

Custom Chatbot for any Website with Retrieval Augmented Generation

Applied Machine Learning

Roudranil Das Soham Sengupta Subhashree Saha Ujan Dasgupta

Chennai Mathematical Institute
MSc. Data Science

April 27, 2024

Table of Contents

- 1 Introduction
 - Problem statement
 - When and not to apply LLM's
- 2 Creating the chatbot
 - Data curation
 - Retrieving stored information
 - LLM's used
- 3 Comparison of results
- 4 Conclusion
 - Challenges and trade-offs
 - Division of responsibilities

Problem statement

- The objective is to create a natural language question answering chatbot on any given website (public or private)

Problem statement

- The objective is to create a natural language question answering chatbot on any given website (public or private)
- With the help of
 - Large Language Models
 - LangChain
 - Embeddings, Information retrieval and vector stores.

When to and not to apply LLM's



Data curation

- For our problem the data was curated from Recipes around the world (wikibooks.com)

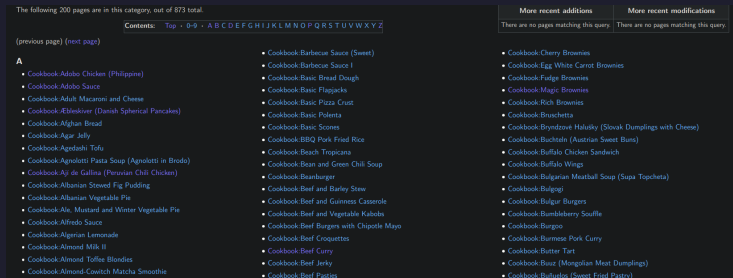


Figure: Screenshot of how the recipes were in the website

- Data collection: we used a web crawler (scrapy) to crawl all the pages on this site, extract the text and save them in a JSON file.

Data curation

```
1  [  
2    {  
3      "url": "https://en.wikibooks.org/wiki/Cookbook:Aubergine_and_Onion_Vegetable_Pie",  
4      "text": "Cookbook:Aubergine and Onion Vegetable Pie\ndes ..."  
5    },  
6    {  
7      "url": "https://en.wikibooks.org/wiki/Cookbook:Palatschinken_(Czech/Austrian_Crepes)",  
8      "text": "Cookbook:Palatschinken (Czech/Austrian Crepes)\nde ..."  
9    },  
10   { ... },  
11  ]
```

Figure: Preview of the JSON file where the recipes are saved.

Data curation

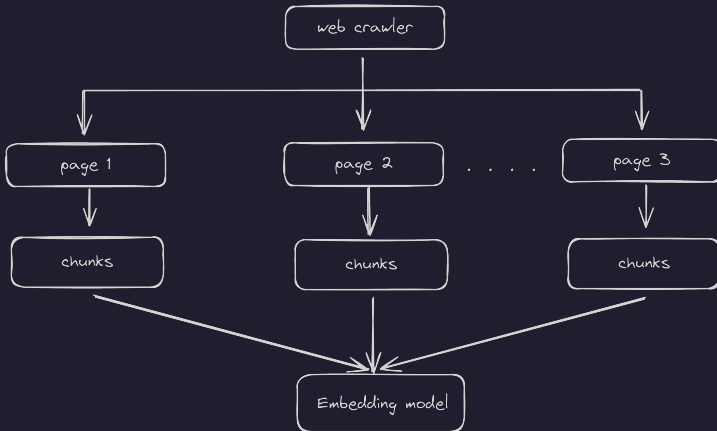
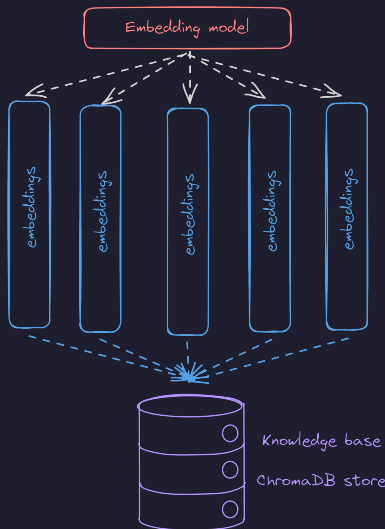
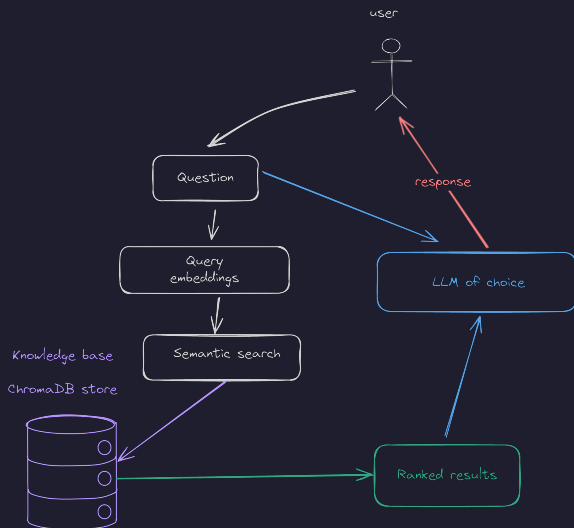


