Milestone 2 Experiment Log
Pourfect AI

We are using the following model specifications for our experiments:

- EMBEDDING_MODEL = "text-embedding-004"
- GENERATIVE_MODEL = "gemini-1.5-flash-001"
- EMBEDDING_DIMENSION = 256
- 10 documents as context

Experiment 1 - LLM Model with and without RAG documents

- We first test Gemini 1.5's ability to answer questions about cocktails without any additional context from our RAG model. Then, we compare the output with providing context from the top 10 documents from the RAG model.
- The first user prompt without context is "Answer the following query: {query}"" and the model provided the following output:
    - Query 1: Can you give me a recipe for an Espresso Martini?
    - LLM Response: The Espresso Martini is a coffee-flavored cocktail, typically served in a martini glass.
    - **Ingredients:**
    - * 1 1/2 oz Vodka
    - * 1 oz Coffee Liqueur (such as Kahlúa)
    - * 1 oz Espresso, freshly brewed and chilled
    - **Preparation:**
    - 1. Combine all ingredients in a cocktail shaker filled with ice.
    - 2. Shake vigorously until well-chilled.
    - 3. Strain into a chilled martini glass.
    - **Garnish:**
    - * Coffee beans
    - The Espresso Martini is a strong and flavorful cocktail that is perfect for a night out. The coffee flavor is balanced by the sweetness of the coffee liqueur and the smoothness of the vodka.

    - Query 2: Can you give me a recipe for a refreshing, fruity, summer mocktail?
    - Query: Can you give me a recipe for a refreshing, fruity, summer mocktail?
    - LLM Response: ## Strawberry Basil Spritzer
    - **Type:** Mocktail
    - **Glass:** Highball glass
    - **Ingredients:**
    - * 1 cup sliced fresh strawberries
    - * 1/2 cup fresh basil leaves
    - * 1/2 cup sugar
    - * 1 cup water
    - * 1/2 cup sparkling water
    - * Ice

    - **Preparation:**
    - 1. Combine strawberries, basil, sugar, and water in a saucepan.
    - 2. Bring to a boil over medium heat, then reduce heat and simmer for 5 minutes.
    - 3. Remove from heat and let cool completely.
    - 4. Strain the mixture through a fine-mesh sieve, discarding the solids.
    - 5. Fill a highball glass with ice.
    - 6. Add the strawberry basil syrup and top with sparkling water.
    - 7. Garnish with a fresh strawberry and basil sprig.
    - **Description:** This refreshing mocktail is perfect for a hot summer day. The sweetness of the strawberries and the herbaceousness of the basil create a delightful flavor combination. The sparkling water adds a touch of fizz, making it a truly refreshing drink.

- The second user prompt with context is "Here are some text chunks that may contain relevant information: {context}. Based on the above information, please answer the following question: {query} " and the model provided the following output:
    - Query 1: Can you give me a recipe for an Espresso Martini?
    - LLM Response: The Espresso Martini is a classic cocktail that has become quite popular.
    - **Espresso Martini**
    - * **Category:** Classic
    - * **Glass:** Martini glass
    - * **Ingredients:**
    - * 1 shot Vodka
    - * 1 shot Kalhua
    - * 1 shot Espresso
    - * **Method:** Shake and fine strain into a chilled Martini glass.
    - * **Garnish:** 3 coffee beans
    - It is important to use freshly brewed espresso in this cocktail, as substituting coffee or other beverages will alter the flavor profile. While some bartenders add a touch of cream on top, it is not essential to the recipe.

    - Query 2: Can you give me a recipe for a refreshing, fruity, summer mocktail?
    - LLM Response: Here is a recipe for a refreshing, fruity, summer mocktail:
    - **Name:** Watermelon Slush
    - **Type:** Mocktail
    - **Glass:** Highball glass
    - **Ingredients:**
    - * 2 cups Watermelon
    - * 2 wedges Lime
    - * Lime Juice to taste
    - **Preparation:**
    - 1. Blend the watermelon until smooth.
    - 2. Add lime juice to taste.
    - 3. Pour into a highball glass filled with ice.
    - 4. Garnish with a lime wedge.

    - **Description:** This mocktail is refreshing and sweet with a citrus bite, perfect for summer bonfires and beach trips.
- For query 1, the models provide a relatively similar output because an Espresso martini is a very common cocktail that Gemini likely already can provide instructions to.
- For query 2, when we describe a type of drink but don't give the name of a popular cocktail like Espresso martini, the LLM with context from RAG provides a more unique drink with a better description. While Gemini without context suggests sparkling water with strawberries, the model with context provides a more unique and appealing drink (a watermelon slush).

