RAG System Optimization Report

1. Problem Description

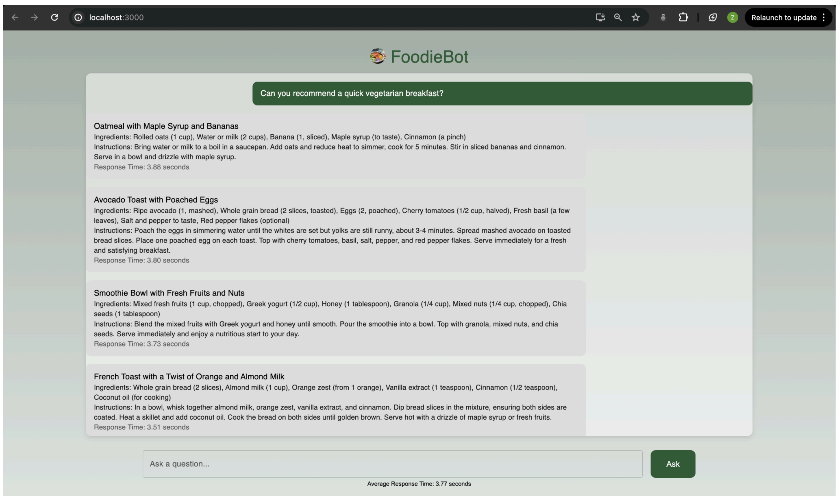
I developed a RAG-based chatbot for providing recipe recommendations. After initial development, I identified issues with the accuracy of responses and response time.

2. Issues Before Optimization

Accuracy Issues:

Before optimization, the system included some irrelevant answers. For example:

1. **Question**: "Can you recommend a quick vegetarian breakfast?"



Can you recommend a quick vegetarian breakfast?

Oatmeal with Maple Syrup and Bananas
Ingredients: Rolled oats (1 cup), Water or milk (2 cups), Banana (1, sliced), Maple syrup (to taste), Cinnamon (a pinch)
Instructions: Bring water or milk to a boil in a saucepan. Add oats and reduce heat to simmer, cook for 5 minutes. Stir in sliced bananas and cinnamon. Serve in a bowl and drizzle with maple syrup.
Response Time: 3.88 seconds

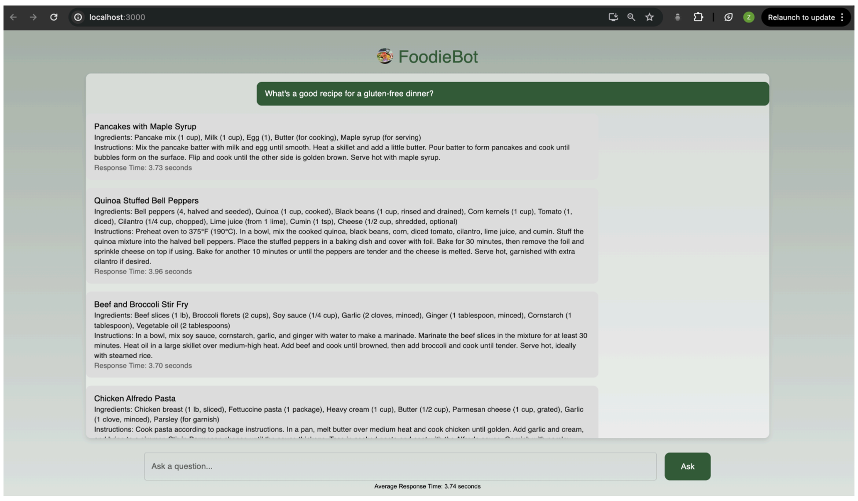
Avocado Toast with Poached Eggs
Ingredients: Ripe avocado (1, mashed), Whole grain bread (2 slices, toasted), Eggs (2, poached), Cherry tomatoes (1/2 cup, halved), Fresh basil (a few leaves), Salt and pepper to taste, Red pepper flakes (optional)
Instructions: Poach the eggs in simmering water until the whites are set but yolks are still runny, about 3-4 minutes. Spread mashed avocado on toasted bread slices. Place one poached egg on each toast. Top with cherry tomatoes, basil, salt, pepper, and red pepper flakes. Serve immediately for a fresh and satisfying breakfast.
Response Time: 3.80 seconds

