-based AI services offered by industry giants such as Google Cloud AI and AWS play a significant role in providing infrastructure and tools for developing and deploying such models. Rasa, a conversational AI platform, also merits consideration as a player in this space, offering solutions for building chatbots and virtual assistants.

Figure 2.1 Major Players In The Same Space

Company Name	Product / Solution Overview	Who is the primary customer?	Key differentiation vs our solution (based on our understanding)
ChatGPT	Answer questions via LLM	public	Our solution is an information retrieval approach based solely on information found in Jeopardy questions.
Google Cloud AI	NLP APIs and tools that provide human-like responses to questions.	public	
AWS	NLP services like Amazon Comprehend and Kendra can be used to build intelligent question-answering systems	public	
Rasa	Conversational AI platform that allows developers to build own chatbots and question-answer systems	Top enterprises such as Starbucks and Adobe	

3. Methods

3.1 Data

The open source dataset contains Jeopardy clues from Season 1 through Season 39 and has ~ 470,000 clues. The data was pulled from the GitHub repository https://github.com/jwolle1/jeopardy_clue_dataset/tree/master. These clues are distributed to over 52,000 categories, about 50MB in Excel Format and around 130MB in JSON. The average number of clues for the first 11 seasons was around 8800 and about 13000 from seasons 12-38. The dataset was already cleaned and filtered out on clues that depend on images, video, or audio. There was still minimal cleanup that was done on our data, where some questions had some notes from the crew. The non-alphanumeric characters were also removed from the dataset to help reduce the noise in the text data.

The dataset had several fields that were available to us which includes round number, clue value, daily double value, category, comments, questions, answers, air date and notes. Since the focus of this project is doing a Large Language Model (LLM), our exploratory data analysis was focused on the categories, questions, and answers.

The exploratory data analysis involved examining the distribution and characteristics of our Jeopardy clues. We looked into the frequency counts and some examples were the number of times each category was used and the most common questions in which we found out that places and locations were the most common ones. We used bar plots to visualize the frequencies and it has helped us to identify patterns in our data. We also did tokenization of our data which helped us analyze which words or tokens are the most common. After tokenization, we also did a bigram analysis of our data to find out which pairing of words is the most common as this can help with the patterns and association within the text.