Experiment 2 - Changing the Prompt to the Model

- We then experimented with using the prompt we were given during class, but changing "cheese" to cocktail. We call this prompt SYSTEM_INSTRUCTION_1
    - Output with system instruction version 1 (SYSTEM_INSTRUCTION_1):
    - Query: Can you give me a recipe for an Espresso Martini?
    - LLM Response: The Espresso Martini is a classic cocktail made with:
    - * 1 shot Vodka
    - * 1 shot Kalhua
    - * 1 shot Espresso
    - The method is to shake and fine strain into a chilled martini glass. Garnish with 3 coffee beans.
- Then, we updated the prompt to improve the LLM output. We removed some of the hard constraints from the first prompt to allow the model to be more creative when designing recipes. We call this prompt SYSTEM_INSTRUCTION_2
    - Output with system instruction version 2 (SYSTEM_INSTRUCTION_2):
    - **Query:** Can you give me a recipe for an Espresso Martini?
    - LLM Response: The Espresso Martini is a classic cocktail that has become quite popular.
    - ****Espresso Martini****
    - * ****Category:**** Classic
    - * ****Glass:**** Martini glass
    - * ****Ingredients:****
    - * 1 shot Vodka
    - * 1 shot Kalhua
    - * 1 shot Espresso
    - * ****Method:**** Shake and fine strain into a chilled Martini glass.
    - * ****Garnish:**** 3 coffee beans
    - It is important to use freshly brewed espresso in this cocktail, as substituting coffee or other beverages will alter the flavor profile. While some bartenders add a touch of cream on top, it is not essential to the recipe.
- This experiment highlights that the model is able to create a more cohesive recipe and provide better descriptions with the second system prompt. We will continue with the prompt engineering process following Milestone 2.

Experiment 3 - Changing the number of documents provided as context to the model

- We ran all of these experiments using SYSTEM_INSTRUCTION_2
- We first ran the model with 10 documents from the vector database:
    - **Query:** Can you give me a recipe for an Espresso Martini?
    - LLM Response: The Espresso Martini is a classic cocktail, typically served in a martini glass.
    - ****Ingredients:****
    - * 1 shot Vodka
    - * 1 shot Kalhua
    - * 1 shot Espresso
    - ****Preparation:****
    - Shake the ingredients together and fine strain into a chilled martini glass.
    - ****Garnish:****
    - 3 coffee beans
    - ****Description:****
    - The Espresso Martini is a strong and flavorful cocktail that combines the boldness of espresso with the smoothness of vodka and the sweetness of Kalhua. It's a popular after-dinner drink and is often enjoyed as a pick-me-up.
    - ****Note:****
    - The text chunks provide multiple recipes for the Espresso Martini, but they all share the same core ingredients and method. Some recipes suggest adding a dash of Angostura bitters or a dollop of cream, but these are optional additions.
- Then we ran the model with 25 documents:
    - **Query**: Can you give me a recipe for an Espresso Martini?
    - LLM Response: The Espresso Martini is a classic cocktail, typically served in a martini glass.
    - ****Ingredients:****
    - * 1 shot Vodka
    - * 1 shot Kalhua
    - * 1 shot Espresso
    - ****Method:****
    - * Shake all ingredients with ice and fine strain into a chilled martini glass.
    - ****Garnish:****
    - * 3 coffee beans
    - ****Description:****
    - This is a simple and straightforward recipe that doesn't require much modification. The espresso should be freshly made and not substituted with coffee or other alternatives. The Kalhua adds sweetness to the drink, so there is no need for additional sugar. Some bartenders like to add a hand-wiped cream on top, but this can be a bit much after a while.
- Finally we ran the model with 50 documents:
    - **Query:** Can you give me a recipe for an Espresso Martini?
    - LLM Response: The Espresso Martini is a classic cocktail, typically served in a martini glass.