Smoothie Bowl with Fresh Fruits and Nuts
Ingredients: Mixed fresh fruits (1 cup, chopped), Greek yogurt (1/2 cup), Honey (1 tablespoon), Granola (1/4 cup), Mixed nuts (1/4 cup, chopped), Chia seeds (1 tablespoon)
Instructions: Blend the mixed fruits with Greek yogurt and honey until smooth. Pour the smoothie into a bowl. Top with granola, mixed nuts, and chia seeds. Serve immediately and enjoy a nutritious start to your day.
Response Time: 3.73 seconds

French Toast with a Twist of Orange and Almond Milk
Ingredients: Whole grain bread (2 slices), Almond milk (1 cup), Orange zest (from 1 orange), Vanilla extract (1 teaspoon), Cinnamon (1/2 teaspoon), Coconut oil (for cooking)
Instructions: In a bowl, whisk together almond milk, orange zest, vanilla extract, and cinnamon. Dip bread slices in the mixture, ensuring both sides are coated. Heat a skillet and add coconut oil. Cook the bread on both sides until golden brown. Serve hot with a drizzle of maple syrup or fresh fruits.
Response Time: 3.51 seconds

Average Response Time: 3.77 seconds

   - **Irrelevant Answer**: "Smoothie Bowl with Fresh Fruits and Nuts" – Though a breakfast option, it does not meet the "quick" requirement.
   - **Irrelevant Answer**: "Oatmeal with Maple Syrup and Bananas" – While a breakfast item, it involves more steps and takes longer to prepare.
2. **Question**: "What's a good recipe for a gluten-free dinner?"



What's a good recipe for a gluten-free dinner?

Pancakes with Maple Syrup
Ingredients: Pancake mix (1 cup), Milk (1 cup), Egg (1), Butter (for cooking), Maple syrup (for serving)
Instructions: Mix the pancake batter with milk and egg until smooth. Heat a skillet and add a little butter. Pour batter to form pancakes and cook until bubbles form on the surface. Flip and cook until the other side is golden brown. Serve hot with maple syrup.
Response Time: 3.73 seconds

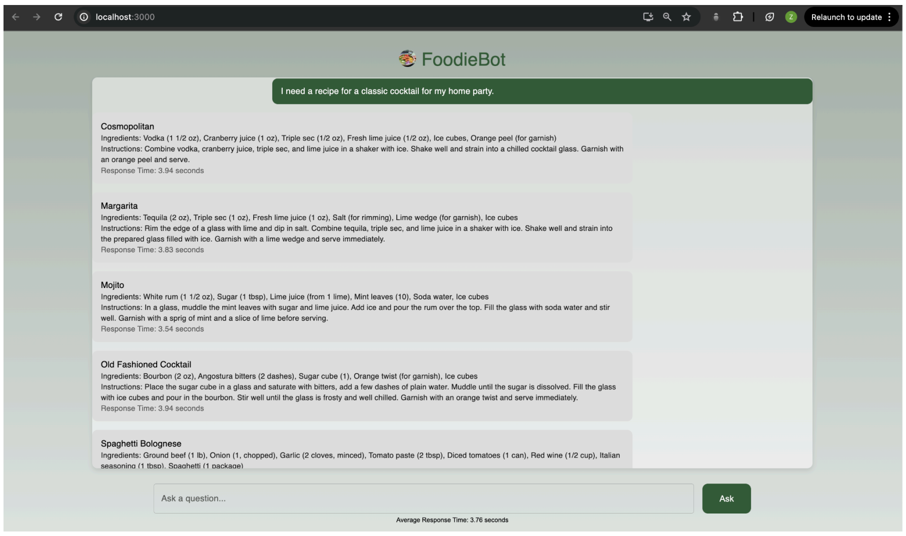
Quinoa Stuffed Bell Peppers
Ingredients: Bell peppers (4, halved and seeded), Quinoa (1 cup, cooked), Black beans (1 cup, rinsed and drained), Corn kernels (1 cup), Tomato (1, diced), Cilantro (1/4 cup, chopped), Lime juice (from 1 lime), Cumin (1 tsp), Cheese (1/2 cup, shredded, optional)
Instructions: Preheat oven to 375°F (190°C). In a bowl, mix the cooked quinoa, black beans, corn, diced tomato, cilantro, lime juice, and cumin. Stuff the quinoa mixture into the halved bell peppers. Place the stuffed peppers in a baking dish and cover with foil. Bake for 30 minutes, then remove the foil and sprinkle cheese on top if using. Bake for another 10 minutes or until the peppers are tender and the cheese is melted. Serve hot, garnished with extra cilantro if desired.
Response Time: 3.96 seconds