Figure 3.1 Top 20 most common categories

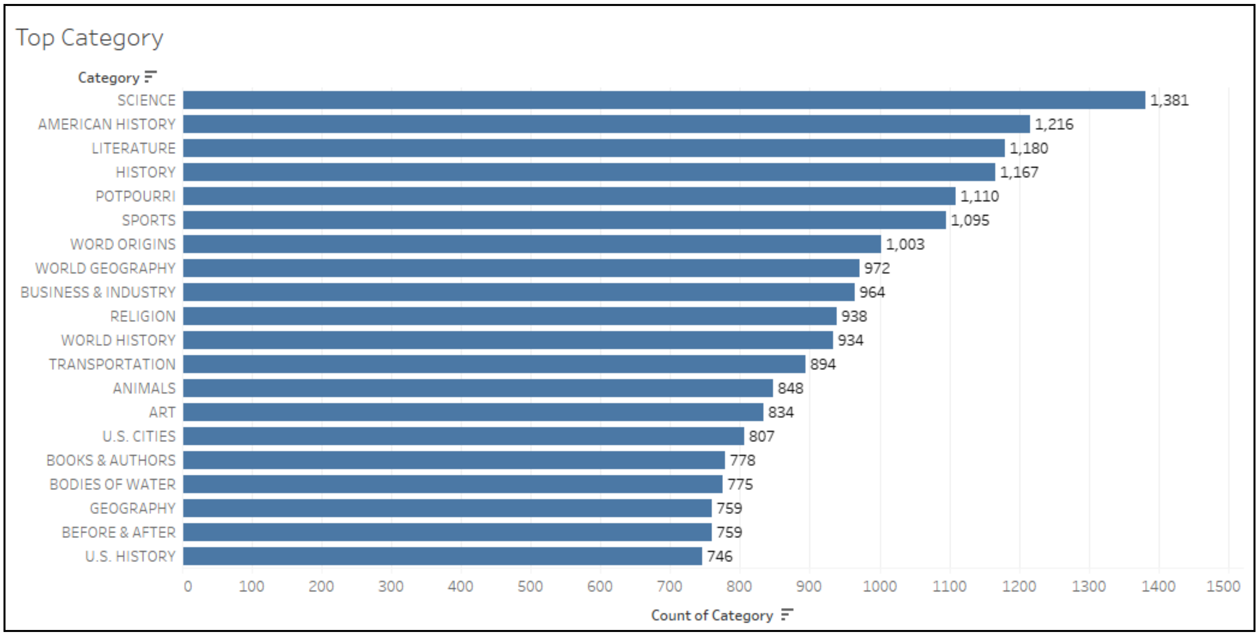


Figure 3.2 Top 20 most common questions

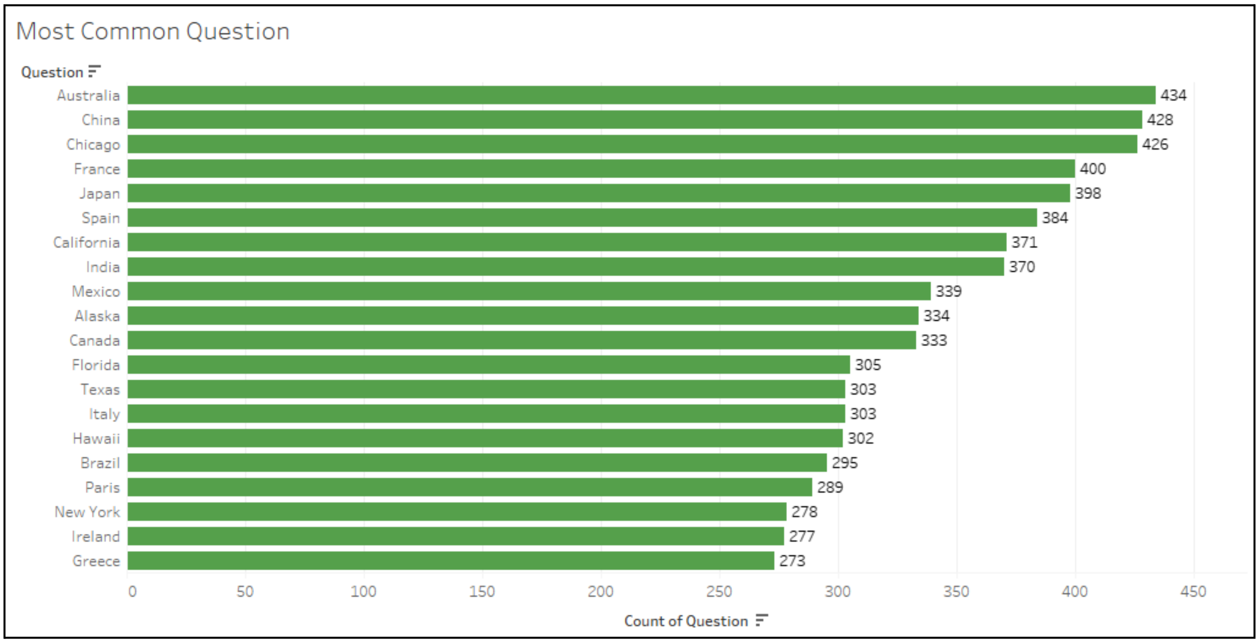


Figure 3.3 Word Cloud of answers when tokenized

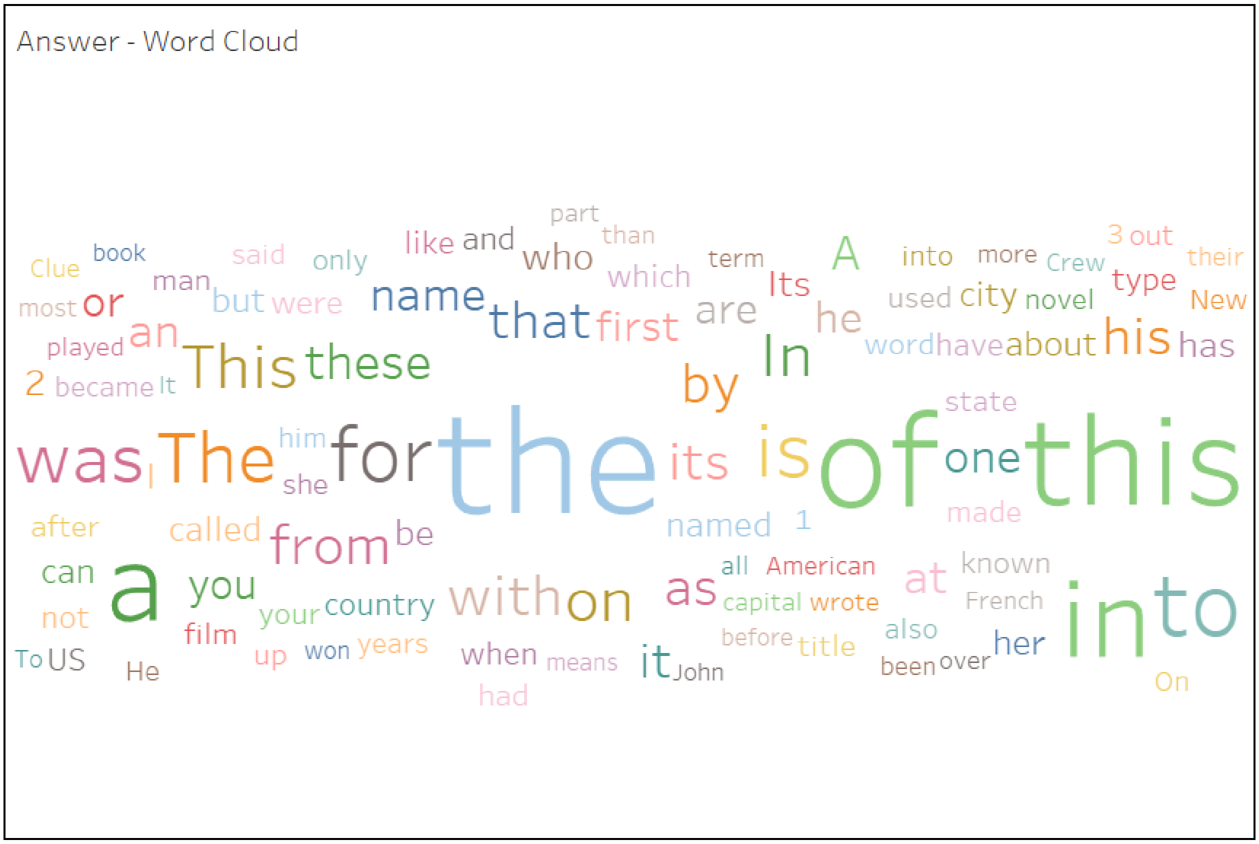


Figure3.4 Top 20 most common bigrams

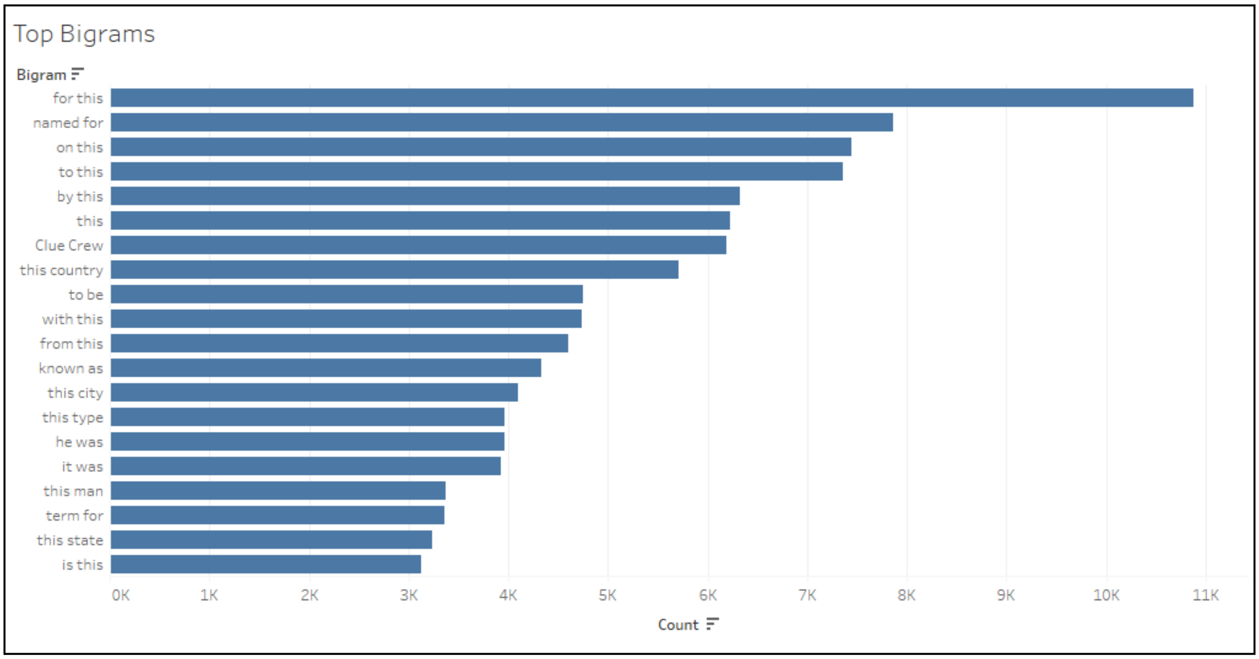
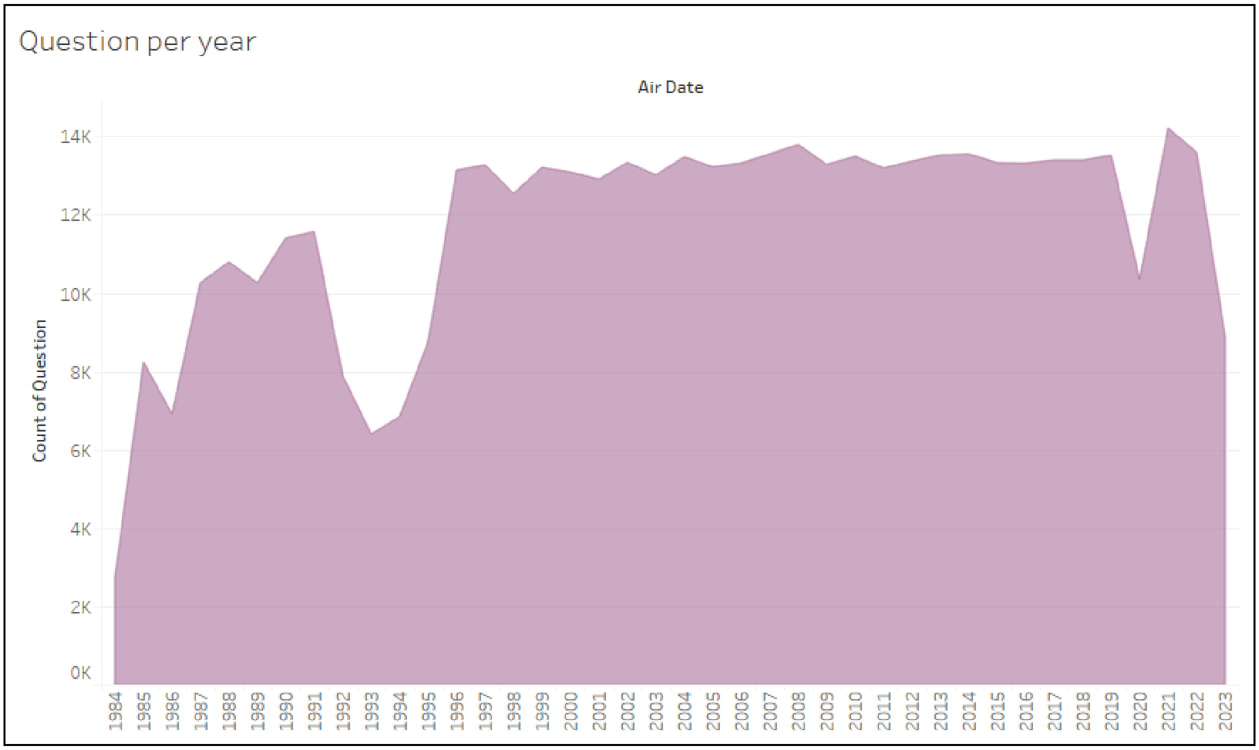


