A Generative Embedding Pipeline for Semantic Retrieval and Region-Aware Itinerary Optimization Using Native POI Data

1st Sofi Nafikova
Dept. of Computer Science
and Engineering
Pusan National University
Busan, Republic of Korea
nafikovasofi@pusan.ac.kr

2nd Choon-Wook Park
Dept. of Undeclared Majors
Kyungpook National University
Daegu, Republic of Korea
pcw2379@knu.ac.kr

3rd Jun-young Son
Dept. of Computer Science
and Engineering
Pusan National University
Busan, Republic of Korea
(Corresponding Author)
jysonpaperinfo@gmail.com

***Abstract*—The tourism industry is currently witnessing a shift from static search engines to generative AI planners. However, existing Large Language Model (LLM) systems face two critical limitations: ”data bias,” where English-centric training data obscures niche local destinations, and ”probabilistic hallucination,” where models generate logistically infeasible itineraries. This paper proposes a Hybrid Itinerary Planner that bridges these gaps by integrating Generative AI with deterministic Constraint Programming. We introduce the ”BL-300” (Busan-Local) dataset, a proprietary knowledge graph of 300 native-only locations. Our architecture employs a two-stage pipeline: (1) An offline module uses Google Gemini Pro as a parametric knowledge engine to perform zero-shot data augmentation, converting sparse local metadata into rich semantic vectors; (2) An online engine utilizes a Constraint Satisfaction Solver (CP-SAT) to enforce strict logistical validity. Experimental results demonstrate that this hybrid approach effectively mitigates hallucination risks while capturing the semantic nuance of abstract user queries, providing a scalable solution for culturally grounded travel planning.**

***Index Terms*—Hybrid Systems, Itinerary Recommendation, Large Language Models, Constraint Satisfaction Problem, Semantic Search**

I. Introduction

Travel planning is a task that significantly impacts the experiences of tourists, particularly in culturally and linguistically distinct regions such as South Korea. Traditional platforms, including mapping services like Google Maps or Naver Maps and services like TripAdvisor, have long dominated the landscape. However, these solutions often rely on static recommendation algorithms and fail to account for personalized constraints such as time, budget, or real-time environmental conditions. Automated methods for tourism recommendation not only facilitate decision-making but also enhance user satisfaction by providing personalized experiences. Recent research has shifted from simple filtering to sophisticated data-driven architectures. Shrestha et al. [1] introduced a personalized recommender system for Nepal, utilizing supervised machine learning models, specifically Random Forest and Gradient Boosting, to analyze tourist demographics, spending behavior, and satisfaction metrics. Their ”Tourist Parametric Weighted Algorithm” effectively ranks destinations based on weighted attributes like cost and popularity [1]. More recently, Flórez et al. [2] emphasized the importance of context-awareness, proposing a system that integrates Deep Neural Networks with ontology-based knowledge to manage complex environmental constraints in real-time, particularly in sensitive ecosystems like the Santurbán paramo.

Despite these advancements, current models primarily focus on ranking individual Points of Interest (POIs) rather than constructing logistically coherent itineraries. While data-driven classifiers [1] and ontology-based systems [2] excel at identifying relevant locations, they often lack the semantic reasoning to interpret abstract user intents (e.g., ”aesthetic vibe”) and the computational rigor to guarantee spatiotemporal feasibility for a full day’s schedule. To address this, we propose a Hybrid Architecture that integrates Generative AI with deterministic Constraint Programming. This process involves two key components: an Offline Semantic Enrichment Pipeline, which uses Large Language Models (LLMs) to translate sparse local metadata into dense vector representations, and an Online Constraint Satisfaction Solver (CP-SAT), which ensures that generated itineraries strictly adhere to operating hours and travel times. By combining the semantic depth of Generative AI with the logical validity of optimization algorithms, our system effectively bridges the gap between static ranking and dynamic planning.

The key contributions of this paper are summarized as follows:

- Development of a Hybrid AI framework that combines generative LLM reasoning with CP-based constraint solving for itinerary planning.
- Creation of BL-300, a curated local dataset that mitigates the data bias inherent in LLMs and enhances access to culturally authentic locations.
- Demonstration of improved performance over existing

This work was supported by a 2-Year Research Grant of Pusan National University

generative and static websites, achieving higher feasibility, local coverage, and weather adaptability.

