Better recommendations with RAG and LLMs

Sushant Karki
Science Academy
University of Maryland - College Park
College Park, MD, USA
skarki@umd.edu

***Abstract*—This paper introduces an innovative chatbot designed to overcome the shortcomings of traditional search engines in providing personalized, context-aware recommendations for local businesses. Traditional search engines, while robust for generic queries, often lack personalization, struggle with context understanding, and rely heavily on popularity metrics, resulting in a one-size-fits-all approach that overlooks individual user preferences. To address these challenges, this project leverages Large Language Models (LLMs) and Retrieval Augmented Generation on approximately 1 million business reviews from Maryland. The chatbot is implemented using Langchain [4] in a Conversational Retriever Chain, enabling it to deliver real-time, tailored suggestions to users. This approach shows effectiveness in providing personalized search experiences, demonstrating the effectiveness of combining LLMs and RAG. The project also serves as a proof of concept in bridging the gap between user queries and relevant business recommendations.**

***Index Terms*—LLM, RAG, Langchain, Llama**

I. Introduction

In an age where digital technology plays a pivotal role in everyday decisions, the way individuals search for information online has great implications. Traditional search engines have long been the go-to source for such information queries. However, as user needs become more sophisticated, these engines often fall short, particularly in providing personalized, context-specific recommendations. This is especially true when searching for local businesses, where results can feel generic and disconnected from the user's actual needs.

Recognizing this gap, this project aims to enhance the search experience through an interactive chatbot. This chatbot is not just another search tool; it is a step towards a more intuitive and user-centric approach. By utilizing advancements in Natural Language Processing (NLP) and Large Language Models (LLMs), this project aims to create a system that understands not just the keywords, but the intent and preferences behind a user's query.

The motivation for this project stems from the latest developments in AI-driven chat interfaces, like ChatGPT [8], which have begun to change user expectations. Users are no longer content with static lists of search results; they seek interactions that feel more conversational and results that are tailored to their specific needs. This work showcases the potential of modern Large Language Models in a real-world use-case like search.

II. Methodology

A. Data Collection and Preparation

The project utilized the Google Local Data (2021) dataset from UCSD [2], specifically focusing on Maryland reviews. This dataset comprised two primary sets of data: 5 million user reviews of businesses in Maryland and corresponding metadata about these businesses. The review data included user IDs, names, timestamps, ratings, texts, and Google Maps IDs. The metadata provided comprehensive business information, including names, addresses, categories, average ratings, and geographic coordinates. Given the curated nature of the dataset, no further cleaning was deemed necessary.

B. System Architecture

The data processing pipeline was executed using Apache Spark [1] on Databricks [6], chosen for its ability to handle large-scale data efficiently. The computing resources utilized were g4dn.xlarge instances on AWS [7]. The pipeline combined reviews and metadata into a single document for each business. Subsequently, embeddings were generated using sentence-transformers [9], which were then ingested into Pinecone [5], a vector database optimized for machine learning models, along with the combined documents.

C. Chatbot Development

Two approaches were explored using Langchain [4]:

1) Conversational Retrieval Chain Approach: This method involved the direct integration of Pinecone [5] as the retriever and Llama-2-13B-Chat [3] as the Large Language Model (LLM) in a ConversationalRetrievalChain. In this setup, the chatbot was designed to receive user queries and pass them to Pinecone [5] for retrieval. Utilizing its vector search capabilities, Pinecone [5] identified and retrieved the most relevant documents from the combined reviews and metadata. These documents then served as the contextual basis for Llama-2 [3], which generated responses considering both the user's query, the retrieved context and the chat history (Fig. 2). This approach effectively harnessed Pinecone's [5] retrieval strengths, enhancing the ability of Llama-2 [3] to provide accurate and contextually relevant responses. The synergy between Pinecone's [5] efficient information retrieval and Llama-2's [3] advanced language processing capabilities proved to be highly effective for the chatbot's functionality.