Figure 3.5 Number of questions per year



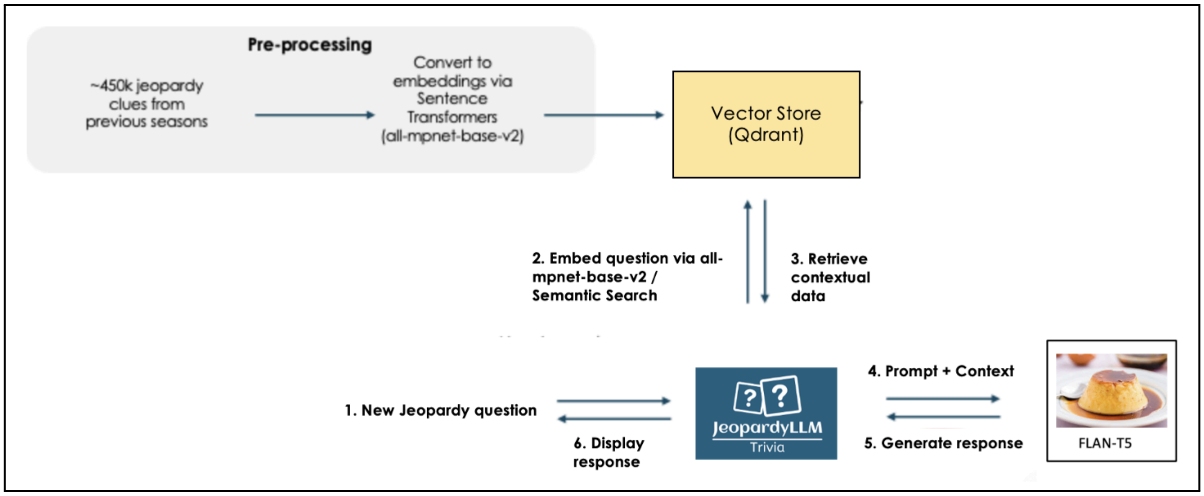
3.2 Model

For ease of understanding, questions and answers are defined as they usually are (rather than the answer -> question style that Jeopardy is known for). All the Jeopardy clues from the first 39 seasons were converted into sentence form with the following format:

'Regarding [category], [question] is [answer]'

Afterwards, each question is converted into embeddings using all-mpnet-base-v2, part of the Sentence Transformers framework (sbert.net). Sentence Transformers is a Python-based collection of pre-trained models. The model of all-mpnet-base-v2 was chosen based on embedding time, accuracy and ease of use. Comparisons of various models can be seen below in the Discussion section. The embeddings are stored in a vector store on Qdrant.

After pre-processing, the model can be conducted as seen in the following Figure:


Pre-processing
~450k jeopardy clues from previous seasons
Convert to embeddings via Sentence Transformers (all-mpnet-base-v2)
Vector Store (Qdrant)
2. Embed question via all-mpnet-base-v2 / Semantic Search
3. Retrieve contextual data
1. New Jeopardy question
6. Display response
JeopardyLLM Trivia
4. Prompt + Context
5. Generate response
FLAN-T5

New answers become embedded with the same all-mpnet-base-v2 model from Sentence Transformers. The ten most similar Jeopardy clues are retrieved from the vector store and combined to form a context. The context is combined with a prompt being 'Use the following pieces of context to answer the question' and the question itself, then fed to an LLM. For the purposes of our model, we use Flan-T5, a powerful text-to-text language model and enhanced version of T5 introduced by Google which has shown proficiency in question answering (Chung 2022). Attempts have been made to fine-tune Flan-T5 on novel sets of data, including SQuAD and Quora questions, although with
variable amounts of success (Toughdata 2023).

The model is tested on Jeopardy 2024 Tournament of Champions Quarterfinals Questions (n = 536). The results from this task will be compared to the correct answers, looking for an exact match in the Jeopardy vein. Similar to the game show, spelling does not matter unless a syllableis added. In addition, last names are sufficient for proper names. Any grammatically correct question with the correct answer within it counts as well. In addition, FLAN-T5 without context achieved only 2.1%, allowing for room for improvement using context. Accuracy with the addition of context improved the accuracy to 31.7%.

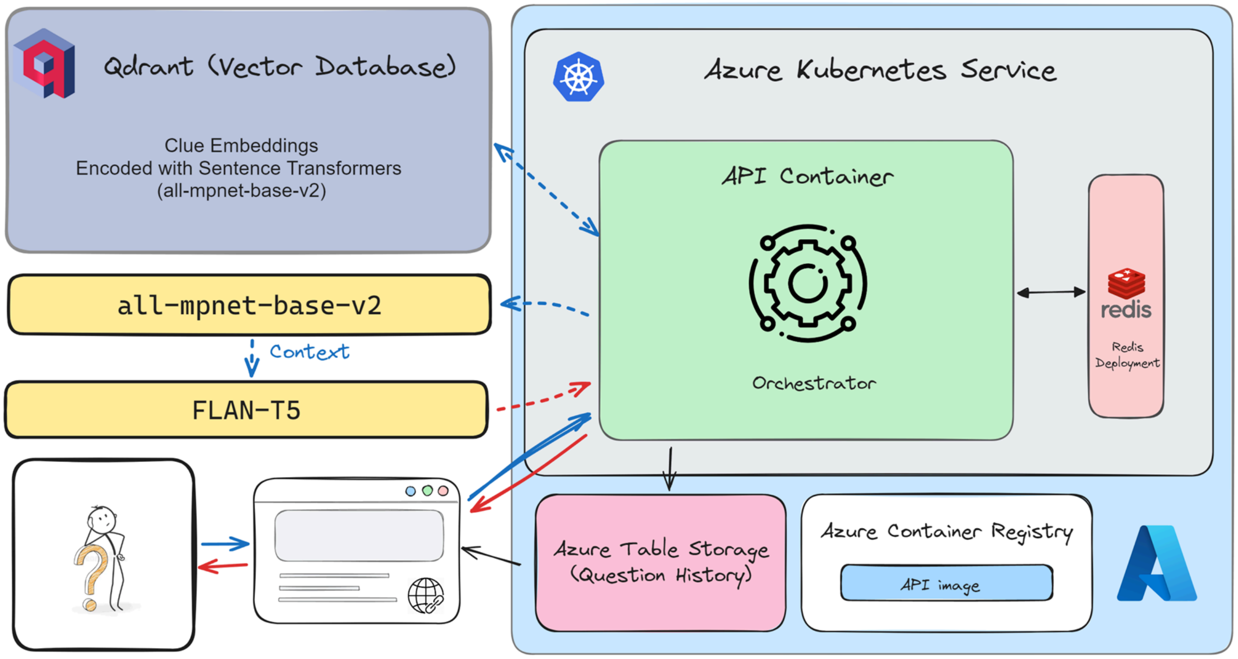
3.3 Architecture

Our backend architecture is a sophisticated ensemble of state-of-the-art vector database technology and advanced machine learning models. At its core, it integrates Qdrant as the vector database to handle and optimize the search of text embeddings, which are crucial for the responsiveness and accuracy of our web application.