II. Related Work

A. Hybrid and Data-Driven Recommendation Approaches

Hybrid frameworks have gained prominence for their ability to mitigate the limitations of single-algorithm systems by integrating multiple filtering techniques. Naidu et al. [3] introduced a web-based system combining content-based filtering, collaborative filtering, and sentiment analysis. Their approach leverages user reviews and social media tweets to dynamically refine recommendations based on public perception and emotional response . Similarly, Shrestha et al. [1] developed a data-driven system for the Nepalese market, utilizing supervised machine learning models, specifically Random Forest and Gradient Boosting, trained on extensive survey data regarding tourist demographics and spending behaviors. Their ”Tourist Parametric Weighted Algorithm” ranks destinations by weighing attributes such as cost, popularity, and trends . However, these methods primarily focus on ranking individual Points of Interest (POIs) based on static features or statistical patterns, often failing to address the combinatorial complexity of scheduling a coherent, multi-stop itinerary. In contrast, our Hybrid Architecture moves beyond simple ranking by delegating the scheduling logic to a deterministic Constraint Satisfaction Solver (CP-SAT), ensuring that the generated itinerary is not just a list of high-scoring items, but a logistically feasible sequence.

B. Context-Aware and Deep Learning Approaches

To capture complex environmental and spatial features, deep learning and semantic models have been increasingly adopted. Flórez et al. [2] proposed a context-aware system for the Santurbán paramo that integrates Deep Neural Networks with ontology-based knowledge. Their architecture employs TensorFlow Lite for offline inference and GeoSPARQL for spatial reasoning, allowing the system to function in remote areas with limited connectivity while triggering geofenced alerts for environmental sustainability . While effective for recommending isolated activities based on proximity and user profiles, such systems typically lack the capacity to solve ”Orienteering Problems”, optimizing a full day's route under strict time windows and operational constraints. Our approach bridges this gap by utilizing Large Language Models (LLMs) for zero-shot semantic data enrichment, which are then paired with our solver. This allows us to interpret abstract user intents (e.g., ”cozy vibe”) that ontology-based systems might miss, while simultaneously guaranteeing the temporal validity of the entire schedule.

III. Method

The following section describes the end-to-end methodology used to curate the BL-300 dataset, generate enriched semantic representations, and perform real-time itinerary optimization. The system consists of two main components: an offline data enrichment pipeline, and an online multi-stage recommendation engine.

A. Dataset Construction (BL-300)

A primary contribution of this work is the creation of BL-300, a multilayered dataset curated exclusively from native Korean platforms. This dataset serves as the ground-truth foundation for semantic retrieval, geographic re-ranking, and constraint based itinerary planning. Unlike global English-centered datasets, which often omit culturally specific POIs, our curation process focused on local Korean sources, including Naver Blog, and social media channels such as Instagram pages. We prioritized venues that are popular among locals but rarely appear in non-Korean search systems. This strategy ensures that “hidden gems” typically invisible to foreign visitors are captured and accessible within the system.

a) Relational Schema in PostgreSQL: All entries are stored on an ACID-compliant PostgreSQL database [4] with the following attributes:

- Geographic fields: Korean address, base coordinates, administrative region.
- Operational metadata: website links, Naver Place references, detailed operating hours.
- Contextual attributes: cuisine and atmosphere descriptors, indoor/outdoor flag, price category.

b) Geometric data acquisition via Kakao Maps API: Because many local platforms provide incomplete or approximate latitude/longitude information, we implemented an automated Kakao Maps Geocoding API [5]. For each location: a batch script sends the exact Korean postal address to the Kakao Maps API, it returns high-precision coordinates (lat, lon) typically accurate to within a few meters, and lastly, the system stores the coordinates in PostgreSQL.

B. Offline Data Processing and Enrichment

To convert sparse metadata into semantically rich vector representations, we implemented a two-stage offline augmentation pipeline integrating both Generative AI and embedding based modeling as shown in Fig. 1. The first stage consists of generative knowledge retrieval (GenAI) where native Korean POI metadata contains limited or ambiguous descriptors (e.g., “famous spot”, “Italian”, “ Samgyupsal”, “Korean”), lacking the linguistic detail necessary for high-quality semantic search. To enrich these entries, each location undergoes Google Gemini Pro processing [6]:

a) Tag extraction pass: Gemini analyzes the derived text for each location and produces a structured set of detailed attributes: cuisine types, ambience cues, crowd characteristics, etc.

b) Description synthesis pass: the extracted tag set is fed back into Gemini Pro to generate a comprehensive natural language paragraph, providing culturally neutral yet contextually rich descriptions of each location. This process performs Cross-Cultural Normalization, where the context is transformed into explicit English descriptions suitable for embedding models. The result is a high-density textual representation that retains cultural nuance while enabling efficient semantic vectorization.