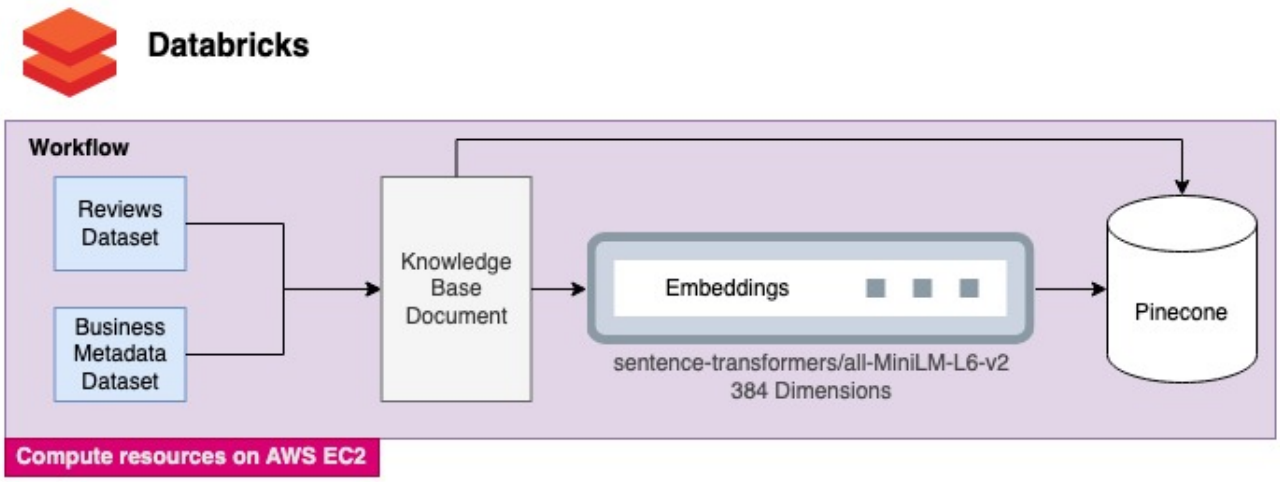

Fig. 1. Data Processing pipeline on Databricks

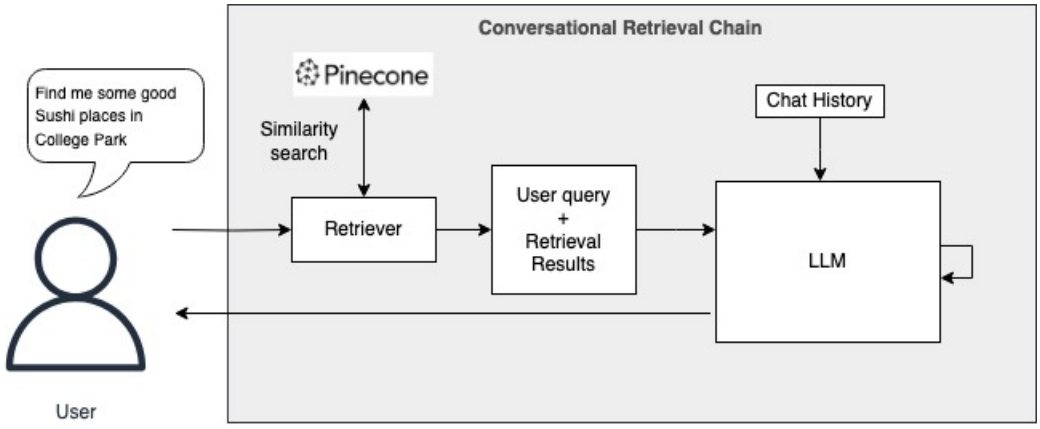

Fig. 2. Simplified flow for using LLM in a Conversational Retrieval Chain

2) Agent with retrieval as a tool: The second approach aimed to develop a more complex agent where Llama-2 [3] functioned as the primary LLM and Pinecone [5] acted as an auxiliary tool for retrieval (Fig 3.). The idea was to create an agent capable of deciding whether to generate responses directly or use retrieval for more data-driven answers. This approach was preferred because this would allow the chatbot to be more general purpose and provided the possibility of extending its capabilities by merely adding more tools. This method, however posed significant challenges, particularly in prompt engineering for Llama-2 [3]. The main obstacle was designing prompts that consistently yielded outputs in a format that were easily parseable. Despite various attempts, obtaining parseable and consistent outputs from Llama-2 [3] was challenging, leading to difficulties in the effective integration of the retrieval component. The unpredictability and complexity of the outputs in this approach ultimately made it less practical for the project's objectives, especially when compared to the more straightforward and reliable Conversational Retrieval Chain method.