We utilize two powerful models from Huggingface, renowned for their efficiency in natural language processing tasks. The 'all-mpnet-base-v2' is our chosen embedding model that processes input vectors with finesse. Post the embedding generation, the FLAN-T5 model steps in as the Large Language Model to predict responses, capitalizing on its vast knowledge base and context comprehension capabilities.

The entire application is reliably hosted on Azure, which offers robust management through its Kubernetes Service and ensures that services like the API container and Redis cache are optimally deployed. The application's image repository is secured in the Azure Container Registry, while Azure Table Storage diligently archives every user interaction. Although the system is designed to cater to 3 concurrent requests within the constraints of a 3GB container and 4GB RAM, it boasts an impressive self-recovery feature that restores functionality approximately 1 minute after a crash.

Figure 3.3.1 System Architecture

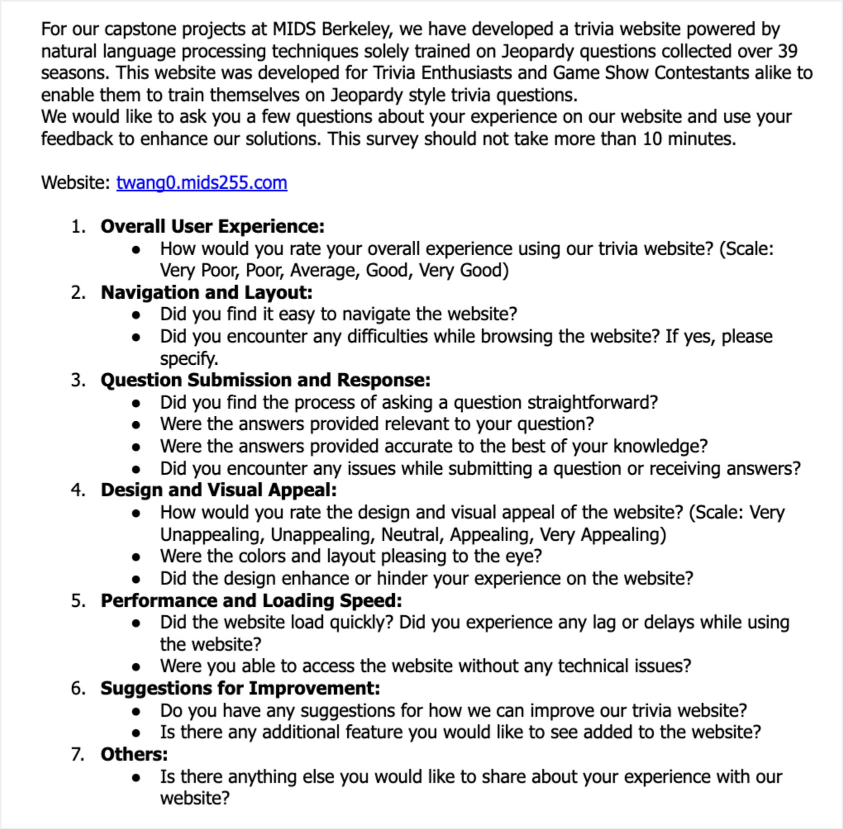


4. Results

4.1 User Testing

To assess the usability of our solution, we designed a comprehensive user survey to cover the essential aspects of user experience, navigation, question submission, design, and performance, as well as gather suggestions for improvement. It was structured in a way that allowed users to provide specific feedback on each aspect of their experience with the trivia website. This helped us identify areas of strength and areas that needed improvement, enabling us to enhance the website's functionality and usability. The completed questionnaire is referenced below

Figure 4.1 User Testing Questionnaire



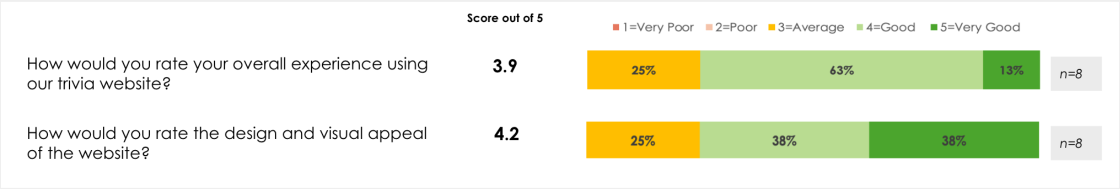
For our capstone projects at MIDS Berkeley, we have developed a trivia website powered by natural language processing techniques solely trained on Jeopardy questions collected over 39 seasons. This website was developed for Trivia Enthusiasts and Game Show Contestants alike to enable them to train themselves on Jeopardy style trivia questions.
We would like to ask you a few questions about your experience on our website and use your feedback to enhance our solutions. This survey should not take more than 10 minutes.

Website: twang0.mids255.com

1. **Overall User Experience:**
   - How would you rate your overall experience using our trivia website? (Scale: Very Poor, Poor, Average, Good, Very Good)
2. **Navigation and Layout:**
   - Did you find it easy to navigate the website?
   - Did you encounter any difficulties while browsing the website? If yes, please specify.
3. **Question Submission and Response:**
   - Did you find the process of asking a question straightforward?
   - Were the answers provided relevant to your question?
   - Were the answers provided accurate to the best of your knowledge?
   - Did you encounter any issues while submitting a question or receiving answers?
4. **Design and Visual Appeal:**
   - How would you rate the design and visual appeal of the website? (Scale: Very Unappealing, Unappealing, Neutral, Appealing, Very Appealing)
   - Were the colors and layout pleasing to the eye?
   - Did the design enhance or hinder your experience on the website?
5. **Performance and Loading Speed:**
   - Did the website load quickly? Did you experience any lag or delays while using the website?
   - Were you able to access the website without any technical issues?
6. **Suggestions for Improvement:**
   - Do you have any suggestions for how we can improve our trivia website?
   - Is there any additional feature you would like to see added to the website?
7. **Others:**
   - Is there anything else you would like to share about your experience with our website?

We conducted user testing among our social circles and received a full response of 11 samples. The results of the user testing were promising. The overall experience rating for the trivia website was 3.9 out of 5, while the design and visual appeal received a score of 4.2. Users appreciated the easy navigation, clean website design, and usability features such as helpful guides and instructions. However, some concerns were raised about repeated recent questions and the lack of indication for wait times, both of which were resolved. Additionally, users requested sample questions, which were also addressed. Issues related to response grammatical correctness and mixed accuracy were considered out of scope due to model and skill limitations. Overall, we were satisfied with the value we were able to drive by conducting user testing.

