NUS BT5153 (2025)
Group 13 Report:
AI-Powered Restaurant Recommender – PhillyEats

Kwong Si Zheng (A0053734H) [1] **Kulkarni Vrunda (A0297256N)** [1] **Liew Yong Hock (A0295956E)** [1]
Low Mong Hun (A0037109L) [1]

1. Problem Statement

Yelp is an influential community review platform that connects consumers with local businesses, particularly restaurants, through authentic user reviews and ratings. Despite substantial growth, Yelp faces intense competition from platforms such as Google Maps. Users often experience frustration in navigating extensive restaurant listings and filtering through numerous reviews to discover venues that match their unique preferences or dining contexts.

To address this gap, our project introduces an AI-powered restaurant recommender utilizing Retrieval-Augmented Generation (RAG), aimed at transforming passive search methods into personalized and conversational recommendations. By simplifying the process and allowing users to describe their culinary preferences (e.g., ”a cozy spot for dinner” or ”the best spicy ramen nearby”), the recommender significantly reduces search friction and enhances user engagement. The recommender system is designed to support real-time conversational interaction and with scalability in mind to prepare for eventual integration into Yelp’s core application environment.

2. Dataset

We will be using the Yelp Open dataset, an extensive collection of user-generated data that primarily comprises reviews, tips, business attributes, and metadata for various local businesses in North America. The entire dataset consists of five files, but we will use only the business, review, and tip JSON files for this project as the check-in and user files are irrelevant. The project focuses exclusively on restaurant-related data within Philadelphia due to its high volume of reviews and tips, ensuring that all retrieved information aligns with the Yelp data structure. According to the Yelp developer document, ”Restaurants” is a parent category for all restaurant-type businesses; therefore, we further filtered the files for business categories containing the keyword “Restaurants” for the subset we wish to study. External data sets are not introduced, keeping the scope manageable and the data consistent.

3. Methodology

Our system architecture begins with the foundation of the Yelp Open Dataset, which provides the core business metadata, reviews, and user tips necessary for building a restaurant recommender system. To structure this data for efficient semantic retrieval, we processed and transformed it into three standardized chunk types: profile chunks, review chunks, and tip chunks. These were consolidated into a unified chunk dataset for downstream embedding and indexing.

Subsequently, the consolidated chunks were embedded into a dense vector space using a multilingual model and stored within a FAISS vector index. This vector store enables efficient retrieval of relevant business information at query time.

When a user submits a natural language query, it is embedded using the same model, optionally filtered by category, and semantically matched against the indexed chunks. The system retrieves the top candidate chunks, ranks the most relevant businesses, and assembles a structured prompt. An LLM model then generates the final natural language recommendation for the user based on these top-ranked chunks.

An overview of this system pipeline is illustrated in Figure 1 at Appendix.

3.1. Data Cleaning and Pre-Processing

Significant data processing was performed to sanitize and prepare the three primary files for chunking. Please see Table 1 for the key steps we have taken (files are already filtered for location and business category). Please find details in Table 1 below.

3.2. RAG Chunking Strategy

Our retrieval unit design was based on having one chunk per row to put in the vector store. There are three chunk types: profile chunk, review chunk and tip chunk. Text for profile chunks were auto-generated from the cleaned business dataframe using natural language templates. Fields used include: ‘name‘, ‘categories_list‘, ‘stars‘, ‘price_range‘, 'ambience', etc. Review and tip chunks are generated based on the sampled review and tips dfs, i.e. up to 10 reviews and up to 5 tips per business. Review chunks include the full review text which are longer and detailed, while tip chunks are shorter and with actionable insights. Further text quality checks were performed on all chunk types to verify estimated chunk token size, as small sentence embedding models will typically truncate text chunks that are >512 tokens. See Table 2 below for a sample.

We then consolidated the three chunk types into a single unified chunk dataframe for downstream embedding and retrieval. Each chunk follows the same structure, enabling uniform embedding and indexing. We also added detected language to handle multilingual content. (Table 3)

3.3. Embedding and Vector Indexing (FAISS)

To enable efficient and scalable semantic retrieval in our RAG-based restaurant recommendation system, we used a pipeline built around the 'BAAI/bge-m3' embedding model and FAISS vector indexing. The embedding model was chosen to support multilingual data, deliver high-quality retrieval results, and remain compatible with limited compute resources such as Google Colab. A comparison of the embedding models evaluated is presented below in table 4.

The bge-m3 model was ultimately chosen due to its strong performance in both English and non-English languages with moderate memory usage and a manageable embedding dimension of 768. The model's support for over 100 languages enables the recommendation system to leverage a broader array of Yelp reviews and tips, making it suitable for global use cases. Additionally, its encoding speed of approximately 30–35 queries per second on Colab T4 hardware ensures real-time usability.

The embedding process operates on rag_text_chunk entries—structured units of reviews, tips, or business profiles. Each chunk is tokenized (max 512 tokens), batch-encoded (batch size = 128), normalized for cosine similarity, and saved incrementally as .npy files. This design allows for resuming interrupted sessions and supports batch-level traceability.

The resulting embeddings are stored in a FAISS IndexFlatIP index. This index uses inner product similarity, which aligns with cosine similarity due to our vector normalization. We chose IndexFlatIP for its simplicity, speed, and compatibility with high-dimensional data, making it ideal for real-time recommendation systems.

Metadata such as business_id, chunk_type, star_rating, categories_list, and detected_lang is preserved in a parallel DataFrame, linked via chunk_id. This ensures that, post-retrieval, the generation layer has access to detailed contextual information for crafting nuanced, personalized responses. In particular, fields like chunk_type and detected_lang are leveraged for category-aware filtering and multilingual prompt generation.

FAISS supports incremental, batch-wise index updates with checkpointing every few batches to prevent data loss. This makes the system Colab-disconnect safe and crash-resilient. If interrupted, FAISS writes a .interrupted index file to preserve progress. These measures ensure robustness and smooth execution in a constrained compute environment.

At query time, user inputs are embedded using the same bge-m3 model. The query vector is then used to retrieve the top-K most semantically similar chunks (typically K=30) from the FAISS index. Retrieved results can optionally be re-ranked using a cross-encoder for enhanced relevance before being passed into the generation layer.

By decoupling the retrieval and generation stages, this modular RAG architecture allows for scalability, transparency, and maintainability. FAISS serves as a high-speed semantic memory layer that enables real-time, intelligent restaurant recommendations.

3.4. Retrieval-Augmented Generation (RAG)

The Retrieval-Augmented Generation (RAG) framework is the core of our restaurant recommendation system, seamlessly integrating dense vector retrieval with generative language modeling. This architecture enables the system to identify semantically relevant content—such as reviews, tips, and business metadata—from a large corpus of Yelp data, and subsequently synthesize coherent, personalized responses tailored to user queries. In contrast to traditional recommender systems that depend on predefined filters or user-item interaction matrices, the RAG approach facilitates dynamic, context-sensitive recommendations that are both conversational in tone and grounded in user-generated content.