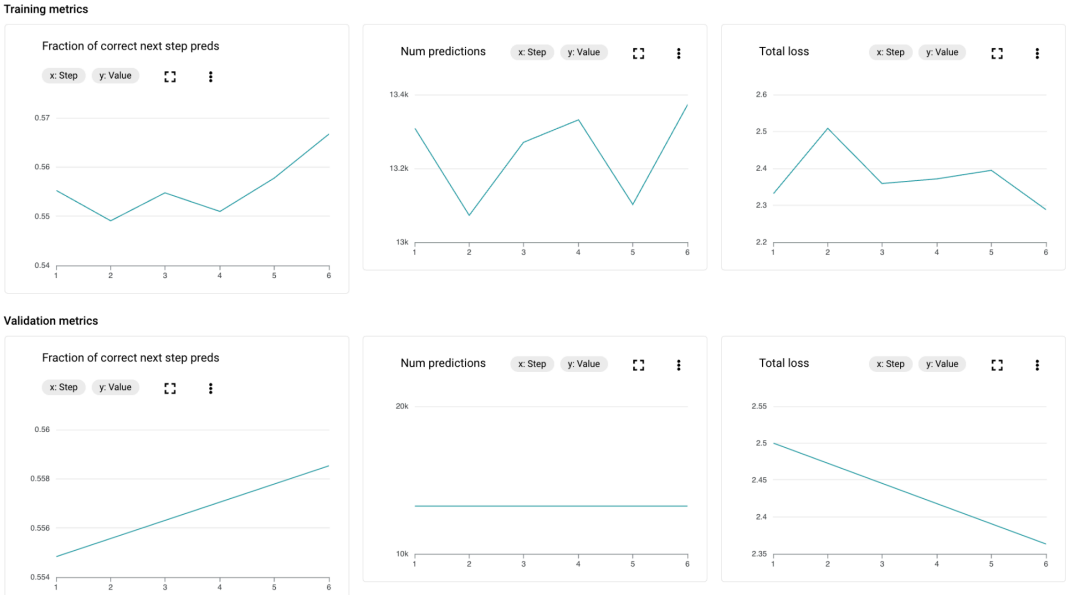
    - **Ingredients:**
    - * 1 shot Vodka
    - * 1 shot Kalhua
    - * 1 shot Espresso
    - **Method:**
    - Shake all ingredients with ice and fine strain into a chilled martini glass.
    - **Garnish:**
    - 3 coffee beans
    - **Description:**
    - The Espresso Martini is a sweet and strong cocktail with a rich coffee flavor. It is a popular after-dinner drink and is often enjoyed as a pick-me-up.
    - **Note:** The text chunks provide multiple recipes for the Espresso Martini, but they all share the same core ingredients and method. Some recipes suggest adding a dash of Angostura bitters or a dollop of cream, but these are optional additions.
- With only 10 documents, the model provided a simple description and added a note about how some recipes suggested different additions among the documents. With the 25 documents, the response was slightly more elaborate and provided information on the ingredients quality as well as an optional garnish modification. With 50 documents, the model response was overall very similar to the response with 25 documents.
- Thus, going forward we will likely use around 25 documents to ensure that we get comprehensive recipes and descriptions without providing any unrelated documents.

Experiment 4: Fine Tuning the LLM: Changing epochs and token size

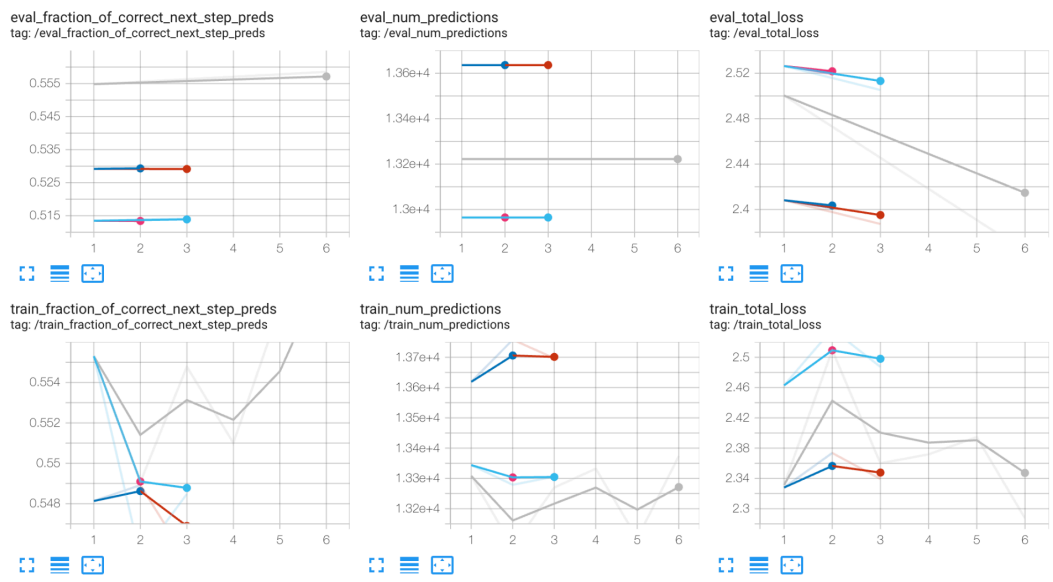
- We define the number of training examples as the number of examples we generate in the finetuning_data.py file. It then runs this for a total of 5 iterations. In this file we also specify the maximum number of tokens.

Fine Tuning Experiment Number	Epochs	Max Token Size	Number of Training Examples	Number of Iterations	Eval Fraction of Preds Correct	Eval Total Loss
1	2	4,096	20	5	0.513	2.52
2	4	4,096	20	5	0.514	2.50
3	2	8,192	20	5	0.529	2.40
4	4	8,192	20	5	0.529	2.39
5	2	8,192	50	5	0.513	2.52
6	4	8,192	50	5	0.559	2.36

Best performing model: Experiment #6



Comparison of all fine tuned models in TensorBoard:



- As we continue with this project, we will fine tune even further based off of experiment #6 as well as integrating the RAG pipeline with the fine-tuned Gemini model to achieve even higher validation metrics.