Figure 4.2 Results



4.2 Website

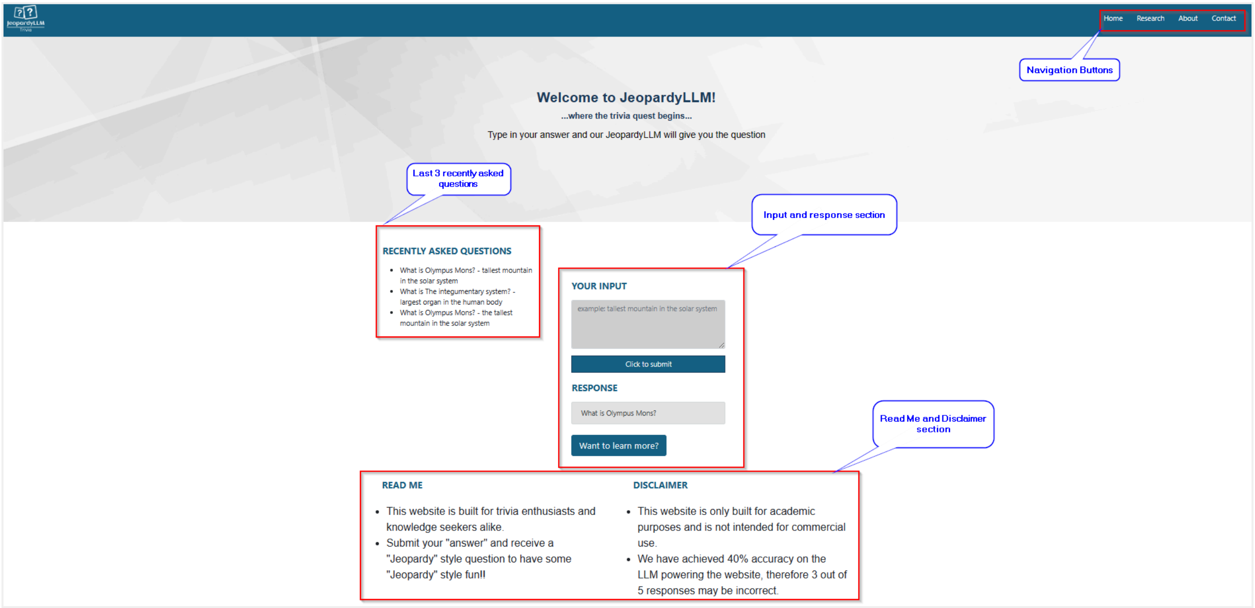
The website features a streamlined design, utilizing the Bootstrap framework, which is a widely recognized, free, and open-source Cascading Style Sheet (CSS) framework. Bootstrap integrates essential web development elements such as HyperText Markup Language (HTML), CSS, and JavaScript, which have been instrumental in our design. We selected Bootstrap to streamline the development process, taking advantage of its diverse range of options for colors, sizes, fonts, and layouts. This framework offers pre-defined style guidelines for all HTML components, ensuring a cohesive visual aesthetic. The website serves as the gateway to the trivia quest, inviting users to begin their interactive experience.

4.2.1 Landing page

The landing page contains the following sections:

- Input and response section - in this section, the user submits an answer as shown in the example and receives a "Jeopardy" style response question. Note that if the user wants to learn more about the response there is a "Want to learn more?" button that opens another tab to a Google search on the topic. See figure 4.2.1b.
- Recently Asked Questions section - here, the user can see the last 3 recently asked questions.
- Read Me and Disclaimer section - this section provides a brief explanation of how to use the website and a disclaimer noting that this website is only for academic purposes and has a 40% accuracy rating.
- There are 4 navigation buttons where the user can explore the research paper content, learn about the team, or contact the team.

Figure 4.2.1a: Website Landing Page



Welcome to JeopardyLLM!

...where the trivia quest begins...

Type in your answer and our JeopardyLLM will give you the question

RECENTLY ASKED QUESTIONS

- What is Olympus Mons? - tallest mountain in the solar system
- What is The integumentary system? - largest organ in the human body
- What is Olympus Mons? - the tallest mountain in the solar system

YOUR INPUT

example: tallest mountain in the solar system

RESPONSE

What is Olympus Mons?

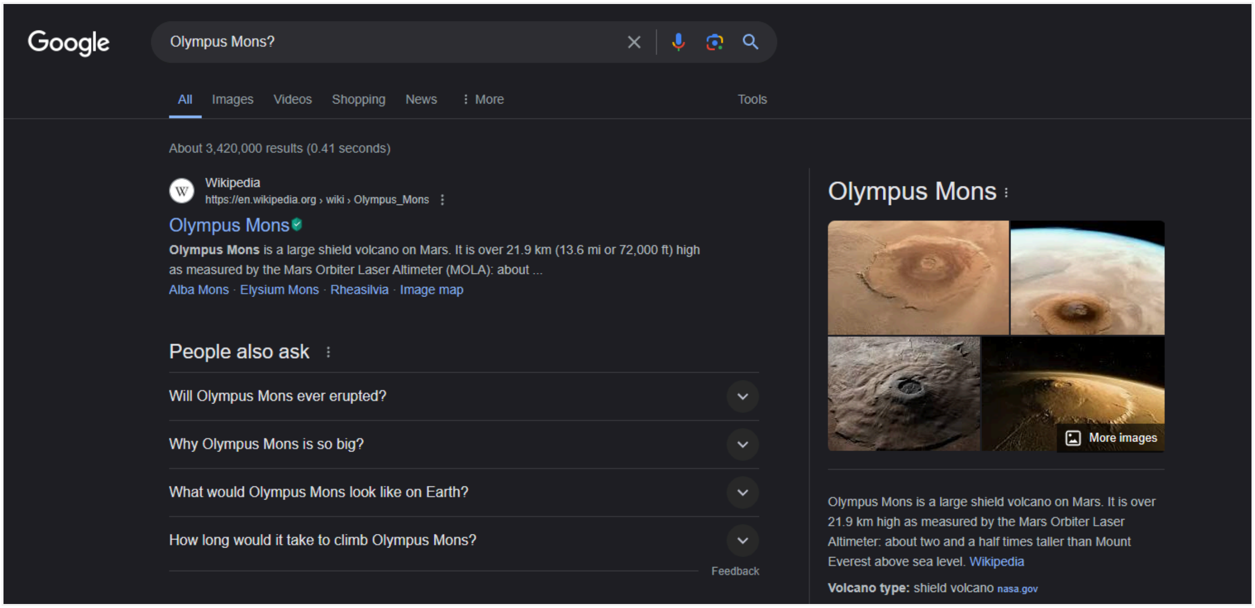
READ ME

- This website is built for trivia enthusiasts and knowledge seekers alike.
- Submit your "answer" and receive a "Jeopardy" style question to have some "Jeopardy" style fun!!

DISCLAIMER

- This website is only built for academic purposes and is not intended for commercial use.
- We have achieved 40% accuracy on the LLM powering the website, therefore 3 out of 5 responses may be incorrect.

Figure 4.2.1b Google search of topic



About 3,420,000 results (0.41 seconds)

Wikipedia
https://en.wikipedia.org › wiki › Olympus_Mons

Olympus Mons

Olympus Mons is a large shield volcano on Mars. It is over 21.9 km (13.6 mi or 72,000 ft) high as measured by the Mars Orbiter Laser Altimeter (MOLA): about ...

Alba Mons · Elysium Mons · Rheasilvia · Image map

People also ask

- Will Olympus Mons ever erupted?
- Why Olympus Mons is so big?
- What would Olympus Mons look like on Earth?
- How long would it take to climb Olympus Mons?