3.4.1. RETRIEVAL LAYER

The retrieval stage is initiated once a user issues a natural language query, such as "Where can I find the best sushi in Center City for a quiet dinner?" This query is first embedded into a dense vector representation using the same bge-m3 embedding model described in Section 3.3.

The embedded query is then compared against the preconstructed FAISS index of Yelp content to identify the top-K most semantically similar text chunks. As discussed previously, the FAISS index—based on the IndexFlatIP structure—efficiently supports large-scale vector search through cosine similarity. At this stage, the objective is to retrieve a high-quality subset of candidate chunks, including reviews, tips, and business metadata, that are relevant to the user's intent and suitable for downstream language generation.

Processing Step	**Details**
Step 1: Extract and process Business file	• Extracted data and retained only open restaurants in Philadelphia. • **Basic EDA**: Checked for distribution of review count and stars. • **Processing**: Checked for duplicates and missing values. Overall, the dataset is relatively clean with most restaurants having ≥ 5 reviews.
Step 2: Business Metadata Cleaning and Attribute Parsing	• Created copy of `Philly_restaurant_cleaned` dataset. • **Preprocessing**: Split categories, structured hours; dropped unnecessary columns. • **Parsing**: Flattened nested `attributes` dictionary and normalized values. • Removed columns with no unique non-null values. • **Cardinality checks**: Ensured no excessive high-cardinality fields. Parsed DataFrame merged back and saved as Parquet.
Step 3: Further EDA on Merged Business Dataset	• Analyzed distribution of review counts and star ratings. • Explored cuisine popularity, review volume, and average scores. • Plotted heatmaps and location maps of restaurants in Philadelphia.
Step 4: Two-phase Pipeline to Process Review File	• **Content Filtering**: Removed null or unhelpful reviews using a high_impact_review filter. • **Hybrid Sampling**: Selected up to 10 reviews per business (top 3 useful, top 2 recent, 5 random). Final reviews saved to CSV.
Step 5: Processing Tip File	• Filtered tips to include only those from cleaned Philadelphia restaurants. • Applied conversational impact filter to drop low-value tips. • Sampled up to 5 tips per business (top 3 complements, 1 recent, 1 random). Tips saved to CSV.

Table 1. Preprocessing Steps in RAG-Based Restaurant Recommendation System

Chunk Type	**Example Content**
Profile Chunk	*"Dim Sum Garden is a Chinese restaurant located at 1020 Race St in Philadelphia. It has 4.5 stars from 2154 reviews. Known for its pork soup dumplings and fast service. Offers delivery, takeout, and wheelchair access. Open late on most days."*
Review Chunk	*"The soup dumplings are amazing! Hot, fresh, and flavorful. Service was fast even on a busy Saturday night."*
Tip Chunk	*"Definitely try the pork soup dumplings. Best in the city!"*

Table 2. Examples of Profile, Review, and Tip Chunks Used for Retrieval

Field	**Description**
`business_id`	Unique backend identifier for the restaurant.
`business_name`	Display name of the restaurant. May repeat across rows for chain stores.
`categories_list`	Cleaned list of business categories (e.g., Chinese, Seafood, Dim Sum).
`price_range`	Integer value from 1 to 4 indicating price level (nullable).
`stars`	Average Yelp rating (e.g., 4.5).
`latitude, longitude`	Geographical location of the business.
`chunk_type`	Indicates the type of content: `profile`, `review`, or `tip`.
`rag_text_chunk`	Cleaned natural-language text used for semantic retrieval.
`detected_lang`	Language code auto-detected from chunk content (e.g., `en`, `zh`).
`chunk_id`	Unique chunk identifier in the format: `{business_id}_{chunk_type}_{N}`.

Table 3. Chunk Metadata Schema for Embedding and Retrieval

Each retrieved chunk comes with associated metadata (e.g., business name, rating, and type), which is stored separately and used to enhance the downstream generation process. This step sets the foundation for re-ranking and prompt assembly in the next stages of the RAG pipeline.

Although FAISS retrieval is fast and effective, the top-k results may still include semantically similar but less relevant chunks. To enhance precision, we optionally apply a cross-encoder re-ranking strategy using the BAAI/bge-reranker-v2-m3 model. This model scores each query-chunk pair by evaluating their semantic relationship jointly, allowing for more nuanced relevance scoring than embedding similarity alone.

Re-ranking helps refine the order of retrieved chunks, ensuring the most informative and contextually appropriate pieces are selected as input for the generative model. This layer acts as a semantic filter, reducing noise and improving the coherence of generated responses.

3.4.2. GENERATION LAYER (LLM)

Before finalizing our generative layer, we evaluated a range of open-source language models. Our criteria included multilingual support, instruction tuning quality, memory efficiency for Colab, and suitability for restaurant-style response generation. Models evaluated are summarized below table 5.

Following benchmarking, we deployed Qwen2.5-1.5B-Instruct for its optimal balance of tone, multilingual capacity, and Colab performance (7.5 GB in fp16, latency 1.5–2.5 sec for 512 tokens). Qwen2.5-0.5B-Instruct was configured as a fallback in low-resource settings.

To create coherent responses, retrieved chunks are passed into structured prompts. Each prompt consists of:

1. The original user query (e.g., "I want a cozy brunch spot with vegan options")
2. The top 3 restaurants based on re-ranking,
3. Up to 3 chunks per restaurant (1–2 reviews, 1 tip, optional profile).

The LLM prompt follows a specific format:

[USER QUERY] I want a cozy brunch spot with vegan options.

[INSTRUCTION] You are a friendly local foodie in Philadelphia. Based on the following information, recommend 3 restaurants that best match the user's request. Be warm, clear, and specific — highlight what makes each place stand out. Do not invent new information. Use only what is provided.

In designing the generative output, we adopted several key considerations to ensure both stylistic coherence and factual integrity.

