Evaluating User Intent Classification and Hybrid Retrieval in a RAG-based Conversational Tourism Recommender System

Akshat Tandon*, *Ashmi Banerjee**

Technical University of Munich, Germany

RecSys acm.org

RecTour 2025
ACM RecSys Workshop on Recommenders in Tourism

*Equal Contributions

Agenda

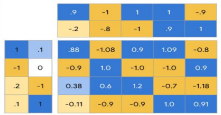
- Introduction & Motivation
- Our Approach: Hybrid RAG-driven Conversational TRS
- Evaluation
- Q&A & Discussion

Introduction

User Desires

- Personalized Recommendations
- Reliable Suggestions
- Smooth User Experience

Traditional RS Challenges



- Heavy reliance on past data
- Cold-start problem
- Scalability issues

⚠️ **Often leads to ...**
Irrelevant or premature recommendation for new and underrepresented users

The Opportunity:

- Conversational interfaces enhance user experience.
- LLMs offer natural language understanding and world knowledge.
- Retrieval-Augmented Generation (RAG) grounds LLMs, preventing hallucinations.

Goal: Build a conversational tourism recommender that's intelligent, adaptive, and grounded

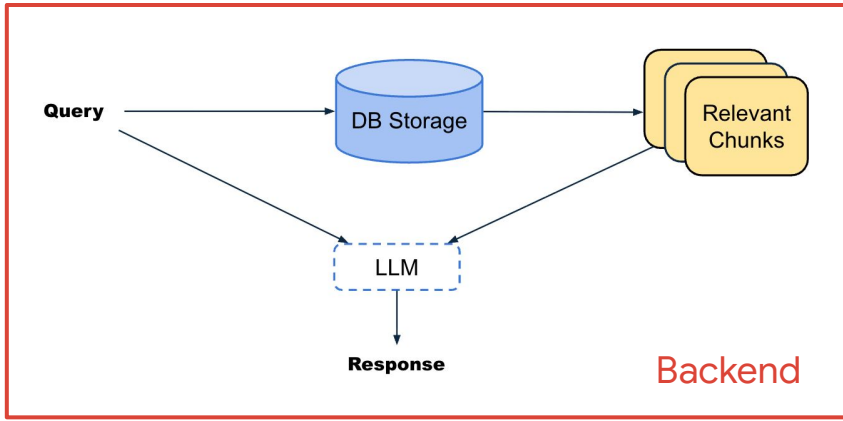
User Interaction Scenario

"Can you suggest a relaxing destination in Europe for early spring?"

Algarve, Seville, Santorini .....

That sounds interesting. I like beaches and less crowd. Can you tell me more about Algarve and its local cuisine there.

Perfect choice — Algarve is one of Europe's most relaxing early-spring beach destinations .....



Data Preparation

Data Sources

- Wikivoyage articles ~ 160 TXT files (structured & unstructured
- Tripadvisor API: Green hotels and attractions
- Total: Over 160 European cities

Preprocessing

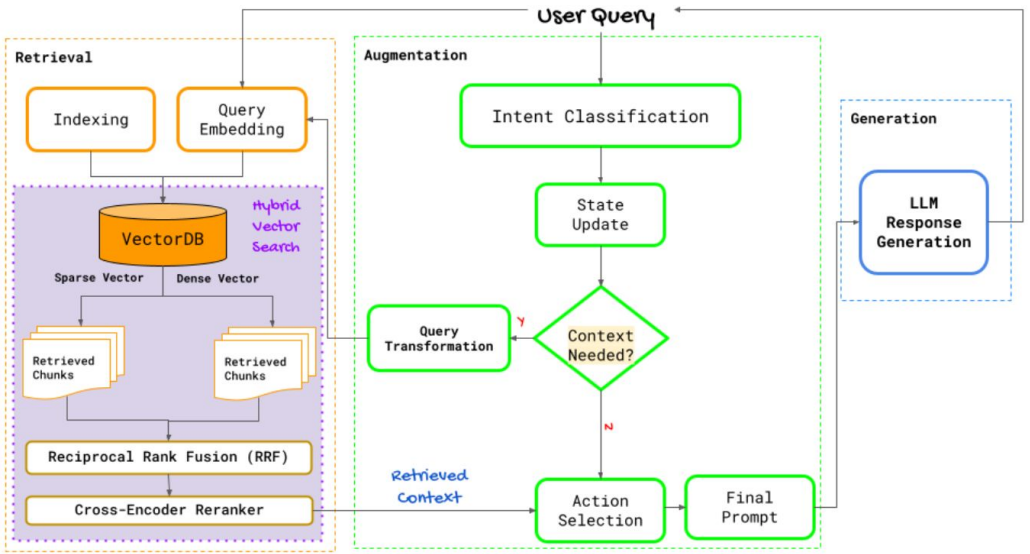
1. Clean Up
   - Remove uncommon headings from articles
   - Filter relevant features for hotels and attractions
2. Context aware and recursive chunking of content