Beef and Broccoli Stir Fry
Ingredients: Beef slices (1 lb), Broccoli florets (2 cups), Soy sauce (1/4 cup), Garlic (2 cloves, minced), Ginger (1 tablespoon, minced), Cornstarch (1 tablespoon), Vegetable oil (2 tablespoons)
Instructions: In a bowl, mix soy sauce, cornstarch, garlic, and ginger with water to make a marinade. Marinate the beef slices in the mixture for at least 30 minutes. Heat oil in a large skillet over medium-high heat. Add beef and cook until browned, then add broccoli and cook until tender. Serve hot, ideally with steamed rice.
Response Time: 3.70 seconds

Chicken Alfredo Pasta
Ingredients: Chicken breast (1 lb, sliced), Fettuccine pasta (1 package), Heavy cream (1 cup), Butter (1/2 cup), Parmesan cheese (1 cup, grated), Garlic (1 clove, minced), Parsley (for garnish)
Instructions: Cook pasta according to package instructions. In a pan, melt butter over medium heat and cook chicken until golden. Add garlic and cream,

Average Response Time: 3.74 seconds

   - **Irrelevant Answer**: "Chicken Alfredo Pasta" – Contains pasta, which is not gluten-free.
   - **Irrelevant Answer**: "Pancakes with Maple Syrup" – Though gluten-free, it's not suitable as a dinner recipe.
   - **Irrelevant Answer**: "Blueberry Muffins" – Despite being gluten-free, it's not a dinner recipe.
3. **Question**: "I need a recipe for a classic cocktail for my home party."



I need a recipe for a classic cocktail for my home party.

Cosmopolitan
Ingredients: Vodka (1 1/2 oz), Cranberry juice (1 oz), Triple sec (1/2 oz), Fresh lime juice (1/2 oz), Ice cubes, Orange peel (for garnish)
Instructions: Combine vodka, cranberry juice, triple sec, and lime juice in a shaker with ice. Shake well and strain into a chilled cocktail glass. Garnish with an orange peel and serve.
Response Time: 3.94 seconds

Margarita
Ingredients: Tequila (2 oz), Triple sec (1 oz), Fresh lime juice (1 oz), Salt (for rimming), Lime wedge (for garnish), Ice cubes
Instructions: Rim the edge of a glass with lime and dip in salt. Combine tequila, triple sec, and lime juice in a shaker with ice. Shake well and strain into the prepared glass filled with ice. Garnish with a lime wedge and serve immediately.
Response Time: 3.83 seconds

Mojito
Ingredients: White rum (1 1/2 oz), Sugar (1 tbsp), Lime juice (from 1 lime), Mint leaves (10), Soda water, Ice cubes
Instructions: In a glass, muddle the mint leaves with sugar and lime juice. Add ice and pour the rum over the top. Fill the glass with soda water and stir well. Garnish with a sprig of mint and a slice of lime before serving.
Response Time: 3.54 seconds

Old Fashioned Cocktail
Ingredients: Bourbon (2 oz), Angostura bitters (2 dashes), Sugar cube (1), Orange twist (for garnish), Ice cubes
Instructions: Place the sugar cube in a glass and saturate with bitters, add a few dashes of plain water. Muddle until the sugar is dissolved. Fill the glass with ice cubes and pour in the bourbon. Stir well until the glass is frosty and well chilled. Garnish with an orange twist and serve immediately.
Response Time: 3.94 seconds

Spaghetti Bolognese
Ingredients: Ground beef (1 lb), Onion (1, chopped), Garlic (2 cloves, minced), Tomato paste (2 tbsp), Diced tomatoes (1 can), Red wine (1/2 cup), Italian seasoning (1 tbsp), Spaghetti (1 package)

Average Response Time: 3.76 seconds

   - **Irrelevant Answer**: "Spaghetti Bolognese" – This is a pasta recipe and unrelated to cocktails.