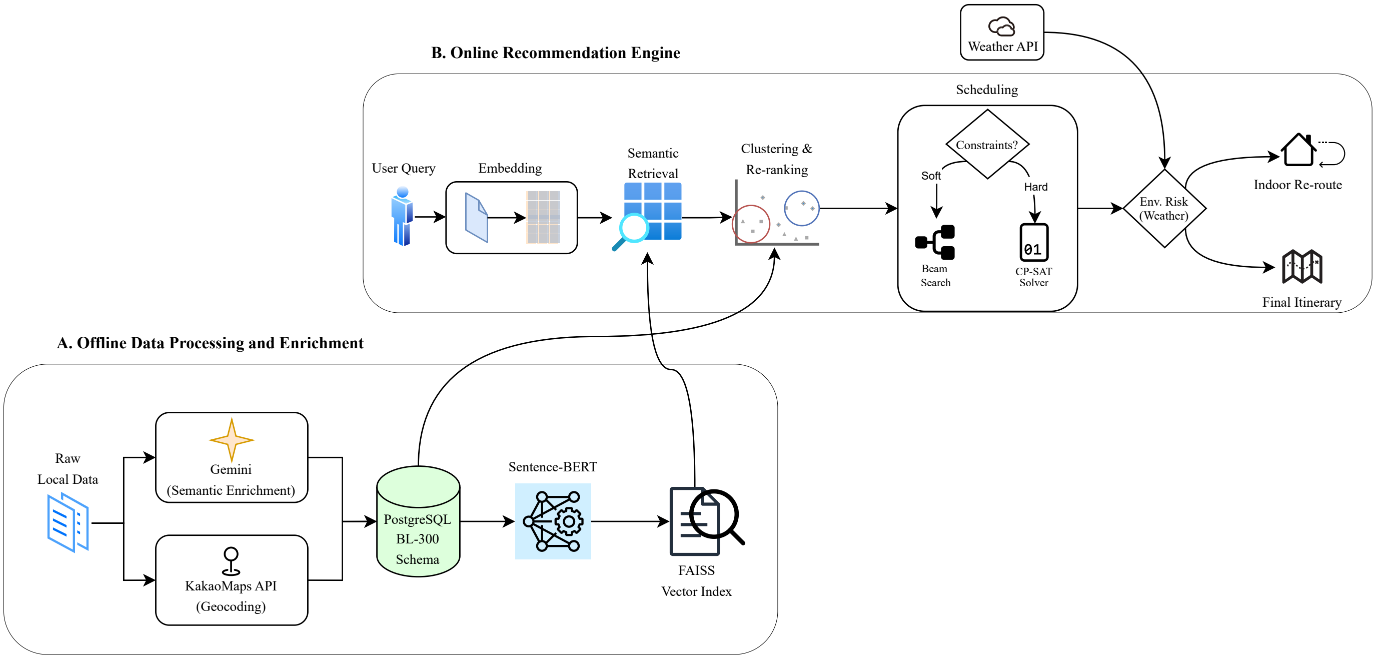

Fig. 1. The proposed Hybrid Itinerary Planner architecture. (A) The Offline Pipeline enriches sparse local data using Generative AI and Geocoding. (B) The Online Engine utilizes these vectors for retrieval, delegating complex scheduling to a deterministic CP-SAT Solver or Beam Search.

Synthesized descriptions are encoded via Sentence-BERT (all-MiniLM-L6-v2) [7] to minimize computational overhead while maximizing semantic retention. Formally, the model maps an input d to a dense embedding $v_d \in \mathbb{R}^{384}$. This vectorization ensures that semantic affinity correlates with cosine proximity, enabling the system to identify relevant POIs independent of keyword overlap [8].