Olympus Mons

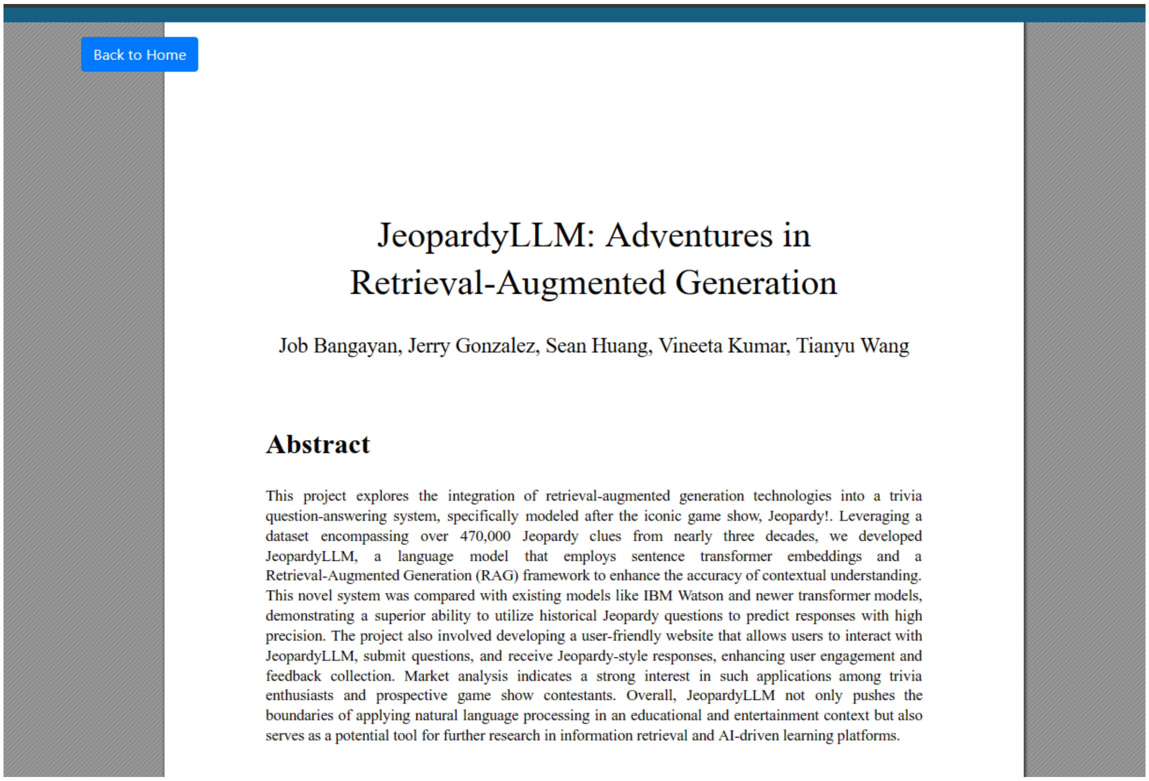
Olympus Mons is a large shield volcano on Mars. It is over 21.9 km high as measured by the Mars Orbiter Laser Altimeter: about two and a half times taller than Mount Everest above sea level. Wikipedia

Volcano type: shield volcano nasa.gov

4.2.2 Research section

Figure 4.2.3 illustrates what the user would see when hitting the research tab on the landing page. The paper details our mission, the methods used, and the results obtained when using a Jeopardy style question-answer dataset to train an LLM using a RAG pipeline.

Figure 4.2.2: Jeopardy LLM project paper



JeopardyLLM: Adventures in Retrieval-Augmented Generation

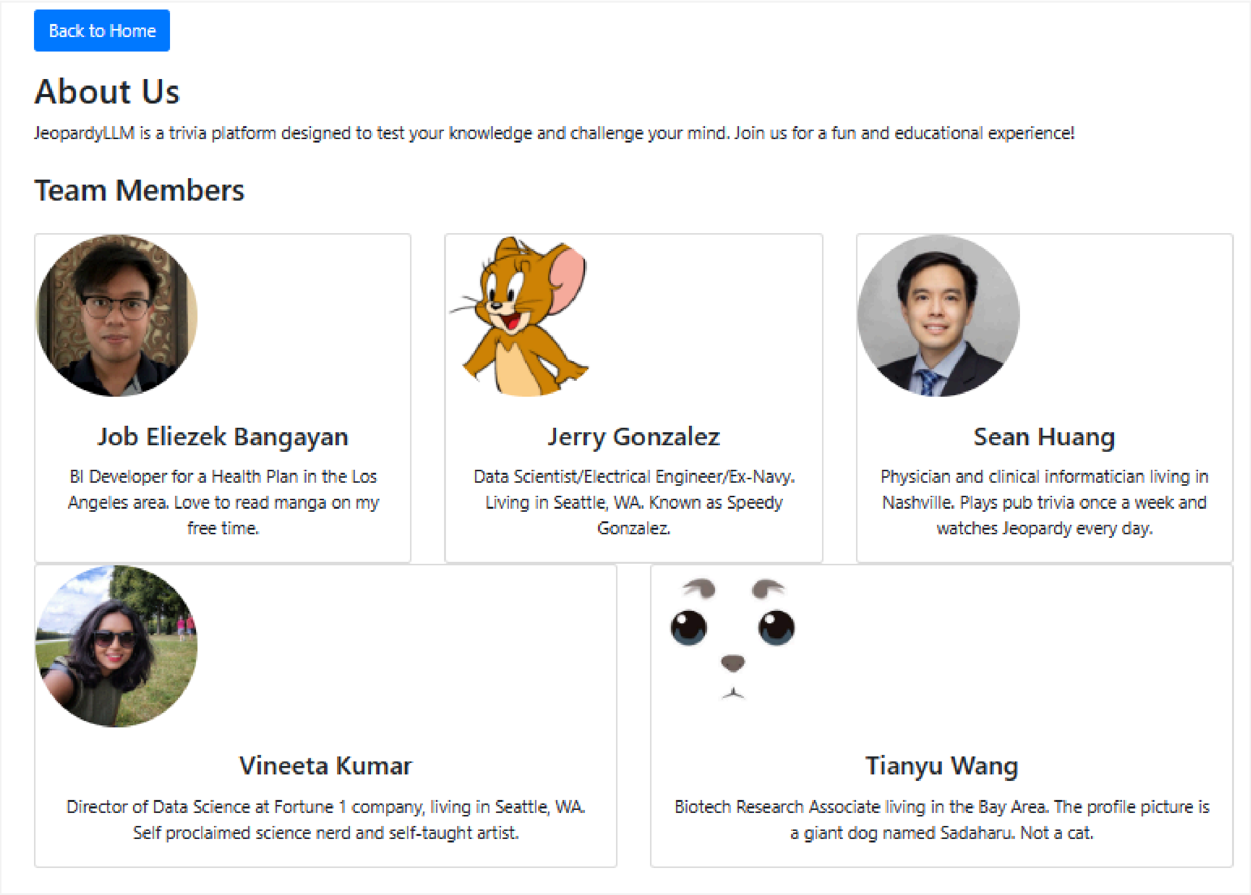
Job Bangayan, Jerry Gonzalez, Sean Huang, Vineeta Kumar, Tianyu Wang

Abstract

This project explores the integration of retrieval-augmented generation technologies into a trivia question-answering system, specifically modeled after the iconic game show, Jeopardy!. Leveraging a dataset encompassing over 470,000 Jeopardy clues from nearly three decades, we developed JeopardyLLM, a language model that employs sentence transformer embeddings and a Retrieval-Augmented Generation (RAG) framework to enhance the accuracy of contextual understanding. This novel system was compared with existing models like IBM Watson and newer transformer models, demonstrating a superior ability to utilize historical Jeopardy questions to predict responses with high precision. The project also involved developing a user-friendly website that allows users to interact with JeopardyLLM, submit questions, and receive Jeopardy-style responses, enhancing user engagement and feedback collection. Market analysis indicates a strong interest in such applications among trivia enthusiasts and prospective game show contestants. Overall, JeopardyLLM not only pushes the boundaries of applying natural language processing in an educational and entertainment context but also serves as a potential tool for further research in information retrieval and AI-driven learning platforms.

4.2.3 About section

Figure 4.2.3 has a brief description of the team members.



