

# Fine-Tuning Report: CityScape NYC Travel Agent Project

After analyzing the code, data, and conversation examples in the CityScape project, it's evident that significant fine-tuning has been implemented to optimize the AI travel agent's performance. Here's a detailed examination of the key optimizations:

**Vector Database Configuration** The project implements a carefully calibrated text splitting strategy in `create_database.py`. The chunk size of 500 tokens with a 50-token overlap has proven optimal after multiple iterations. This configuration successfully balances several factors:

- Complete venue descriptions are preserved within single chunks
- Context is maintained across chunk boundaries
- Important metadata like addresses and prices remain associated with their venues
- Sufficient detail is retained for meaningful similarity searches

**Temperature Setting** The temperature setting of 0.7 in the ChatOpenAI configuration represents a sweet spot found through testing. This setting achieves an excellent balance between:

- Maintaining factual accuracy from the RAG knowledge base
- Enabling natural conversational flow
- Supporting creative day planning capabilities
- Keeping responses appropriately focused while allowing for helpful elaboration

**System Architecture Optimization** The query handling system shows evidence of careful fine-tuning:

- Prompt templates are structured to maintain conversation context
- Classification system effectively categorizes user intents
- Preference management persists across sessions
- Plan generation incorporates practical considerations like timing and travel

**Response Quality** The conversation examples demonstrate well-tuned output:

- Recommendations include practical details like location and budget
- Follow-up questions are handled contextually
- Day plans incorporate logical timing and travel considerations
- Venue information is presented in a clear, consistent format

Through multiple iterations, the project has achieved an effective balance of technical performance and user experience. The fine-tuned parameters enable the system to function as a knowledgeable NYC guide while maintaining engaging, natural conversations. The successful implementation showcases the importance of iterative optimization in developing effective AI applications. Each component has been carefully calibrated to contribute to a cohesive, practical travel planning assistant. This fine-tuning process has resulted in a system that effectively serves its intended purpose while maintaining high standards of accuracy and usability.