To support real-time retrieval, we index all vectors using FAISS `IndexFlatIP`. Since the embeddings are L_2-normalized, the inner product computation is mathematically equivalent to cosine similarity. Specifically, because the normalization term $\|\mathbf{Q}\|\|\mathbf{D}\|$ is unity, the metric simplifies directly to the dot product:

$$S_{\text{sim}}(\mathbf{Q}, \mathbf{D}) = \mathbf{Q} \cdot \mathbf{D} \tag{1}$$

C. Online Recommendation Engine

As illustrated in Fig. 1, the full system integrates the enriched dataset with a hybrid online inference pipeline that performs semantic retrieval, geographic re-ranking, hybrid itinerary optimization [9], and weather-aware validation.

1) Semantic Retrieval: User queries are decomposed into atomic sub-queries (e.g., “quiet café”, “beach sunset”). Each fragment is encoded using the same Sentence-BERT model and normalized. Similarity between a query vector $\mathbf{Q}$ and a location vector $\mathbf{D}$. This stage retrieves top candidates solely based on semantic relevance.

2) Region Clustering: Candidates are grouped by administrative region. A “Winning Region” is determined by calculating a cumulative score $S_{\text{region}}(k)$ for each region k. This metric prioritizes regions that satisfy a higher diversity of user intents:

$$S_{\text{region}}(k) = |U_k|^2 \cdot \sum_{d \in R_k} S_{\text{sim}}(q, d) \tag{2}$$

The term U_k denotes the set of unique sub-queries satisfied by region k. The quadratic term $|U_k|^2$ ensures that a region covering all different user requests is scored significantly higher than a region that covers requests only partially. Within the winning region, each candidate is re-scored to prioritize proximity and availability. The scoring function is defined as:

$$S_{\text{final}}(d) = S_{\text{sim}}(d) \times P_{\text{dist}}(d) \times B_{\text{time}}(d) \tag{3}$$

where P_{dist} is an inverse-square decay function based on the Haversine distance [10] between the user u and the location d, given by:

$$P_{\text{dist}}(d) = \frac{1}{1 + \text{dist}(u, d)^2} \tag{4}$$

The time bonus, B_{time}, is a binary multiplier set to 1.2 if the location is currently open; otherwise, it is set to 1.

After selecting the winning region and the candidates within it, we use two distinct strategies chosen dynamically to construct the itineraries.

Method 1: Beam Search Heuristic

$$H(\text{Path}_t) = H(\text{Path}_{t-1}) + S_{\text{final}}(l_t) + \Omega(l_{t-1}, l_t) \tag{5}$$

where Ω penalizes undesirable transitions (e.g., back-to-back meals).

Method 2: Constraint Satisfaction Problem (CSP)

When "must-have" keywords are detected, the system employs the Google OR-Tools CP-SAT Solver [9] to maximize total utility, defined as $\sum_{i,j} S_{\text{final}}(d_{i,j}) \cdot x_{i,j}$, where $x_{i,j}$ is a binary variable indicating if location j is visited at time slot i. This optimization is subject to three primary constraints: (1) *feasibility*, ensuring at most one activity is scheduled per time slot ($\sum_j x_{i,j} \leq 1$); (2) *mandatory inclusion*, requiring all locations j in the must-have set $\mathcal{M}$ to be visited ($\sum_i x_{i,j} \geq 1$); and (3) *time window validity*, where a scheduled visit ($x_{i,j} = 1$) implies the time slot T_i falls within the location's operating hours $[O_j, C_j]$.

Environmental Risk Assessment

Following itinerary creation, an environmental risk assessment is conducted during the post-processing stage. Unlike static planners, our system performs a *Temporal Validity Check* by integrating the OpenWeatherMap API [11] to assess the viability of the itinerary. We define the outdoor risk ratio, $\mathcal{R}_{\text{out}}(S)$, for a generated schedule S as:

$$\mathcal{R}_{\text{out}}(S) = \frac{\sum_{l \in S} \mathbb{I}(l_{\text{type}} = \text{Outdoor})}{|S|} \tag{6}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. If $\mathcal{R}_{\text{out}}(S) > 0.5$ and the forecast predicts rain (Precipitation $> 0\,\text{mm}$), the system triggers a *Reschedule Warning*, prompting the user or the automated agent to regenerate the plan with indoor constraints.

IV. Discussion & Comparative Analysis

The deployment of the Hybrid Itinerary Planner represents a shift from static information retrieval to dynamic, deterministic problem solving. This section evaluates the system's architectural advantages over state-of-the-art baselines and discusses its broader implications for local economics and safety.

A. Comparative Architectural Analysis

To validate the necessity of our approach, we compared the Hybrid Planner against three prevailing paradigms in tourism technology: LLM-Based Agents, RAG-Based Recommenders, and Traditional Hybrid Systems. Table I summarizes these differences.