Figure: Crawling process

Retrieving stored information



Retrieving stored information



LLM's used

- In this work we use and compare LLM's of 3 different sizes:
 - TinyLlama 1B
 - Zephyr 3B
 - WizardLM 7B
- Each of the models are downloaded and loaded in 8-bit precision after quantization. This is done to reduce GPU memory footprint.
- We use e5-large-v2 sentence embeddings. The embedding dimension is 1024.
- Finally we compare the outputs of the 3 LLM's on the same prompt.
- The pipeline was built with the LangChain framework.

Comparison of results

Prompt: *Give the cooking procedure (with ingredients and instructions) for a chicken curry dish in the cookbook.*

```

Cookbook: Spicy Chilli Chicken
Desc
Chile chicken, cooked Kerala style.
Ingredients
1 pound boneless chicken meat
1 tbsp masala
1 cup lemon juice
Salt to taste
3 tbsp canola oil
1 1/2 ea. red onion, chopped
3/4 tbsp red chilli powder
3/4 tsp turmeric powder
1 tsp smoked paprika
1 tsp coriander
3 medium tomatoes, chopped
3/4 tbsp minced ginger
2 tbsp minced garlic
6 tbsp tomato ketchup
8 tbsp soy sauce
4 Indian green chiles, chopped
2 green bell peppers, sliced
Procedure
Marinate the chicken with lemon juice, masala, and salt for 1 hour.
In a large pot, fry the marinated chicken in canola oil until golden brown (5-7 minutes).
Remove the chicken from the pot and cut into pieces.
Into the remaining oil, add the onions, red chilli powder, turmeric, paprika, and coriander
powder, and saute
until wilted.
Add the tomatoes, ginger, and garlic, and cook till the tomatoes are softened and breaking
apart.
Put the ketchup, soy sauce, fried chicken pieces, green chiles, and green peppers in the pot,
and simmer until
peppers are soft and chicken is fully-cooked.
Serve over steamed rice.

```

Figure: TinyLlama 1B

Comparison of results

Prompt: *Give the cooking procedure (with ingredients and instructions) for a chicken curry dish in the cookbook.*

The cooking procedure for a chicken curry dish in the Spicy Chilli Chicken cookbook involves the following steps using the listed ingredients:

1. Marinate boneless chicken meat with lemon juice, masala, and salt for 1 hour.
2. Fry the marinated chicken in canola oil until golden brown. Remove the chicken from the pot and cut it into pieces.
3. In the same oil, add red onions, red chili powder, turmeric powder, smoked paprika, and coriander powder.
Sauté them until they become wilted.
4. Add chopped tomatoes, ginger, and garlic, and cook until the tomatoes soften and break apart.
5. Mix tomato ketchup, soy sauce, fried chicken pieces, Indian green chiles, and green bell peppers into the pot. Simmer everything together until the peppers become soft and the chicken is fully cooked.
6. Serve the spicy chicken curry over steamed rice.