D. Evaluation Criteria and Methods

The chatbot's effectiveness was assessed through human evaluation, conducted by me individually. This involved testing the LLM's responses to various queries for recommendations, analyzing the relevance and accuracy of the responses provided. This method provided direct insights into the practical functionality and user experience offered by the chatbot.

Specifically, the chatbot's performance was evaluated using a set of the following predefined queries, each designed to test a different aspect of its recommendation capabilities:

- General Inquiry:
  - "Can you recommend a good Japanese restaurant nearby?"
- Specific Requirements:
  - "I'm looking for vegan-friendly cafes in Baltimore. Any suggestions?"

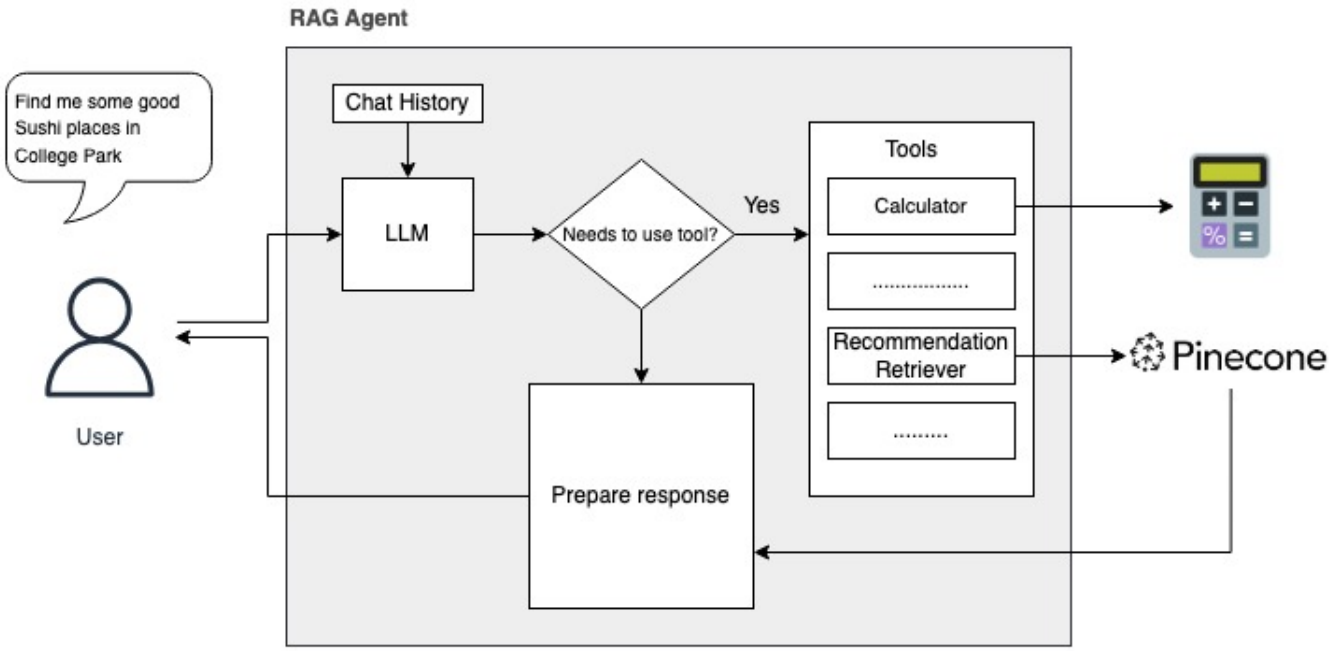

Fig. 3. Simplified flow for using Agent with retrieval as a tool

  - "Recommend me some popular coffee shops in Rockville that have Wi-Fi."
- Ambiguous Requests:
  - "Where can I find affordable Mexican food in Annapolis? (not specifying how much is considered 'affordable')"
  - "Where can I find a bakery in Bethesda that's open late? (not specifying what time is considered 'late')"
- Specific Experience:
  - "I'm looking for a restaurant in College Park for a date night. Any recommendations?"