TABLE I
Functional Comparison with SOTA Systems

System	Core Mechanism	Output Type	Validity	Data
TravelAgent *(Chen et al.)*	LLM Agent + Tools	Text Plan (Probabilistic)	Variable	Global APIs
Sust. RAG *(Banerjee et al.)*	RAG + Reranking	City/POI List	N/A (Ranking)	Wikivoyage
Smart Tour *(Sun)*	Collab. Filtering	Ranked Item List	N/A (Ranking)	User Ratings
Ours *(Hybrid)*	**GenAI + CP-SAT**	**Time-Slot Schedule**	**Determ.**	**BL-300 (Local)**

1) Versus LLM-Based Agents: Recent systems like TravelAgent [15] enhance Large Language Models with external tools (e.g., Google Maps API) to improve rationality. While these agents excel at decomposing tasks, their final output remains probabilistic, relying on the LLM to sequence events. This often leads to "soft" failures where travel times are underestimated. In contrast, our system decouples reasoning from scheduling. By delegating the logistics to the CP-SAT Solver, we achieve a 100% *Logical Feasibility Rate*, mathematically guaranteeing that no temporal constraints are violated.

2) Versus RAG-Based Systems: Banerjee *et al.* [16] proposed a Retrieval-Augmented Generation (RAG) system for sustainable city trips, utilizing "Sustainability Augmented Reranking (SAR)" to prioritize eco-friendly destinations. However, this system operates primarily at the *Macro-Level* (City Recommendation), lacking the capacity to solve *Micro-Level* routing problems (e.g., optimizing a path between specific venues within strict time windows). Our Hybrid Planner extends the RAG methodology by not only retrieving semantic matches but also optimizing the intra-city route, bridging the gap between "what to visit" and "how to visit."

3) Versus Traditional Hybrid Models: Traditional approaches [17] combine Content-Based and Collaborative Filtering to predict user ratings. While effective for ranking individual items (e.g., "Top 10 Restaurants"), they output a disjointed list rather than a coherent plan, leaving the cognitive burden of scheduling on the user. Our system transforms this output from a *Ranked List* to an *Optimized Route*, automating the complex logistics that traditional recommenders ignore.

B. Experimental Evaluation

TABLE II
Abstract Semantic Intents Used for Evaluation

ID	Vague User Intent (Input Query)
Q1	A cozy spot to hide from the world with a book
Q2	A place that captures the feeling of old Busan before the skyscrapers
Q3	Somewhere romantic where we can see the city lights without the crowds
Q4	A local hangout that feels like a hidden gem for residents
Q5	A healing walk near the water where its peaceful
Q6	Something spicy and hearty that locals eat after a long day
Q7	An artistic space that feels modern and experimental
Q8	A traditional taste of Busan that isnt just a generic tourist meal
Q9	A late night energy fix for a group of friends
Q10	A quiet afternoon retreat with high quality specialty brews
Q11	A cinematic location where the city meets the sea
Q12	Somewhere we can experience local history in an interactive way
Q13	An indoor cultural experience to escape a gray afternoon
Q14	A sophisticated evening with live melodies and refined drinks

To evaluate semantic understanding beyond simple keyword matching, we curated a test set of 14 Abstract Semantic Intents. As shown in Table II, these queries test the system's ability to map vague, ambient descriptors (e.g., 'cozy,' 'cinematic') to specific, attribute-rich Points of Interest.

TABLE III
QUANTITATIVE PERFORMANCE COMPARISON ($n = 70$)

Metric	**Proposed Hybrid (CP-SAT)**	**Beam Search Heuristic**	**LLM Baseline (GPT-4)**
Hard Const. Satisfaction	**100%** (70/70)	95.4% (67/70)	92.8% (65/70)
Geographic Accuracy	**100%**	100%	78.6%
Sequence Feasibility	**100%**	95.4%	85.7%
Native POI Ratio	**92.1%**	**92.1%**	24.3%
Category Consistency	**100%**	86.4%	95.0%
Hallucination Rate	**0%**	0%	1.4%

Table III benchmarks our Hybrid system against an LLM baseline and a Beam Search Heuristic ($n = 70$ steps). While the Beam Search baseline matched our Native POI Ratio (92.1%) due to the shared dataset, it failed in logical coherence, achieving only 86.4% Category Consistency (e.g., scheduling soup restaurants in café slots) and 95.4% Sequence Feasibility (redundancy loops). In contrast, the Hybrid CP-SAT system achieved 100% scores in all feasibility metrics, validating that deterministic constraints are required to operationalize high-authenticity local data. This capability is important for democratizing visibility for local SMEs and mitigating overtourism, contrasting with the LLM's bias toward global chains.