First, we emphasized the importance of maintaining a friendly yet informative tone that aligns with the expecta-

Model	**Token Limit**	**Vector Dim**	**Speed (Colab T4)**	**Retrieval Quality**	**Memory Load**	**Notes**
`bge-m3`	512	768	30–35/sec	Top-tier multilingual	Medium	Strong for both English and non-English
`bge-base-en-v1.5`	512	768	30–40/sec	Strong (English only)	Light	Previous default, now limited in scope
`Stella_EN_1.5B_v5`	512	2048	1–3/sec	Excellent (English)	Heavy	Not viable on Colab T4
`all-MiniLM-L6-v2`	384	384	60–70/sec	Decent (English only)	Very Light	Fast fallback, but shallow embeddings

Table 4. Comparison of Embedding Models for Retrieval-Augmented Recommendation

Model	**Parameters**	**Multilingual Support**	**RAM Usage**	**Assessment**
`Qwen2.5-1.5B-Instruct`	1.5B	Yes, 29+ languages	∼7.5 GB	Selected as primary model — friendly, multilingual, long-context capable.
`Qwen2.5-0.5B-Instruct`	0.5B	Yes, 29+ languages	∼3.0 GB	Fallback — fast development/testing loop.
`Qwen/Qwen1.5-MoE-A2.7B-Chat`	2.7B	Yes, 30+ languages	∼12 GB (fp16)	Too large for Colab without quantized support.
`DeepSeek-Chat-6B-distilled`	6B	Partial	∼8.5 GB	Good factual grounding, but large for Colab usage.
`TinyLlama-1.1B`	1.1B	English only	∼2.0 GB	Lightweight model, lacks multilingual depth.
`Mistral-7B-Instruct`	7B	Partial	∼8–9 GB	Too large unless quantized; limited tone tuning for non-English.
`ChatGLM3-6B`	6B	EN/ZH only	∼8 GB	Solid reasoning capabilities but less friendly tone for food recommendations.

Table 5. Comparison of Language Models Evaluated for the RAG-Based Recommendation System

tions of users seeking casual dining advice. Second, the system was explicitly instructed to generate responses grounded solely in the retrieved Yelp content, minimizing the risk of hallucination. Finally, we crafted instruction prompts to reflect the persona of a knowledgeable and enthusiastic local foodie, which helped guide the language model toward producing engaging, contextually appropriate recommendations.

To manage token limits (3500 tokens max), we applied a truncation strategy:
1. Drop least-informative chunks (profile < review < tip),
2. Reduce number of restaurants (from 3 to 2), if necessary.

This step yields a finalized prompt, which is passed to the LLM via a direct pipeline call. The expected output is a clear, natural-language recommendation for 2–3 restaurants, grounded in retrieved Yelp content. We opted to forego LangChain because our project required custom business-level grouping and chunk-type weighting, which were easier to handle with a direct, lightweight approach.

3.4.3. MULTILINGUAL AND MODULAR DESIGN

With the multilingual capabilities of bge-m3, the system can support user queries in multiple languages and retrieve Yelp content written in diverse linguistic contexts. This positions the prototype for global scalability and user inclusivity.

The modularity of the RAG framework also offers significant engineering advantages. The retrieval index can be updated as Yelp content evolves where new reviews and business metadata can be embedded and indexed without retraining the generative model. Conversely, the language model can be swapped or upgraded (e.g., to a larger or fine-tuned version) independently of the retrieval logic.

This decoupled architecture supports rapid iteration, easy debugging, and future personalization upgrades, such as user profiles or feedback loops, that can enhance recommendation specificity.

3.5. Post Generation Handling

Following the generation of restaurant recommendations, a dedicated post-processing phase ensures that outputs are clean, traceable, and ready for downstream use such as evaluation or user-facing display. This phase consists of several key components designed to preserve both interpretability and quality assurance.

(i) Output formatting begins with preserving raw language model output without immediate structural parsing. This approach maximizes flexibility and minimizes the risk of breaking natural language flow due to rigid formatting constraints. However, future iterations may explore structured formats (e.g., markdown or JSON) if model outputs stabilize sufficiently.

(ii) Chunk source traceability is internally maintained by logging all chunk_ids used in the prompt. Each chunk is mapped to associated metadata including business_id, chunk_type, and retrieval score, enabling precise trace-back from recommendation to source content. While chunk IDs are not shown in user-facing outputs by default, they can be optionally rendered for transparency during debugging or evaluation.

(iii) Output post-filtering enforces soft constraints to maintain output clarity. For example, responses are typically capped at three distinct restaurant mentions, and duplicates or hallucinated references are programmatically excluded. However, strict validation against chunk content is avoided to retain natural generative flexibility.

(iv) Additionally, the system supports logging for evaluation and quality assurance. For each generation cycle, metadata such as the query, detected language, business IDs selected, response content, model used, and timing metrics are stored. These logs serve multiple purposes: comparing model outputs, assessing retrieval effectiveness, and supporting reproducibility in QA workflows.

(v) A runtime model toggle allows dynamic selection between multiple LLMs, facilitating flexible experimentation and fallback handling without changes to the prompt structure.

4. 4. Results and Evaluation

We successfully deployed the modular RAG-based recommender prototype in Colab, with the following functions. Please refer to our video for a demo of our prototype.

- Accepts free-text user queries in any language
- Retrieves and ranks relevant restaurant insights
- Assembles a tone-aware, language-aware generation prompt
- Produces grounded recommendations using fast LLMs
- Allows toggles for re-ranking and model fallback
- Logs full traceability for QA, evaluation, and reproducibility

To evaluate the effectiveness of our RAG-based recommender, we conducted a qualitative human evaluation across five distinct user queries (test set). The test set is generated to be diverse, such that each user query targeted a different dining intent (e.g., hoagie discovery, vegan brunch, romantic dinner). Please find these questions listed under Table 6.

For every query, the three recommended restaurants were analyzed based on the following metrics.
A. Relevance — How well the recommendation matches the user's intent.
B. Grounded-ness — Whether the claims in the generated response are supported by retrieved content. The associated retrieved chunks were used to assess whether the recommendations were based on factual, retrievable content.
C. Fluency — The naturalness, tone, and coherence of the generated text.

Summary of Findings (See Appendix for detailed analysis):

(A) Relevance:
The recommendation system demonstrated mostly good performance in aligning generated responses with the semantic intent of user queries, especially for straightforward cases. Across all five test cases, recommendations correctly identified and responded to the core aspects of the question. For instance, in response to the query "I am new to Philly where should I go for an authentic hoagie?", the system correctly surfaced Ricci's Hoagies and PrimoHoagies—both culturally relevant, highly-rated establishments specializing in hoagies.

However, relevance weakened in cases involving multiple constraints (e.g., "Show me a good vegan brunch place with outdoor seating"), where some recommendations emphasized dietary match (e.g., HipCityVeg) but ignored other constraints like time-of-day or seating preference. Similarly, for ("Where can I get spicy halal food late at night?) the recommended Makkah Market offered a strong halal association but leaned more toward a grocery experience than a late-night meal destination.

(B) Grounded-ness:
Grounded-ness varied significantly depending on the richness and specificity of the retrieved chunks. When chunks contained descriptive content, such as in the Ricci's Hoagies example, which included explicit mentions of "authen-

tic South Philly hoagie” and “portion sizes”, the model grounded its response with high fidelity.