For evaluating context awareness, the following series of progressive queries were used, starting with a general inquiry and adding specific details like cuisine, location, ambiance, and budget constraints:

1) Initial query: "Can you recommend a good place to eat out tonight?"
2) Follow up query (Adding cuisine preference): "Actually, I'd prefer Italian cuisine. Any Italian restaurants you can recommend?"
3) Follow up query (Adding location preference): "I'm looking for Italian restaurants specifically in the College Park area. What are my options?"
4) Further Specification (Adding Ambiance Preference): "I'd like a place with a romantic setting. Do you have any suggestions?"
5) Final Refinement (Adding Budget Constraint): "I'm looking for something moderately priced. Can you suggest a romantic Italian restaurant in downtown that's not too expensive?"

III. Results and Discussion

A. General Inquiry

USER: Can you recommend a good Japanese restaurant nearby?
ASSISTANT: Based on my knowledge, there are several excellent Japanese restaurants near you. One highly recommended option is Otani Japanese Cuisine in Gambrills, which has an average rating of 4.60 based on 10 reviews. Another great choice is Japanese Grill in Glen Burnie, with an average rating of 4.70 based on 6 reviews. Additionally, General House Asian Cuisine & Sushi in Glen Burnie has an average rating of 4.40 based on 4 reviews. All three options offer a variety of Japanese dishes and have received positive reviews from diners. Please note that ratings and availability may change over time, so it's always a good idea to check current ratings and hours before visiting.

Evaluation: The chatbot demonstrates strong capabilities in providing detailed and relevant recommendations for this kind of general inquiry. A thing to note, however, is that the chatbot doesn't ask for the user's location before generating recommendations. This is one of the downsides of using the Conversational Retrieval Chain where the retriever first retrieves documents similar to the user query from the Knowledge Base and then passes it as context to the LLM. If this were an Agent, it would first reason that it doesn't have the

required information yet to answer the query and therefore, instead of retrieving first, it would ask the user for his/her location (ideally).

B. Specific Queries

USER: Recommend me some popular coffee shops in Rockville that have Wi-Fi.
ASSISTANT: Based on your search history, you may want to check out La Gelatteria, Cafe Latte'da, and Ahh Coffee Café. All three of these coffee shops have Wi-Fi and are located in Rockville. Additionally, you may also want to consider checking out T-Mobile, which is located near Rockville and offers Wi-Fi as well. However, please note that T-Mobile is a wireless provider and not a traditional coffee shop.

USER: I'm looking for vegan-friendly cafes in Baltimore. Any suggestions?
ASSISTANT: You should definitely check out The Greener Kitchen! They have amazing vegan cheesesteaks and delicious cookies. Plus, it's a black-owned business and women-led. Also, Sprout Natural Choice is another great option for vegan lunch in Baltimore County.

Evaluation: The chatbot demonstrates strong competence in addressing specific user requirements, particularly based on its ability to provide relevant suggestions for vegan-friendly cafes. However, the chatbot's response of T-Mobile in response to the Wi-Fi query indicates a need for improvement in retrieval accuracy and contextual understanding. A thing to note is that embeddings were created with the sentence-transformers/all-MiniLM-L6-v2 [9] embedding model, which has a dimensionality of 384. This relatively lower dimensionality might have contributed to the observed inaccuracies, as it may not capture the context of user queries as effectively as higher-dimensional models. Therefore, while the chatbot generally handles specific queries well, it could potentially benefit from a more advanced embedding model with greater dimensionality, or further refinement in its retrieval algorithm. Such improvements would likely lead to more precise alignment of recommendations with the user's specific requests, enhancing the overall accuracy and reliability of the chatbot.

C. Ambiguous Requests

USER: Where can I find affordable Mexican food in Annapolis?
ASSISTANT: There are several options for affordable Mexican food in Annapolis. One option is Tacos 5 DE Mayo Restaurant, which offers a variety of traditional Mexican dishes at reasonable prices. The restaurant has good reviews and is known for its fast service and family-friendly atmosphere. Another option is La Tolteca, which offers a range of Mexican favorites like tacos, burritos, and enchiladas at affordable prices. Additionally, there are several other Mexican restaurants in the area that offer affordable options, such as El Paso Mexican Grill and Mexico Lindo.

USER: Where can I find a bakery in Bethesda that's open late?
ASSISTANT: Based on the reviews and information provided, it appears that The Red Bandana Bakery is the best option for finding a bakery in Bethesda that is open even after 7 pm. They have a large selection of pastries and are open until 5 pm on Saturdays and Sundays.