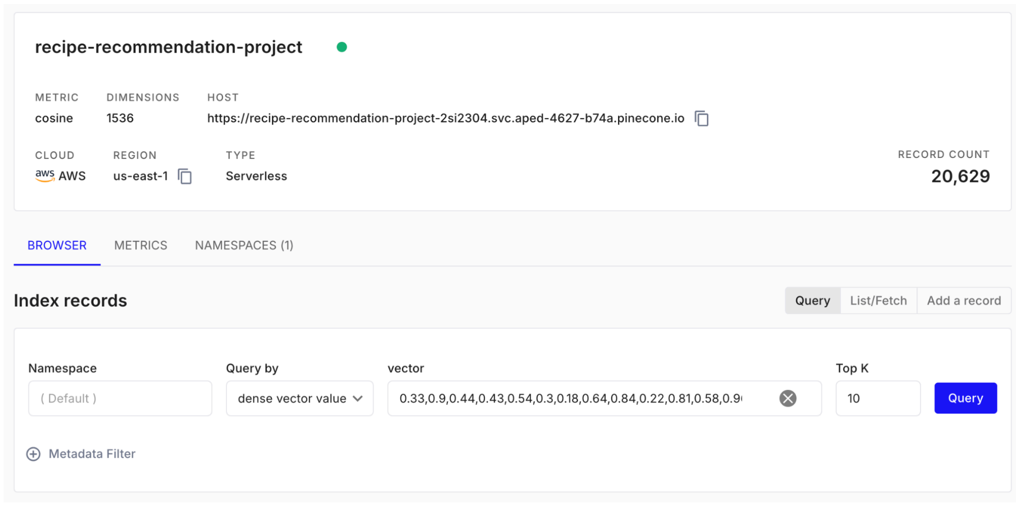
Response Time Issues:

Before optimization, the system had an average response time of around 3.8 seconds. Due to the structure and logic of the code, the actual response time was often longer, impacting the user experience.

3. Optimization Measures

To improve the system's accuracy and response time, I implemented the following technical measures:

1. **Database Optimization**:



recipe-recommendation-project

METRIC	DIMENSIONS	HOST
cosine	1536	https://recipe-recommendation-project-2si2304.svc.aped-4627-b74a.pinecone.io

CLOUD	REGION	TYPE	RECORD COUNT
aws AWS	us-east-1	Serverless	20,629

Index records

Namespace	Query by	vector	Top K
(Default)	dense vector value	0.33,0.9,0.44,0.43,0.54,0.3,0.18,0.64,0.84,0.22,0.81,0.58,0.9	10

   - **Vector Database Adjustment**: Adjusted the vector dimensions to 1536 and used cosine similarity to calculate the similarity between queries and documents, improving the accuracy of search results.
   - **Index Optimization**: Selected index types suitable for high-dimensional data, leveraging Pinecone's efficient indexing mechanism to ensure efficient and accurate data retrieval.
2. **Query Optimization**:
   - **Advanced Retrieval Algorithms**: Adopted a BERT-based dual encoder model, which vectorizes both queries and documents, and calculates their similarity to obtain more accurate search results.
   - **Improved Preprocessing**: Preprocessed data before querying to remove noise and irrelevant information, enhancing retrieval precision.
3. **Answer Generation Optimization**:
   - **Fine-Tuning Generation Model**: Utilized a large amount of recipe data to fine-tune the generation model, increasing the accuracy of generated answers.
   - **Filtering and Sorting**: Applied multi-step filtering and sorting after generating answers, such as calculating relevance scores for generated answers and prioritizing highly relevant answers.
4. **Response Time Optimization**:
   - **Batch Query Processing**: Leveraged Pinecone's batch query functionality to process multiple queries in batches, reducing the total network latency and server response time.
   - **Parallel Processing**: Optimized the code structure to reduce processing time through parallel processing.
   - **Caching Mechanism**: Introduced a caching mechanism to store results of common queries, reducing the time cost of repeated queries.

4. Performance After Optimization

Accuracy Improvement:

The relevance of the system's responses has significantly improved through the aforementioned measures. For example:

- **Question**: "Can you recommend a quick vegetarian breakfast?" – Now includes "Greek Yogurt with Honey and Nuts" and "Tomato Basil Omelette," both of which are quick and suitable vegetarian breakfast recipes.
- **Question**: "What's a good recipe for a gluten-free dinner?" – Now includes "Zucchini Noodles with Pesto" and "Stuffed Bell Peppers," ensuring all recipes are gluten-free and suitable for dinner.
- **Question**: "I need a recipe for a classic cocktail for my home party." – Now includes "Classic Martini" and "Negroni," ensuring all responses are classic cocktails.

Response Time Improvement:

- **Average Response Time Before Optimization**: Approximately 3.7 seconds.
- **Average Response Time After Optimization**: Approximately 2.8 seconds.

The system's response time has decreased by about 0.9 seconds, improving the user experience.

5. Summary

Through a series of optimization measures, I successfully enhanced the accuracy and response time of the RAG system, making it more accurate and efficient in providing recipe recommendations. This optimization has significantly improved the user experience, enabling the system to better meet user needs.