About Us

JeopardyLLM is a trivia platform designed to test your knowledge and challenge your mind. Join us for a fun and educational experience!

Team Members

Job Eliezek Bangayan

BI Developer for a Health Plan in the Los Angeles area. Love to read manga on my free time.

Jerry Gonzalez

Data Scientist/Electrical Engineer/Ex-Navy. Living in Seattle, WA. Known as Speedy Gonzalez.

Sean Huang

Physician and clinical informatician living in Nashville. Plays pub trivia once a week and watches Jeopardy every day.

Vineeta Kumar

Director of Data Science at Fortune 1 company, living in Seattle, WA. Self proclaimed science nerd and self-taught artist.

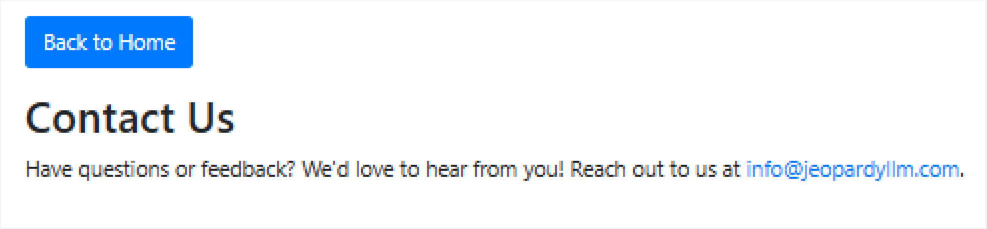
Tianyu Wang

Biotech Research Associate living in the Bay Area. The profile picture is a giant dog named Sadaharu. Not a cat.

Figure 4.2.3 Team Members

4.2.4 Contact Us section

Figure 4.2.4 instructs the user how to contact the team for questions and feedback.



Contact Us

Have questions or feedback? We'd love to hear from you! Reach out to us at info@jeopardyllm.com.

Figure 4.2.4 How to contact the team

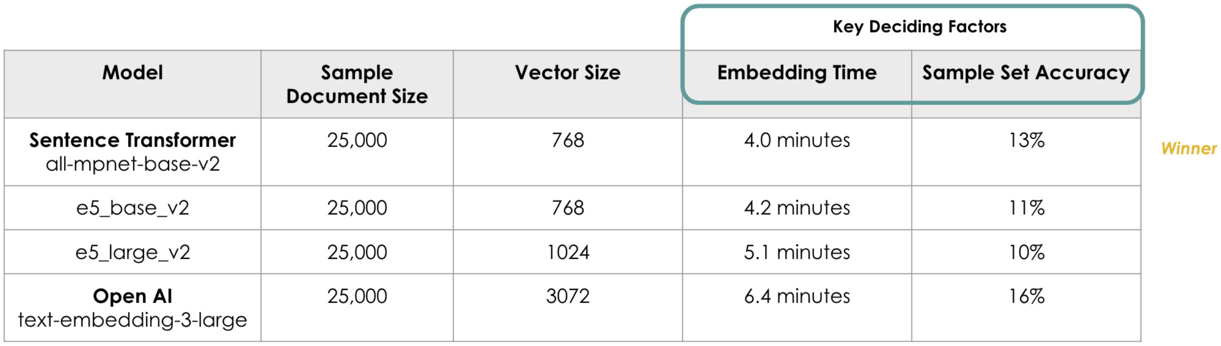
5. Discussion

5.1 Embedding Model Research

From the plethora of cutting-edge embedding models available, we needed to select the most suitable embedding model for the project. To keep the process streamlined and accommodate the large dataset required for training, we narrowed down our criteria to two key metrics: **embedding time** and **question-and-answer accuracy**.

We conducted experiments using a sample document containing 25,000 questions and evaluated four popular embedding models as shown in the table below. Upon analysis, we observed that the OpenAI model, `text-embedding-3-large`, demonstrated a 23% higher accuracy compared to the Sentence Transformer model all-mpnet-base-v2. However, it also exhibited a 60% increase in embedding time due to the larger vector size of 3072 for `text-embedding-3-large` vs 768 for `all-mpnet-base-v2`. Considering both metrics and acknowledging that the OpenAI model is a paid service, we ultimately opted to proceed with the open-source alternative from Sentence Transformer i.e. `all-mpnet-base-v2`.

Figure 5.1 Embedding Model Performance Comparison

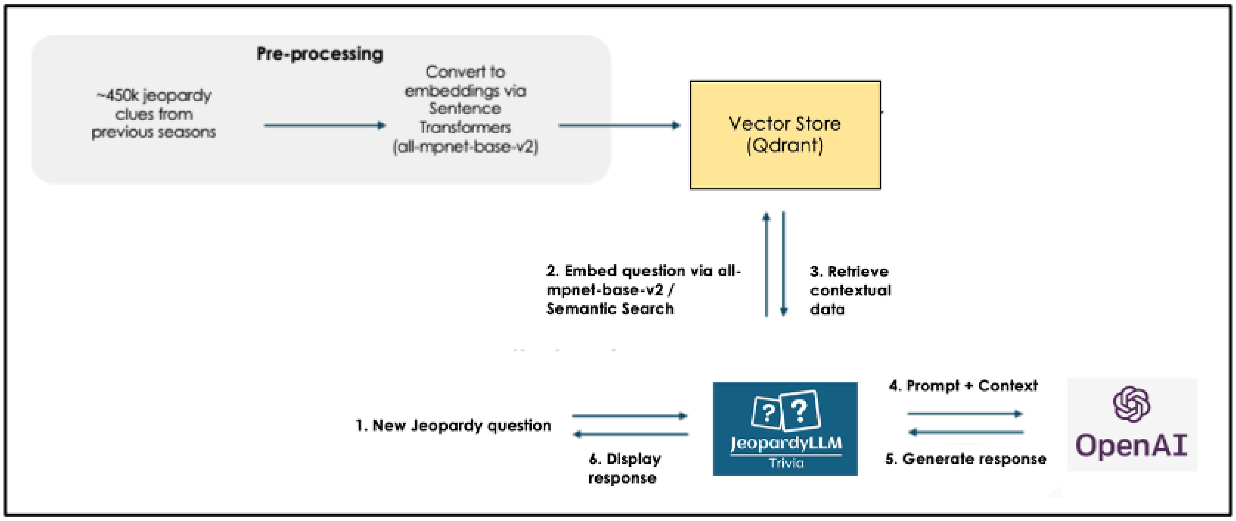


			Key Deciding Factors		
Model	Sample Document Size	Vector Size	Embedding Time	Sample Set Accuracy	
Sentence Transformer all-mpnet-base-v2	25,000	768	4.0 minutes	13%	*Winner*
e5_base_v2	25,000	768	4.2 minutes	11%	
e5_large_v2	25,000	1024	5.1 minutes	10%	
Open AI text-embedding-3-large	25,000	3072	6.4 minutes	16%	