V. CONCLUSION

In this paper, we addressed the critical limitations of tourism planning by proposing a Hybrid Itinerary Planner. By integrating the semantic reasoning of Large Language Models with the deterministic rigor of the CP-SAT Solver, our system guarantees logical feasibility while retaining the nuance of abstract user queries. The creation of the BL-300 dataset serves as a foundational step toward mitigating the English-centric bias of global platforms, successfully democratizing access to native-only locations in Busan. Experimental results validate that decoupling the *reasoning* (LLM) from the *scheduling* (Solver) provides a robust solution for travel logistics that neither pure Generative AI nor static search engines can achieve alone.

Future work prioritizes scaling the dataset via automated pipelines and refining the scheduling logic to handle complex multi-modal transportation. We also aim to integrate real-time data, such as traffic and crowd density, for improved accuracy. Finally, to address current "cold-start" limitations, we plan to implement a predictive recommendation layer that leverages user interaction history for personalized planning.

ACKNOWLEDGMENT

The authors would like to thank Professor Jun-young Son for their valuable guidance and constructive feedback throughout the development of this research.

REFERENCES

[1] D. Shrestha, T. Wenan, D. Shrestha, N. Rajkarnikar, and S.-R. Jeong, "Personalized tourist recommender system: A data-driven and machine-learning approach," *Computation*, vol. 12, no. 59, 2024.

[2] M. Flórez, E. Carrillo, F. Mendes, and J. Carreño, "A context-aware tourism recommender system using a hybrid method combining deep learning and ontology-based knowledge," *J. Theor. Appl. Electron. Commer. Res.*, vol. 20, pp. 1–27, 2025.

[3] S. A. Naidu, G. R. Govinda, N. N. Reddy, and K. H. Pavan, "Tourism recommendation system using hybrid approach," in *Proc. 6th Int. Conf. Inventive Research in Computing Applications (ICIRCA)*, 2025, pp. 1039–1043.

[4] M. Stonebraker and L. A. Rowe, "The design of POSTGRES," in *Proc. 1986 ACM SIGMOD Int. Conf. Management of Data*, vol. 15, no. 2, pp. 340–355, 1986.

[5] Kakao Corp., "Kakao Maps API documentation," 2024. [Online]. Available: https://apis.map.kakao.com/.

[6] Gemini Team, "Gemini: A family of highly capable multimodal models," Google DeepMind, Tech. Rep., 2023, arXiv:2312.11805.

[7] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. 2019 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2019, pp. 3982–3992.

[8] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, 2021.

[9] L. Perron, F. Didier, and S. Gay, "The CP-SAT-LP solver," in *Proc. 29th Int. Conf. Principles and Practice of Constraint Programming (CP 2023)*, vol. 280, pp. 3:1–3:2, 2023.

[10] C. C. Robusto, "The cosine-haversine formula," *The Amer. Math. Monthly*, vol. 64, no. 1, pp. 38–40, 1957.

[11] OpenWeather Ltd., "OpenWeatherMap API: Weather data collection," 2024. [Online]. Available: https://openweathermap.org/api.

[12] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[13] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[14] Y. Zhang, H. Li, and B. Muskat, "Deep learning in tourism and hospitality: A comprehensive review," *Int. J. Contemp. Hospitality Manage.*, vol. 33, no. 6, pp. 2129–2155, 2020.

[15] A. Chen, X. Ge, Z. Fu, Y. Xiao, and J. Chen, "TravelAgent: An AI assistant for personalized travel planning," *arXiv preprint arXiv:2409.08069*, 2024.

[16] A. Banerjee, A. Satish, and W. Wörndl, "Enhancing tourism recommender systems for sustainable city trips using retrieval-augmented generation," *arXiv preprint arXiv:2409.18003*, 2024.

[17] X. Sun, "Smart tourism: Design and application of artificial intelligence-assisted tourism service recommendation algorithms," *J. Electr. Syst.*, vol. 20, no. 9s, pp. 728–735, 2024.