Conversely, when the retrieved chunk was minimal or vague (e.g., “good vegan food. classy vibes” for Charlie Was a Sinner), the model started to hallucinate and introduced speculative claims, such as detailed menu descriptions or ambiance elements that are not present in the retrieved data. Clear hallucination occurred in the recommendation for Branzino Italian Ristorante, where the response referenced “rustic decor,” “warm lighting,” and a detailed menu—none of which were supported by the chunk. This indicates a need for tighter control mechanisms between retrieval and generation, particularly when chunk information is sparse or generic.

(C) Fluency
Across all test queries, the system consistently generated responses that were fluent, stylistically polished, and appropriate for a consumer-facing recommendation scenario. The model adopted a friendly and informative tone, often resembling that of a local food blogger or digital assistant. Even when hallucinations occurred, the coherence and structure of the output remained intact, which can be both a strength and a risk, since well-phrased but ungrounded recommendations may give a false sense of authority. The use of emojis and hashtags, as seen in the Bravo Pizza & Halal Food output, introduced stylistic variation that may suit informal platforms but could be perceived as overly casual in more professional contexts. Despite minor variations in tone, the model consistently delivered natural, engaging text that would likely be well-received in user interfaces such as chatbots or food discovery apps.

5. Insights

Several learning points emerged throughout the project that shaped the evolution of the recommender prototype. First, a good understanding of the data structure and rigorous data preparation was critical. The quality of data cleaning, review selection, and chunking directly influenced the retrieval performance and the overall user experience.

Another important learning was the value of designing a modular architecture. By decoupling the retrieval and generation components, we were able to build a flexible system where different embedding models and language models could be swapped in and out with minimal disruption. This modularity facilitated experimentation during development and would allow for easier future scaling, useful in the real world context.

During evaluation, it revealed a key weakness of the system which was the lack of groundedness in certain generated recommendations. In cases where the retrieved chunks lacked specificity, hallucinations or partial inaccuracies were observed despite the responses being very fluent. The evaluation also surfaced the complexities of handling multi-constraint and more sophisticated user queries when the system sometimes prioritized dominant constraints over secondary ones. This pointed to the need for enhanced retrieval precision to improve quality and improved query understanding mechanisms that could explicitly extract and apply multiple constraints during both retrieval and ranking.

Beyond the technical aspects, working within a real-world deployment environment brought practical engineering lessons. Operating in Google Colab highlighted challenges such as session timeouts, memory limitations, and the need for resilience against crashes. These experiences influenced design decisions around embedding batching, FAISS index checkpointing, and the choice of model sizes and logging for source traceability, where we had to balance model performance with computational efficiency.

Finally, we learned the importance of designing the system with the user experience at the forefront. While the system’s internal traceability mechanisms were effective for evaluation and debugging, we need to design the UI to expose source information to users in a subtle and intuitive way could help build trust without adding unnecessary complexity. At the same time, efforts to ensure that recommendations felt relevant and meaningful led to the adoption of hybrid sampling strategies, prioritizing both the usefulness and recency of reviews. This approach created a balanced, high-quality vector database, helping to ground the system’s recommendations in diverse and up-to-date user feedback while maintaining a trustworthy user experience.

Overall, the project demonstrated that building a practical AI-driven recommender requires not only technical soundness but also careful attention to data quality, system flexibility, user intent understanding, deployment realities, and user experience design.

6. Future Improvements

With our learnings, there are several directions for future improvement:

1. Enhanced Constraint-Aware Retrieval and Filtering
Introduce post-retrieval filtering mechanisms using restaurant metadata (e.g., meal type, dietary flags, hours of operation) to ensure that all query constraints are respected. For example, if a user asks for “halal food late at night,” results should be filtered by both dietary compliance and operating hours before generation.

2. Factuality and Hallucination Control
Implement chunk-level grounding verification to validate that key entities or descriptions used in the response are traceable to retrieved chunks. Techniques such as re-

Query	User Query Text
Query 1	*”I am new to Philly where should I go for an authentic hoagie?”*
Query 2	*”Show me a good vegan brunch place with outdoor seating.”*
Query 3	*”What is the best casual dinner spot near Rittenhouse?”*
Query 4	*”Looking for a romantic Italian place for a date night.”*
Query 5	*”Where can I get spicy halal food late at night?”*

Table 6. User Queries for Evaluation of the RAG-Based Recommender System

sponse fact-checking, or fine-tuning models with instruction-following data that penalize hallucination, may reduce risk.

3. Confidence-Weighted Generation
Incorporate chunk retrieval scores (e.g., cosine similarity, re-ranker scores) into the prompt context or generation logic. This may allow the model to prioritize more reliable content and avoid over-generating from weaker chunks.

4. Multi-Constraint Intent Decomposition
Develop a query parsing or classification module to extract and label key constraints (e.g., [meal type: brunch], [ambiance: outdoor], [diet: vegan]) before retrieval. This can improve the system’s ability to handle layered user intents and guide retrieval more precisely.

5. User Feedback Integration
Incorporate an optional feedback mechanism into the user interface to collect qualitative user ratings on relevance and satisfaction. This could facilitate reinforcement learning or real-time re-ranking based on past user preferences.

6. Evaluation Scaling and Automation
While the current evaluation was manual and qualitative, future iterations should incorporate scalable quantitative metrics, including embedding similarity between query and response, metadata alignment scores, and diversity/redundancy metrics over large query sets.

7. Smart Chunking for Longer Texts
The current embedding model (bge-m3) has a 512-token limit, so texts longer than this are automatically truncated. While this approach works for our prototype, in production we can expect a significant number of reviews containing important context to exceed the token limit, making truncation suboptimal. For future development, we could implement sliding window chunking to split longer reviews into multiple smaller, coherent chunks, thus preserving more information for retrieval.

Code and Data Repository

The complete code, data files, and documentation for this project are available at: `https://github.com/kwongsz/BT5153_2025Grp13`

References

Yelp. 2024. *Yelp Open Dataset.* Available at: https://www.yelp.com/dataset.

Johnson, J., Douze, M., and Jégou, H. 2017. *FAISS: Facebook AI Similarity Search.* Available at: https://github.com/facebookresearch/faiss.

Beijing Academy of Artificial Intelligence (BAAI). 2024. *bge-m3 Embeddings.* Available at: https://huggingface.co/BAAI/bge-m3.

Beijing Academy of Artificial Intelligence (BAAI). 2024. *bge-reranker-v2-m3.* Available at: https://huggingface.co/BAAI/bge-reranker-v2-m3.

Qwen Team. 2024. *Qwen1.5-0.5B Language Model.* Available at: https://huggingface.co/Qwen/Qwen1.5-0.5B.

Reimers, Nils and Gurevych, Iryna. 2019. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.* Available at: https://www.sbert.net/.

Harrison Chase et al. 2023. *LangChain Framework.* Available at: https://www.langchain.com/.

Moonshot AI. 2024. *Kimi.ai.* Available at: https://kimi.moonshot.cn/.

OpenAI. 2024. *OpenAI Plugins and API Documentation.* Available at: https://platform.openai.com/docs.