5.2 RAG Pipeline Research

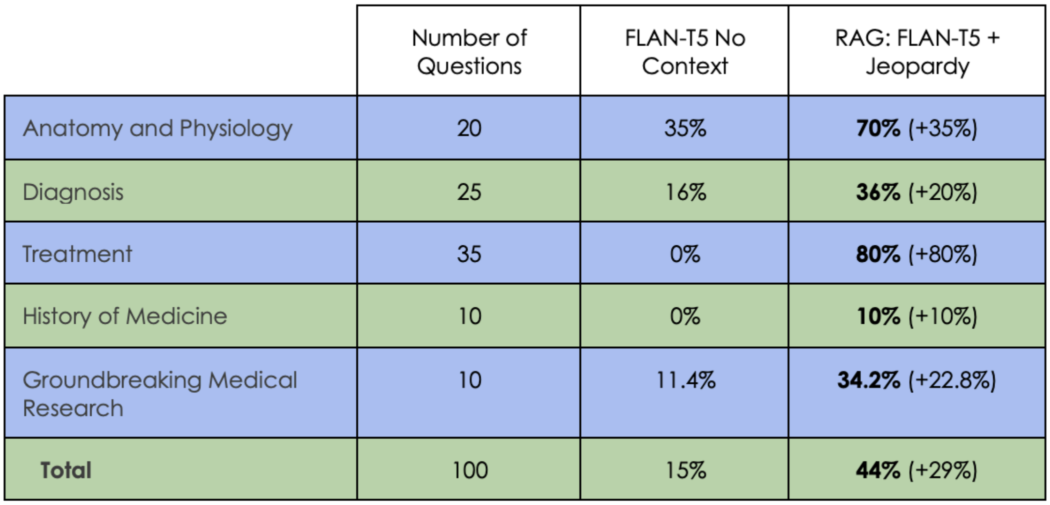
The model is tested on Jeopardy 2024 Tournament of Champions Quarterfinals Questions (n = 536). The results from this task will be compared to the correct answers. FLAN-T5 without context achieved only 2.1%, which is fortunate because it allows room for improvement using context, thereby demonstrating the power of retrieval augmented generation. Accuracy with the addition of context improved the accuracy to 31.7%. This pales however to 39.4% accuracy obtained from only the top ranked embedding by itself and no context. This implies a limitation in FLAN-T5 to answering Jeopardy questions, many of which are oddly clued. Prompt engineering can potentially improve this endeavor, as the base clue can be rewritten in the form of a traditional question.

There was consideration to using OpenAI as the LLM rather than FLAN-T5 given its popularity, as seen in the figure below:


Pre-processing
~450k jeopardy clues from previous seasons
Convert to embeddings via Sentence Transformers (all-mpnet-base-v2)
Vector Store (Qdrant)
2. Embed question via all-mpnet-base-v2 / Semantic Search
3. Retrieve contextual data
1. New Jeopardy question
6. Display response
JeopardyLLM Trivia
4. Prompt + Context
5. Generate response
OpenAI

However, accuracy using this process produced a suspiciously high accuracy of 84.1%, reflecting the possibility that much of OpenAI was trained on Jeopardy questions themselves. In comparison, using OpenAI without context produced a very high accuracy of 90.1%, indicating that the context may be introducing more confusion. Each summarizing the context with OpenAI prior to presentation to FLAN-T5 showed a higher accuracy of 40.7%. The decision was made to avoid OpenAI in the production of our website model as we wanted to focus on an LLM that answers solely with Jeopardy questions. There were cost benefits to using an open-source FLAN-T5 for a website over OpenAI as well. It is worth noting that addition of an LLM actually decreased the accuracy from using the embeddings by themselves. This may be reflective of added confusion and how Jeopardy tends to vary its clue structure. This may be improved with attempts at prompt engineering and fine-tuning.

This process can be conducted for external generalizability as one can wonder about performance on non-Jeopardy questions. For our example, we will test on health questions. 100 questions were generated from OpenAI and formatted in a Jeopardy style, 20 questions taken from ‘Anatomy and Physiology’, 25 from ‘Diagnosis’, 35 from ‘Treatment’, 10 from ‘History of Medicine’, and 10 from ‘Groundbreaking Medical Research. Without any context, FLAN-T5 only answered 15% of questions correctly (mostly from Anatomy and Physiology) while the addition of context allowed for a retrieval-augmented generation process of 44% (a difference of 29%). A table is shown below.



	Number of Questions	FLAN-T5 No Context	RAG: FLAN-T5 + Jeopardy
Anatomy and Physiology	20	35%	**70%** (+35%)
Diagnosis	25	16%	**36%** (+20%)
Treatment	35	0%	**80%** (+80%)
History of Medicine	10	0%	**10%** (+10%)
Groundbreaking Medical Research	10	11.4%	**34.2%** (+22.8%)
Total	100	15%	**44%** (+29%)

5.3 Lessons learned

As a novice front-end developer, we found using the Bootstrap framework to be challenging in certain aspects. While Bootstrap provides a comprehensive set of predefined style guidelines that ensure a uniform aesthetic across HTML components, customizing font styles and sizes was particularly tricky.

Additionally, mastering the Bootstrap grid system presented its own set of challenges. This system allows developers to define the number of columns an element should span. However, issues arise if the content exceeds the number of columns allocated within a single row, causing the extra columns to wrap to a new line. This often complicated the layout when the content volume was larger than the designated column count.

In backend development, this was an excellent opportunity to apply what was learned in the Machine Learning Systems Engineering class to a real-world web application development project. Building an architecture for a real LLM-based application goes beyond the basic architecture used in class and requires considerable effort. An enhanced understanding of computing resource distribution is necessary to manage higher CPU and memory demands compared to traditional web applications.

Furthermore, integrating the backend and frontend while collaborating with other team members presented a significant challenge. Building a CI/CD pipeline is a crucial component of rapid web application development. Providing a solution that works for both development and production environments was another accomplishment of this project. With the completion of the web application, code blocks from various team members can be quickly integrated into the system through an automated pipeline.

6. Conclusion

Our Minimum Viable Product (MVP) showcases its capability to generate responses solely based on past Jeopardy questions. Essentially, it serves as a test to determine whether an extensive repository of Jeopardy knowledge suffices for answering questions beyond those posed in the game show. The project has demonstrated moderate success in addressing newer trivia questions using historical ones. The primary benefit lies in providing swift and accurate responses to trivia style questions However, it's worth noting that existing/competitive solutions are expectedly more robust, drawing from larger and more current data sources and powered by higher computing power.

While our solution is initially tailored for Jeopardy-style questions, its versatility opens up numerous potential applications. Beyond trivia, this Q&A system can be adapted to develop personalized questions based on individual chat histories. Furthermore, it can be employed to query website content, research papers, and other text-heavy materials, expanding its utility beyond the realm of game shows.

Future enhancements are envisioned to focus on improving various aspects of the system. This includes enhancing the user interface for the website, refining the accuracy of the embedding and retrieval model, and addressing grammatical inconsistencies in responses generated by the large language model.

As a part of process, we would like to extend our gratitude to the following individuals for their guidance, advice, feedback and time as we worked through our research:

- Danielle Cummings and Fred Nugen, MIDS 210 Instructors
- MIDS 210 Section 11 Peers
- Mark Butler, MIDS 266 Instructor
- Richard Robbins, MIDS W266 TA
- James York-Winegar, MIDS 255 Instructor
- User testers

Disclaimer: This project was performed for academic purposes only without any commercial intent as a part of the Capstone project for UC Berkeley, Masters of Information and Data Science (MIDS) program.

References

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., ... & Wei, J. (2022). Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Hamilton, B, "A Data Exploration of Jeopardy! from 1984 to the Present" (2020). *CUNY Academic Works*. https://academicworks.cuny.edu/gc_etds/4049

O'Leary, D. (2023). An analysis of Watson vs. BARD vs. ChatGPT: The Jeopardy! Challenge. *AI Magazine*.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Salo D. (2019, Dec) NLP Rules and Models for Jeopardy! Questions and Answers. Dan Salo's Blog https://dancsalo.github.io/2019/12/29/jeopardy/

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, *21*(1), 5485-5551.

Tesauro, G., Gondek, D. C., Lenchner, J., Fan, J., & Prager, J. M. (2013). Analysis of watson's strategies for playing Jeopardy!. *Journal of Artificial Intelligence Research*, *47*, 205-251.

ToughData (2023, Aug) Finetuning Flan-T5 for Question Answering using Quora data. Toughdata. https://www.toughdata.net/blog/post/finetune-flan-t5-question-answer-quora-dataset