Figure: Zephyr 3B

Comparison of results

Prompt: *Give the cooking procedure (with ingredients and instructions) for a chicken curry dish in the cookbook.*

Sure! Here is the recipe for Chicken Wings Asian Style from the cookbook you provided:

Ingredients:

- 24 chicken wingettes
- Poultry shake, as needed
- Cornmeal, as needed
- 2 eggs, beaten
- 6 tbsp hoisin sauce
- 1 tbsp dark soy sauce
- 2 tbsp hot sauce
- 1/2 tbsp minced garlic
- 1/2 tbsp minced ginger
- Spicy garlic oil for deep frying

Instructions:

1. Coat the chicken wingettes with egg. Season them liberally with poultry shake.
2. In a separate bowl, mix together the cornmeal, salt, and black pepper.
3. Dip each chicken wingette into the cornmeal mixture, making sure they are fully coated.
4. Heat up the spicy garlic oil in a deep fryer or a large pot over medium heat.
5. Once the oil is heated through, add the chicken wingettes in batches, making sure not to overcrowd the pan.
6. Cook the chicken for 12 minutes, or until they are golden brown and crispy on the outside.
7. Remove the chicken from the oil using a slotted spoon and transfer them onto a wire rack to drain excess oil.
8. In a small bowl, combine the remaining ingredients (hoisin sauce, dark soy sauce, hot sauce, minced garlic, minced ginger, and salt) and stir well.
9. Toss the chicken wingettes in the sauce until they are evenly coated.
10. Serve the chicken warm with your favorite dipping sauce.

Figure: WizardLM 7B

Comparison of results

Prompt: *Give the ingredients for an Asian spicy dish in the cookbook*

In the cookbook, the recipe calls for 2 tbsp coconut oil, 1 big onion, 3 cloves garlic, 2 cm piece ginger, 2 scallions, 50 g red chile, 1 cup basmati rice, 30 g chile powder, 3 tbsp light soy sauce, 2 tbsp fish sauce, 1 tbsp chicken powder, 1 tbsp salt, and 1 tbsp pepper.

Figure: TinyLlama 1B

Comparison of results

Prompt: *Give the ingredients for an Asian spicy dish in the cookbook*

The ingredients for an Asian spicy dish in the cookbook include 2 tablespoons of coconut oil, 1 big onion, 3 cloves of garlic, 2 centimeters of ginger, 2 scallions, 50 grams of red chilies, 1 cup of basmati rice, 30 grams of chili powder, 3 tablespoons of light soy sauce, 2 tablespoons of fish sauce, 1 tablespoon of chicken powder, 1 tablespoon of salt, and 1 tablespoon of pepper.

Figure: Zephyr 3B

Comparison of results

Prompt: *Give the ingredients for an Asian spicy dish in the cookbook*

The ingredients for an Asian spicy dish in the cookbook are:

- 2 tbsp coconut oil
- 1 big onion, diced
- 3 cloves garlic, crushed
- 2 cm piece ginger, minced
- 2 scallions, cut into 1.5 cm pieces
- 50 g red chile, cut into 1 cm pieces
- 1 cup (900 g) basmati rice, cooked
- 30 g chile powder
- 3 tbsp light soy sauce
- 2 tbsp fish sauce
- 1 tbsp chicken powder
- 1 tbsp salt
- 1 tbsp pepper

Figure: WizardLM 7B

Comparison of results

- We see that the 7B model gives much more verbose and well-formatted results than the smaller models.
- However in the scale of sizes of LLM's, size difference between 1B, 3B, 7B is not so significant.
- Hence we do not see extremely large differences between their performances
- However in some cases, the smaller models execute faster than the bigger models.
- In some of the inferences regarding listing some recipes, the smaller models tend to hallucinate, but the larger models do not.

Challenges and trade-offs

We faced the following challenges:

- LLM models are prone to hallucinate, even when feeding a context with RAG
- The process is GPU intensive, and required us to use platforms such as colab and kaggle
- There is a trade-off of performance vs cost - OpenAI api's are better but expensive than the free alternatives
- There is a trade-off of performance vs compute - because of less availability of GPU memory, we can only experiment with smaller models

Division of responsibilities

Name	Responsibility
Ujan Dasgupta	Web scraping, data preprocessing and preparation
Subhashree Saha	Building the pipeline and chatbot framework with Langchain
Soham Sengupta	Experiment performance of different LLM's on different queries
Roudranil Das	Web scraping, code versioning and making the PPT

Code Walkthrough