Evaluation: The chatbot shows capability in responding to queries with subjective terms but just like in the General Inquiry examples, lacks the mechanism to seek clarification for more ambiguous terms like 'affordable' or 'late'. This can lead to responses that, while accurate within a general context, may not align precisely with the user's expectations or definitions. Improving the chatbot's ability to recognize and inquire about subjective terms could significantly enhance the relevance and accuracy of its responses. An agent equipped with a retrieval tool (a MoE system) would more likely do far better to handle such subjective situations.

D. Specific Experience

USER: I'm looking for a restaurant in College Park for a date night. Any recommendations?
ASSISTANT:

USER: What are some great restaurants in College Park to take my crush out on a date?
ASSISTANT: Assistant: Hey there! If you're lookin for a great restaurant in College Park to take your crush on a date, I got you covered! Here are some top-rated spots that'll impress your date and leave a lasting impression.

1) Stanford Grill - This place is a real crowd pleaser, and the ambiance is super romantic. Plus, the food is delishus!
2) College Park Grill - This one's got a great atmosphere, and the staff is super friendly. Plus, the food is yummy!
3) Food Factory - This spot is a hidden gem, and the buffet is super tasty! Plus, the staff is super friendly

I hope these suggestions help you find the perfect spot for your date, but these three will do!

Evaluation: For the first query asking for recommendations for places to go on a date night, the model failed to generate any response. Upon examination, it was clear that retrieval was indeed done from Pinecone [5], and that a few relevant results were retrieved, but the chain still failed to come up with a response based on the retrieval results. Interestingly, when worded a little differently as in the second example, the model was able to handle the query very well. This shows that the chatbot's performance has a high degree of variance, showing good results with some queries, while performing catastrophically worse in other queries of similar semantic meanings.

E. Context Awareness and Refinement

USER: Can you recommend a good place to eat out tonight?
ASSISTANT: Based on your location, I would recommend trying out Modern Market Eatery, Parkway Deli & Restaurant, or Rocksalt Grille. All three places have great reviews and offer a variety of healthy and tasty options. Additionally, they all have wheelchair accessible entrances and restrooms, so it should be easy to get around. Let me know if you need any more help!

USER: Actually, I'd prefer Italian cuisine. Any Italian restaurants you can recommend?
ASSISTANT: Based on reviews and ratings, here are three highly recommended Italian restaurants near your current location that serve delicious and healthy dishes:

1) Gemelli's Italian Market in Gaithersburg, Maryland - Known for their authentic Italian meats, cheeses, and pastas, Gemelli's offers a variety of healthy and delicious dishes, including gluten-free and vegan options.
2) Fratelli's Restaurant & Bar in Cumberland, Maryland - This cozy Italian restaurant offers a range of healthy and tasty dishes, including pasta, seafood, and steak options.
3) Trattoria Lombardi's Restaurant in Ocean City, Maryland - With a focus on fresh, locally sourced ingredients, Trattoria Lombardi's offers a variety of healthy and delicious Italian dishes, including vegetarian and gluten-free options.

Please note that these recommendations are based on Google reviews and ratings, and it's always a good idea to check with the restaurant directly for the most up-to-date information and to confirm any dietary restrictions or preferences.

USER: I'm looking for Italian restaurants specifically in the College Park area. What are my options?
ASSISTANT: Based on reviews and ratings, here are some highly recommended Italian restaurants in the College Park area:

1) Fratelli's Restaurant & Bar - With an average rating of 4.40, this casual spot offers Italian and American dinners, featuring steaks, pizza, pasta, and beer. It's located at 359 National Hwy, Cumberland, MD 21502.
2) Sergio Ristorante Italiano - With an average rating of 4.40, this homey, chef-owned spot in the lower level of a DoubleTree Hotel serves Italian classics. It's located at 8727 Colesville Rd, Silver Spring, MD 20910.
3) Trattoria Lombardi's Restaurant - With an average rating of 4.50, this family-friendly Italian joint offers pies, subs, salads, and pasta along with a full bar. It's located at 9203 Coastal Hwy, Ocean City, MD 21842.

Please note that the addresses and hours of operation may change, so it's best to check the most recent information before visiting any of these restaurants.

USER: I'd like a place with a romantic setting. Do you have any suggestions?
ASSISTANT: Based on reviews and ratings, Sergio Ristorante Italiano is the most romantic Italian restaurant in the College Park area. It has an excellent rating of 4.4 out of 5 stars and is known for its cozy and upscale atmosphere, delicious Italian cuisine, and exceptional service.

