AI Restaurant Recommender

Technical Appendix

A. Model Choices

The final system uses **GPT-5-mini** as the primary reasoning engine. This model was chosen because:

- **Fast inference speed** → suitable for per-restaurant, multi-call workflows
- **Low token cost** → supports displaying dish recommendations for all Top-5 restaurants
- **High consistency** in structured outputs (bullet lists, short descriptions)
- **Strong contextual reasoning** compared to earlier local LLM tests

Compared with the first prototype (Gemma 12B via Ollama), GPT-5-mini provides:

- Better alignment with real-world restaurant contexts
- More fluent and concise descriptions
- Significantly reduced hallucination when grounded in Maps data
- Broader general knowledge beyond Manhattan / NYC dataset

This makes GPT-5-mini essential for the final product's value proposition.

B. Data Sources

The final system combines *real-time geospatial data* with LLM reasoning:

1. Google Geocoding API

- Converts user-entered addresses into lat, lng, and city
- Enables context-aware LLM prompting (e.g., “signature dishes popular in New York Chinese cuisine”)

2. Google Places API (New / v1)

- Retrieves up-to-date restaurant candidates based on:
    - Text query (cuisine)
    - Location bias (circle radius)
    - FieldMask to maintain efficiency

This structured data forms the factual basis that the LLM uses to avoid hallucinations and produce *accurate, personalized recommendations*.

C. Prompt Strategy

AI prompts are specifically engineered for:

1. Grounded Recommendations

Each prompt includes:

- Restaurant name
- Cuisine type
- City (regional flavor context)
- Required format (bullet list, short explanations)

This ensures the LLM enhances—but never invents—information.

2. Controlled Output

Prompts enforce:

- 2–3 signature dishes
- 1-sentence per dish
- No fabricated restaurant attributes
- Clear, readable formatting for UI consumption

3. Low Token Footprint

Since GPT-5-mini is called for each restaurant, prompts are optimized to be short while still rich enough for meaningful recommendations.

D. Engineering Decisions

Key architectural choices shaped the final product:

1. AI-First Rendering

LLM calls are executed **after** ranking the Top-5 restaurants—optimizing both speed and cost.

2. Deterministic Presentation

- Same restaurant consistently produces similar dish suggestions
- Avoids user confusion & improves trustworthiness

3. Token & Cost Management

- GPT-5-mini selected for efficiency
- Prompts trimmed to reduce token usage
- Future improvement: batched prompting or on-click generation

E. Evaluation Methods

1. AI Output Evaluation

- Accuracy: dishes fit the cuisine and match city context
- Usefulness: recommendations feel actionable, not generic
- Consistency: no hallucinated restaurant names or menus

2. System-Level Evaluation

- End-to-end latency (geocoding → search → AI)
- Stability under repeated LLM calls
- Ranking relevance (ratings + review counts validated manually)

Summary

This updated Technical Appendix elevates AI from a supplementary feature to the **core intelligence layer** of the product.
It clearly shows how **GPT-5-mini + Google Maps data** work together to produce a recommendation experience that is grounded, trustworthy, and truly helpful.