System Design

Modular Hybrid RAG-based Conversational Tourism Recommender System (C-TRS) to recommend European cities

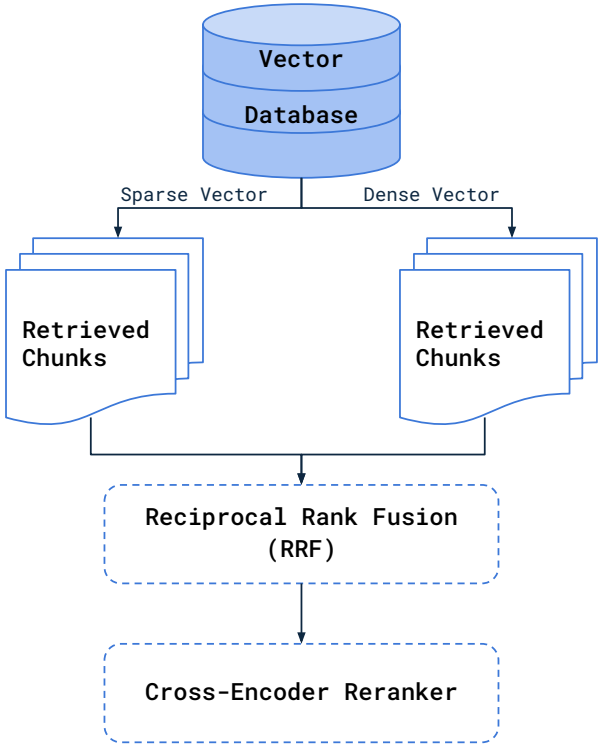
How it Works:

1. **Multi-turn conversational system**.
2. Parses user utterances for **context & intent**.
3. Utilizes **dialogue state** to manage conversation.
4. Retrieves city-level chunks **via hybrid semantic index (dense + sparse)** + optional reranking of chunks.
5. **Augments LLM prompt** with retrieved context.
6. Generates **grounded, context-aware responses**



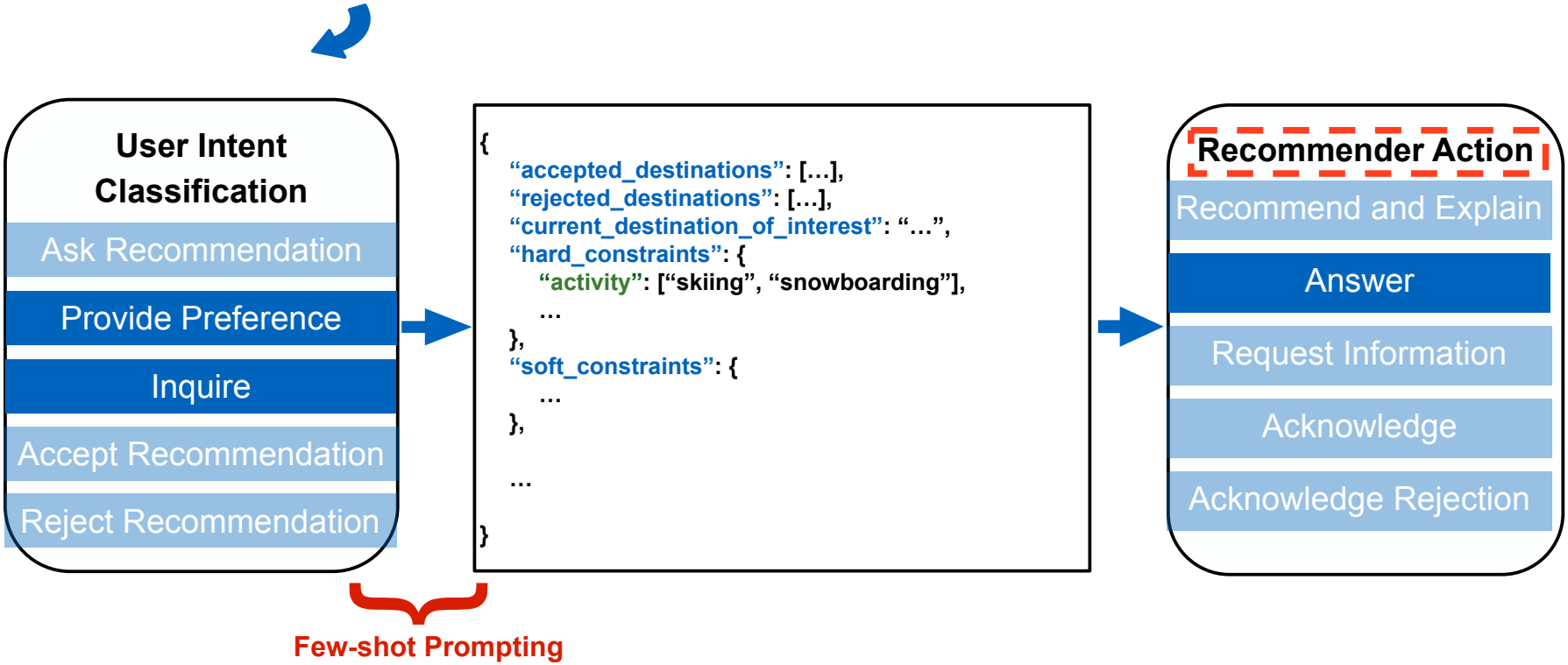
Retrieval - Hybrid Vector Similarity Search