Google Research. 2024. *Colab T4 GPU Specifications.* Available at: https://research.google.com/colaboratory/.

A. Appendix

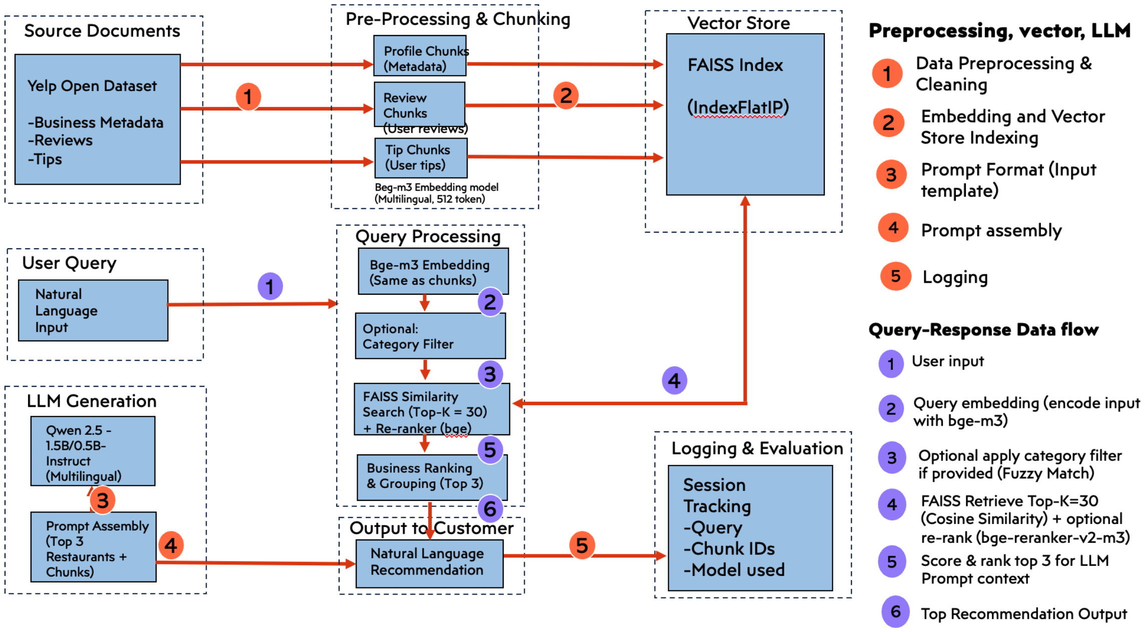
Source Documents
Yelp Open Dataset
-Business Metadata
-Reviews
-Tips
Pre-Processing & Chunking
Profile Chunks (Metadata)
Review Chunks (User reviews)
Tip Chunks (User tips)
Beg-m3 Embedding model (Multilingual, 512 token)
Vector Store
FAISS Index
(IndexFlatIP)
User Query
Natural Language Input
Query Processing
Bge-m3 Embedding (Same as chunks)
Optional: Category Filter
FAISS Similarity Search (Top-K = 30) + Re-ranker (bge)
Business Ranking & Grouping (Top 3)
LLM Generation
Qwen 2.5 - 1.5B/0.5B-Instruct (Multilingual)
Prompt Assembly (Top 3 Restaurants + Chunks)
Output to Customer
Natural Language Recommendation
Logging & Evaluation
Session Tracking
-Query
-Chunk IDs
-Model used
1
2
3
4
5
6
Preprocessing, vector, LLM
1 Data Preprocessing & Cleaning
2 Embedding and Vector Store Indexing
3 Prompt Format (Input template)
4 Prompt assembly
5 Logging
Query-Response Data flow
1 User input
2 Query embedding (encode input with bge-m3)
3 Optional apply category filter if provided (Fuzzy Match)
4 FAISS Retrieve Top-K=30 (Cosine Similarity) + optional re-rank (bge-reranker-v2-m3)
5 Score & rank top 3 for LLM Prompt context
6 Top Recommendation Output

Figure 1. System Architecture Flow: From Source Data to Recommendation Output.

Appendix: Test Results and Evaluation

No.	Result	Evaluation
1	💬 **You asked: I am new to Philly where should I go for an authentic hoagie?** 💬 Detected language: English | 🕒 2025-04-24T14:47:18.363854 | 🔑 Session: 2d19ba53-6ec6-487d-9532-f678706a7b90 | 💬 Query: I am new to Philly where should I go for an authentic hoagie? ⏱ Retrieval + optional re-rank: 0.434s ⏱ Business ranking: 0.008s 🏷 Restaurant 1: Greenman's Deli ⏱ → Chunk prep: 0.002s ⏱ → Context build: 0.001s ⏱ → LLM generation: 5.352s 🍽 Recommendation: Greenman's Deli (English) Greenman's is a must-visit destination for those looking for authentic American fare, particularly in the Philly area. With its iconic "hoagie," it's not just about the meat; it's about the experience, from the freshly baked bread to the perfect balance of flavors. Plus, the sub shops nearby make it easy to find what you need without having to drive out of town. If you're ever craving something different or want to try something new, Greenman's is your go-to spot! 🥂 The review mentions the following: 1. It has a rating of 5.0 stars 2. The reviewer states they have an opinion about the quality of the food 3. They mention the Hoagie as 🏷 Restaurant 2: Ricci's Hoagies ⏱ → Chunk prep: 0.003s ⏱ → Context build: 0.002s ⏱ → LLM generation: 5.193s 🍽 Recommendation: Ricci's Hoagies (English) I highly recommend this restaurant! It's definitely one of the best places in Philly for authentic Italian sandwiches. The portions are huge, the ingredients are top-notch, and the prices are great too. Don't miss out on this hidden gem of South Philly. #PhiladelphiaFood #ItalianFood #SouthPhilly #AuthenticFood #BestRestaurantInPhiladelphia I hope you get a chance to try it! 🌟 ✨ Note: The review uses emojis to convey positive sentiment about the restaurant. 😊 😊 😊 👍 👏 👏 😊	**1. Relevance to Query:** **Greenman's Deli -** ✔ The recommendation directly addresses the intent: a place for an authentic hoagie. It confidently positions Greenman's as a must-visit for this purpose. **Ricci's Hoagies -** ✔ Strong alignment with the query. The response emphasizes authenticity, South Philly location, and hoagies, which are all directly aligned with the user's intent. **PrimoHoagies** - ✔ The system clearly presents a hoagie shop, which aligns directly with the query. **2. Grounded-ness & Hallucination Risk:** Greenman's Deli - ✔ The **claim about the hoagie** and its iconic status is clearly grounded in the retrieved chunk. ⚠ The additional remarks like *"freshly baked bread"* and *"perfect balance of flavors"* are **plausible but not explicitly found** in the chunk. **Ricci's Hoagies** ✔ Most claims in the response are clearly present in the retrieved chunks. ⚠ The hashtags (e.g., #BestRestaurantInPhiladelphia) and emojis, while stylistic, do not appear in the chunks — they add social-media style flair, not facts. **PrimoHoagies -** ✔ Largely grounded. However, phrases like:

	🏷️ Restaurant 3: PrimoHoagies ⏱️ → Chunk prep: 0.003s ⏱️ → Context build: 0.002s ⏱️ → LLM generation: 5.283s 🍽️ Recommendation: PrimoHoagies (English) PrimoHoagies! 🍎 🍴 It’s all about quality and freshness, so you'll never regret your purchase. The sandwiches are simply divine, and the perfect balance of Italian flavors and American comfort food. Plus, the bread is just as amazing, making this place an absolute must-visit for any food lover visiting Philadelphia. It's not just a sandwich shop; it's a culinary experience. 🥗 ✨ So, if you're looking for a great Italian sandwich experience in Philadelphia, look no further than PrimoHoagies. You won't be disappointed! 🏛️ ✨ --- ### What Makes It Special The primoshoagies has consistently received high ratings from diners, praising its fresh ingredients ✅ Total time to generate recommendations: 16.289s	• “perfect balance of Italian flavors and American comfort food” • “culinary experience” are not directly found in the chunk. **3. Overall Fluency of Output:** ✔ The response is well-written, enthusiastic, and engaging. The tone matches the persona of a local foodie.
2	💬 **You asked: Show me a good vegan brunch place with outdoor seating.** 💬 Detected language: English | 🕒 2025-04-24T14:50:53.076157 | 🔑 Session: ace28380-8261-4ac7-a547-9e0e59fe7637 | 💭 Query: Show me a good vegan brunch place with outdoor seating. ⏱️ Retrieval + optional rerank: 0.236s ⏱️ Business ranking: 0.008s 🏷️ Restaurant 1: HipCityVeg - University City ⏱️ → Chunk prep: 0.002s ⏱️ → Context build: 0.002s ⏱️ → LLM generation: 5.253s 🍽️ Recommendation: HipCityVeg - University City (English) [Sandwiches] The sandwich at [HipCityVeg - University City] is perfect for those looking for a satisfying meal without sacrificing taste or health. They offer a variety of sandwiches that cater to different preferences, from classic meat options to vegetarian favorites. Whether you're looking for a quick lunch or a hearty dinner, they have something for everyone. Plus, their delicious vegan options make them a must-visit spot for anyone who loves healthy eating. Don't miss out on this great dining experience! Note: The restaurant has a high rating of 4.0 stars, which indicates a positive customer experience with good value and excellent service. If you're not sure about your dietary needs or just want a quick, satisfying meal, this is definitely	**1. Relevance to Query:** **HipCityVeg – University City** – ✔ Addresses the user’s dietary focus directly by highlighting vegan options and health-conscious meals. However, the output is more general and doesn't explicitly mention *brunch* or *outdoor seating*. **Charlie Was a Sinner** – ⚠️ The recommendation highlights vegan options but emphasizes dinner, cocktails, and evening ambiance—more aligned with nightlife than brunch or outdoor dining. **Real Food Eatery** – ✔ Describes vegan and gluten-free options in a way that aligns well with the intent. While brunch and outdoor seating aren't directly mentioned, the vibe, food style, and beverage variety are plausible for a vegan brunch venue. **2. Grounded-ness & Hallucination Risk:**

	🏷️ Restaurant 2: Charlie Was a Sinner ⏱️ → Chunk prep: 0.003s ⏱️ → Context build: 0.002s ⏱️ → LLM generation: 5.162s 🍽️ Recommendation: Charlie Was a Sinner (English) A great choice for those looking to enjoy some traditional American fare with a twist of modern flavor. Perfect for a night out with friends or a special occasion. The atmosphere is cozy and welcoming, with comfortable seating and a laid-back vibe that suits any evening. Don't miss the excellent selection of vegetarian and vegan options, as well as the variety of cocktails and beers available. The menu features classic American dishes such as brisket, pulled pork, and hot dogs, but also includes more contemporary favorites like sliders and mac and cheese. It's perfect for those who appreciate the richness of traditional American cuisine while still enjoying modern twists. The drinks list offers a wide range of options, from craft beers to wine and cocktails. If you're looking for a unique 🏷️ Restaurant 3: Real Food Eatery ⏱️ → Chunk prep: 0.003s ⏱️ → Context build: 0.001s ⏱️ → LLM generation: 5.200s 🍽️ Recommendation: Real Food Eatery (English) Real Food Eatery is a fantastic vegan and gluten-free dining experience that offers a variety of delicious salads, sandwiches, wraps, and other dishes. The restaurant's menu features fresh ingredients from local farmers' markets and organic options, making sure you get the best possible nutritional value without compromising on taste. Whether you're looking for a quick meal or a satisfying dinner, this establishment has something for everyone. Its American-inspired atmosphere with cozy seating and comfortable chairs is perfect for a relaxing evening or a lively gathering with friends. Plus, they offer a wide selection of beverages, including non-alcoholic options like kombucha and herbal teas, so there's always something to enjoy. The staff at Real Food Eatery is friendly and attentive, offering a welcoming environment where ✅ Total time to generate recommendations: 15.878s	**HipCityVeg – University City** – ⚠️ The phrase "classic meat options" and detailed sandwich variety are **not reflected in the chunk**, which only says "vegan junk food." The rest feels generic and potentially LLM-inferred. **Charlie Was a Sinner** – ⚠️ Menu claims like *brisket, pulled pork, hot dogs, sliders, mac and cheese*, and the **cocktail list** are likely hallucinations—not supported by the short chunk "good vegan food. Classy vibes." **Real Food Eatery** – ⚠️ Most of the recommendation reads like a templated healthy-eating pitch. The chunk supports "healthy fast food," but nothing in the chunk confirms *organic*, *farmers market sourcing*, or *kombucha*. These are likely stylistic embellishments. **3. Overall Fluency of Output:** ✔ All three recommendations are well-written, grammatically correct, and use a friendly, foodie tone that suits the product experience you're going for.
3	💬 **You asked: What is the best casual dinner spot near Rittenhouse?** 🗨️ Detected language: English | 🕒 2025-04-24T15:27:24.323881 | 🔑 Session: bf353619-5b65-4d54-b292-cb190ff551b7 | 💬 Query: What is the best casual dinner spot near Rittenhouse? ⏱️ Retrieval + optional rerank: 0.485s	**1. Relevance to Query:** **The Foodery at Rittenhouse** – ✔ The recommendation is well-aligned with a request for a casual dinner near Rittenhouse. It mentions dinner suitability, a casual vibe, and practical amenities like Wi-Fi and group seating.