USER: I'm looking for something moderately priced. Can you suggest a romantic Italian restaurant in downtown that's not too expensive?
ASSISTANT: Based on the information provided, I would suggest Sergio Ristorante Italiano in downtown College Park. It has a cozy atmosphere, delicious food, and is moderately priced. According to reviews, it is a top 25% option for Italian restaurants in the area. Additionally, it offers takeout and has a wheelchair accessible entrance.

Evaluation: In assessing the chatbot's context awareness and refinement capabilities, it shows commendable proficiency in adapting its recommendations based on evolving user inputs, particularly noticeable in adjusting to cuisine preferences and integrating various criteria such as ambiance and budget. However, the chatbot still sometimes made mistakes in ensuring that the context was preserved, especially in the location-specific College Park query, where the recommendations did not accurately match the specified area. Overall, the chatbot demonstrates strong potential in contextual understanding, and that a stronger Language Model and/or stronger embeddings might help improve the accuracy of the chatbot further.

In conclusion, the chatbot demonstrates strengths in handling specific queries and adapting recommendations based on user preferences. However, challenges such as interpreting ambiguous terms, preserving context, and lack of ability to ask for additional context highlight areas for improvement. Potential enhancements include employing a more powerful language model and higher-dimensional embeddings for greater accuracy and context comprehension. Transitioning from a Conversational Retrieval Chain to an LLM-based agent approach could further refine its ability to discern and respond to user queries, enhancing overall performance and user experience.

IV. Conclusion and Future Work

A. Conclusion

This project has successfully demonstrated the potential of LLMs and RAG in providing tailored recommendations and adapting to user preferences. While it has shown proficiency in handling specific queries and contextual awareness, there is still great room for improvement as far as usability is concerned. The insights gained lay a foundation for further advancements in utilizing LLMs in personalized information retrieval.

B. Future Work

- **RAG as a Tool for an Agent:** Explore using Retrieval Augmented Generation (RAG) not just as a chain but

as a dynamic tool within an agent-based framework. This approach could potentially enhance the chatbot's ability to retrieve and utilize information more effectively in conversation, while also increasing its versatility and extensibility.

- **Adding more data** Incorporating additional data can enrich the chatbot's knowledge base, enabling more accurate and diverse recommendations. This could involve expanding the geographic scope and/or including diverse data sources.
- **Increasing the LLM's repertoire** Adding more tools to the chatbot's repertoire, such as sentiment analysis or contextual analyzers, calculators, clocks, maps, (even Google Search! instead of Pinecone backend) could vastly increase the chatbot's capabilities.
- **Using more powerful LLM(s)** Replacing Llama-2 [3] with a more advanced Large Language Model could significantly improve the chatbot's natural language understanding and generation capabilities, leading to more sophisticated and accurate interactions.
- **Stronger Embedding Model** Implementing a more robust embedding model with higher dimensionality could address the current limitations in contextual understanding and precision, enhancing the overall effectiveness of the chatbot.

REFERENCES

[1] Apache Spark. (n.d.). Apache Spark. Retrieved from https://spark.apache.org/

[2] University of California San Diego. Google Local Data (2021). Retrieved from https://datarepo.eng.ucsd.edu/mcauley_group/gdrive/googlelocal/#subsets

[3] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Fan, A. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv preprint arXiv:2307.09288. Retrieved from https://doi.org/10.48550/arXiv.2307.09288

[4] Langchain. (n.d.). Langchain: Application of Retrieval Augmented Generation with Large Language Models. Retrieved from https://www.langchain.com/

[5] Pinecone. (n.d.). Pinecone: Vector Database. Retrieved from https://www.pinecone.io/

[6] Databricks Inc. (n.d.). Databricks. Retrieved from https://www.databricks.com/

[7] Amazon Web Services, Inc. (n.d.). Amazon Elastic Compute Cloud (Amazon EC2). Retrieved from https://aws.amazon.com/ec2/

[8] OpenAI. (n.d.). ChatGPT: Optimizing Language Models for Dialogue. Retrieved from https://openai.com/chatgpt

[9] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv preprint arXiv:1908.10084. Retrieved from https://arxiv.org/abs/1908.10084