	⏱️ Business ranking: 0.008s 🏷️ Restaurant 1: The Foodery at Rittenhouse ⏱️ → Chunk prep: 0.002s ⏱️ → Context build: 0.002s ⏱️ → LLM generation: 5.212s 🍽️ Recommendation: The Foodery at Rittenhouse (English) The Foodery at Rittenhouse is an excellent choice if you're looking for a dining experience that offers a mix of nightlife, bars, food, and restaurants. With its unique atmosphere and diverse menu options, it's perfect for those who appreciate both the casual atmosphere and the variety of culinary experiences available. Its outdoor seating area provides a refreshing break from the hustle and bustle of city life, while the free Wi-Fi and group-friendly seating make it a great spot for socializing with friends or family. Plus, the alcohol service and happy hour specials ensure you have a great time without breaking the bank. Whether you're in the mood for classic American fare or want something new and exciting, this restaurant is sure to satisfy your cravings. ☀️ You seem to be using the pipelines sequentially on GPU. In order to maximize efficiency please use a dataset 🏷️ Restaurant 2: Casta Diva ⏱️ → Chunk prep: 0.003s ⏱️ → Context build: 0.002s ⏱️ → LLM generation: 5.159s 🍽️ Recommendation: Casta Diva (English) I highly recommend the Casta Diva! It’s a cozy Italian restaurant in Philadelphia that offers an intimate dining experience, making it perfect for dates or a nice dinner with friends. The restaurant has a warm ambiance, perfect for those who want to enjoy some quality time together without the distractions of loud parties. The food at Casta Diva is delicious and expertly crafted, and the prices are reasonable, making it affordable even for families. Plus, they offer BYOB policy, which is great if you're traveling with children or have other guests. Don’t miss out on this charming and welcoming spot! It seems like there might be a misunderstanding regarding the rating provided. The review mentions a "cash only" policy, but the actual review does not 🏷️ Restaurant 3: Day By Day ⏱️ → Chunk prep: 0.003s ⏱️ → Context build: 0.002s ⏱️ → LLM generation: 5.267s 🍽️ Recommendation: Day By Day (English) If you're looking for a great brunch spot near Rittenhouse that offers a simple yet elegant atmosphere, try [REVIEW]'s [NAME]. This cozy location provides a perfect blend of casual comfort and sophisticated	**Casta Diva** – ⚠️ While it fits the “dinner near Rittenhouse” aspect and highlights ambiance, the tone leans more toward **intimate or date night** rather than casual, everyday dining. **Day By Day** – ⚠️ This response focuses heavily on **brunch and breakfast**, even though the query asks about dinner. While the casual vibe fits, the **meal timing** doesn't match the request. **2. Grounded-ness & Hallucination Risk:** **The Foodery at Rittenhouse** – ✔️ Almost entirely grounded. The chunk supports claims of **outdoor seating, Wi-Fi, group-friendliness, alcohol, and happy hour specials**. The only minor extrapolation is the vague “classic American fare,” which isn’t directly quoted but is a safe assumption. **Casta Diva** – ✔️ The description of a cozy, cash-only BYOB restaurant is grounded in the chunk. ⚠️ However, references like “traveling with children” or assuming it's family-friendly are **not supported** in the chunk and feel speculative. **Day By Day** – ⚠️ The chunk talks about **brunch**, **ambiance**, and **outdoor/indoor seating**, which aligns with what’s mentioned in the recommendation. However, the generated output includes some **generic phrases** (like “Please let me know if you...”), and it **mentions dinner**, which the chunk does not support. **3. Overall Fluency of Output:** ✔️ All three responses are fluent, coherent, and written in a welcoming, informative tone.

	elegance, making it an ideal choice for any occasion. From breakfast to lunch, they serve up delicious sandwiches, salads, and desserts that perfectly complement your meal. Don't miss out on this unique culinary experience! 🍽️ 🍽️ 🍴 If you need more recommendations or have specific preferences, feel free to ask, and I'll do my best to suggest another great option. Enjoy your day by Day By Day! ☀️ Would you like me to create a review based on this information? Please let me know if you ✅ Total time to generate recommendations: 16.151s	
4	**🗨️ You asked: Looking for a romantic Italian place for a date night.** 💬 Detected language: English | 🕒 2025-04-24T15:29:58.085234 | 🔑 Session: 8ededaa2-e861-485a-aa4e-a94530fad0a7 | 🗨️ Query: Looking for a romantic Italian place for a date night. ⏱️ Retrieval + optional rerank: 0.288s ⏱️ Business ranking: 0.008s 🏷️ Restaurant 1: PTG Catering ⏱️ → Chunk prep: 0.002s ⏱️ → Context build: 0.001s ⏱️ → LLM generation: 5.139s 🍽️ **Recommendation: PTG Catering (English)** ☀️ [Restaurant Name]: A cozy, charming Italian restaurant with an intimate atmosphere where you can feel at home. The menu offers a variety of dishes including classic Italian favorites such as pasta, pizza, and risotto. The service is attentive, and the prices are reasonable. Perfect for a casual gathering or even a romantic dinner out. The ambiance is perfect for any occasion - whether it's a casual brunch or a formal dinner party. The staff is knowledgeable about Italian food and wine and make your meal experience unforgettable. If you're looking for a unique dining experience, this place is definitely worth checking out! 🍽️ 🍽️ It seems like there might be some confusion regarding the specific type of Italian cuisine being referred to. Could you 🏷️ Restaurant 2: L'anima ⏱️ → Chunk prep: 0.003s ⏱️ → Context build: 0.002s ⏱️ → LLM generation: 5.239s	**1. Relevance to Query:** **PTG Catering** – ✔️ The recommendation highlights a **cozy**, **charming Italian setting**, explicitly describing it as suitable for a **romantic dinner**. Strong alignment with the "romantic Italian date night" query. **L'anima** – ✔️ Focuses on **authentic Italian cuisine**, **cozy atmosphere**, and **Mediterranean flavors**. It hits both the **romantic vibe** and **cuisine**, making it a good fit. **Branzino Italian Ristorante** – ✔️ Emphasizes **Italian classics**, **cozy outdoor seating**, and **elegant decor**. The tone and content match perfectly with a romantic dinner date setting. **2. Grounded-ness & Hallucination Risk:** **PTG Catering** – ✔️ The chunk supports "small, quiet setting…great for date night." ⚠️ However, the output includes mentions of brunch, wine knowledge, and formal dinner parties, which are **not reflected in the chunk** and seem inferred or generalized. **L'anima** – ⚠️ The chunk is quite minimal ("authentic Italian cuisine with a cool environment"). While the general sentiment in the recommendation aligns with the chunk, references to **seafood**, **Mediterranean focus**, and **transportive Italian experience** go beyond what is supported.

	🍽️ **Recommendation: L'anima (English)** If you're looking for an authentic Italian cuisine experience in the heart of Philadelphia, look no further than L'anima. This restaurant is a must-visit for those who love seafood, Mediterranean flavors, and a cozy atmosphere. The staff here is warm and welcoming, making your dining experience truly memorable. Don't miss out on this culinary gem that will transport you to Italy while enjoying delicious seafood dishes and fresh Mediterranean-inspired cuisine. Whether you're craving a hearty pasta dish or a delicate seafood platter, there's something for everyone at L'anima. 🍽️ ✨ The restaurant has been consistently ranked at 4.0 stars by reviewers, which indicates excellent service and quality of food. It's perfect for any occasion—whether it's 🏷️ Restaurant 3: Branzino Italian Ristorante ⏱️ → Chunk prep: 0.003s ⏱️ → Context build: 0.002s ⏱️ → LLM generation: 5.258s 🍽️ **Recommendation: Branzino Italian Ristorante (English)** Branzino Italian Ristorante! 🌟 It's the perfect spot for anyone looking to indulge in delicious Italian cuisine while enjoying a warm and welcoming atmosphere. With their extensive menu featuring classic Italian dishes like pasta, pizza, and risotto, as well as modern interpretations, Branzino offers something for everyone. The cozy outdoor seating provides the perfect setting for those trying to stay active during your meal. Don't miss out on this hidden gem! 🍽️ The restaurant has a rustic yet elegant decor with warm lighting and comfortable seating. The staff is attentive and knowledgeable, making you feel right at home. Their menu includes a variety of options, from traditional dishes to unique twists on classic Italian flavors. Whether you're a fan of ✅ Total time to generate recommendations: 15.951s	**Branzino Italian Ristorante** – ⚠️ The chunk is a **light-hearted personal review** about loving Italian food and noticing the **outdoor seating**, but the full generated response adds detailed decor descriptions, staff quality, and menu items not directly present in the chunk. These may be accurate, but they aren't explicitly grounded. **3. Overall Fluency of Output:** ✔️ All responses are fluent, stylistically polished, and match the tone of a “local foodie guide.”
5	💬 **You asked: Where can I get spicy halal food late at night?** 💬 Detected language: English | 🕒 2025-04-24T15:31:05.009221 | 🔑 Session: ca46df54-11b0-4056-82e8-f69c51483c46 | 💬 Query: Where can I get spicy halal food late at night? ⏱️ Retrieval + optional rerank: 0.263s ⏱️ Business ranking: 0.007s 🏷️ Restaurant 1: Tandoori Of India ⏱️ → Chunk prep: 0.002s	**1. Relevance to Query:** **Tandoori Of India** – ✔️ Highlights **Indian food** and **cozy atmosphere**, and the chunk specifically mentions it as a **late-night spot**, making it a good match. However, the recommendation does **not explicitly mention halal** or spice level. **Bravo Pizza & Halal Food** – ✔️ Includes **halal** prominently in both the name and the recommendation. The chunk mentions **fast delivery** and implies **late-night availability**. Though the output emphasizes friendliness and

	⏱ → Context build: 0.001s ⏱ → LLM generation: 5.241s 🍽 **Recommendation: Tandoori Of India (English)** If you're looking for a delicious Indian food experience with a cozy atmosphere, Tandoori Of India is your perfect choice! The restaurant offers a unique blend of tandoori dishes that are both flavorful and visually appealing. The service is attentive, and the ambiance is warm and inviting. Whether you're a foodie or just want to savor some traditional Indian cuisine, this place has something for everyone. Plus, they offer a variety of dessert options, making it an ideal spot for a post-meal treat. Don't miss out on their delicious flavors! Note: This restaurant does not serve alcohol. However, if you're visiting during holidays or celebrations, please be aware that the menu may change slightly due to seasonal ingredients availability. Enjoy your meal 🏷 Restaurant 2: Bravo Pizza & Halal Food ⏱ → Chunk prep: 0.003s ⏱ → Context build: 0.002s ⏱ → LLM generation: 5.190s 🍽 **Recommendation: Bravo Pizza & Halal Food (English)** Bravo Pizza & Halal Food - A deliciously satisfying fusion of pizza and halal cuisine. The warm, inviting atmosphere and fresh, flavorful ingredients make it an ideal choice for those seeking a unique dining experience without compromising their dietary preferences. It's a great place to grab a quick meal or treat yourself after a busy day at work. Plus, the friendly staff and exceptional customer service ensure you leave feeling pampered. Don't miss out on this incredible culinary journey with Bravo Pizza & Halal Food! 🍎 🍕 #HalalFood #Deliveries #Pizzadelight #DelishMeals #HalalDine #Pizzafriendly #HalalDiet #DeliciousDelivers #HalalDinner #Halal 🏷 Restaurant 3: Makkah Market ⏱ → Chunk prep: 0.003s ⏱ → Context build: 0.002s ⏱ → LLM generation: 5.211s 🍽 **Recommendation: Makkah Market (English)** [Restaurant Name] offers an excellent selection of halal meats and dairy products, with a diverse range of spices that make up the spice aisle. The restaurant also has a variety of vegetarian and vegan options. Plus, their samosas are quite tasty and the prices are reasonable for such high-quality dishes. If you're looking for a halal dining experience, this place is definitely worth checking out! 🌟 This review highlights the restaurant's	variety, it could more directly emphasize “spicy” or “late-night.” **Makkah Market** – ⚠ Focuses on halal food and spice variety (in terms of ingredients), but the output leans more toward a **grocery/deli** vibe than a **restaurant or late-night dining experience**. “Spicy” and “late night” aren’t clearly addressed in the recommendation. **2. Grounded-ness & Hallucination Risk:** **Tandoori Of India** – ✔ Chunk supports “great Indian food” and “late night spot.” ⚠ The mention of desserts, alcohol policies, and seasonal ingredients in the output is not supported by the chunk and feels speculative or boilerplate. **Bravo Pizza & Halal Food** – ⚠ While the review is grounded in terms of delivery and being halal, the LLM's recommendation includes **hashtags, exaggerated phrasing** (e.g., “culinary journey”), and a level of polish that’s not reflected in the chunk. **Makkah Market** – ⚠ The output references **vegetarian and vegan options**, and phrases like “dining experience” and “diverse range of dishes” that aren’t explicitly supported by the chunk, which focuses more on ingredients and store inventory. **3. Overall Fluency of Output:** ✔ All three are fluent and conversational, though **Bravo Pizza's** output veers more into hashtag-heavy, ad-like territory, which may feel less natural depending on your tone preference.

	versatility, offering a wide array of halal meats and dairy products. Its spice aisle, vegetarian and vegan options, and tasty samosas all contribute to making it stand out as a great choice for those seeking halal cuisine. If you're in the Philadelphia area, this should be a must-visit for ✅ Total time to generate recommendations: 15.932s	