- **Hybrid vector search** combines multiple retrieval strategies
  - Dense vectors capture **semantic meaning** and relationships
  - Sparse vectors enable lexical/**keyword** matching
- RRF merges ranked results
- Improves **recall** and **answer quality**

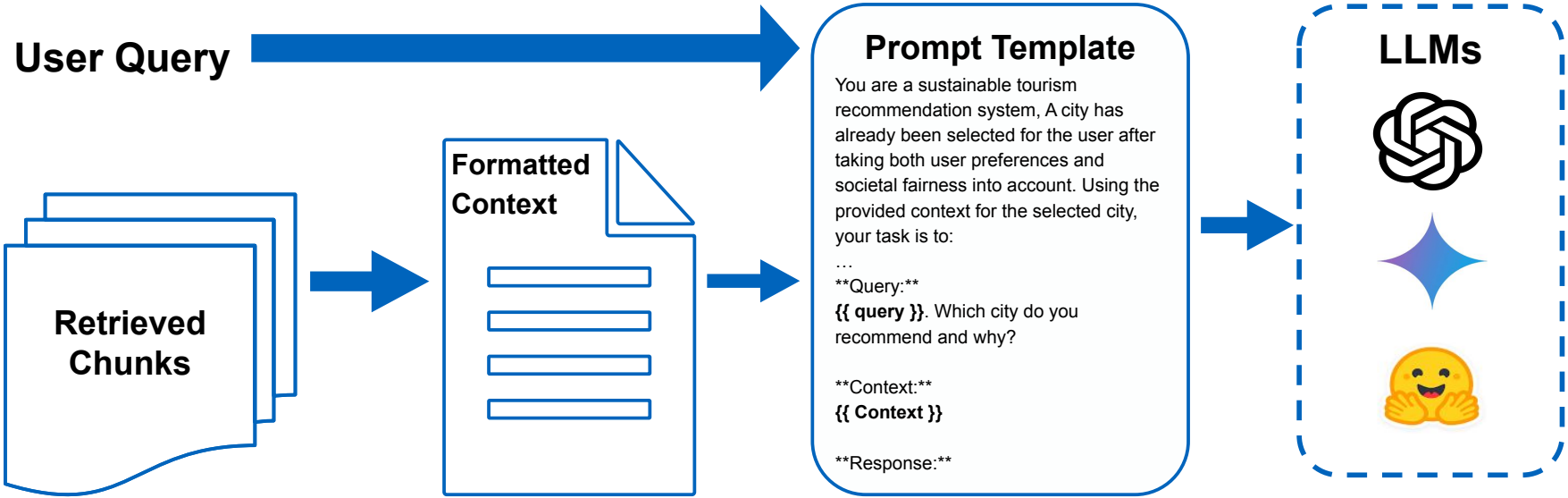


Intent-Aware Conversation Flow: Example

User Query: “Do you also know any locations where we can go skiing or snowboarding?”



RAG - Augmentation



Evaluation - User Intent Classification

Goal: Evaluate the accuracy of user intent classification.

Methods Compared:

- Fine-tuned BERT (Supervised)
- BART-large-MNLI (Zero-Shot LLM)
- GPT-4o-mini (Zero-Shot LLM)
- GPT-4o-mini (Few-Shot LLM) - Our Focus

Dataset: 330 labeled user utterances, split into 80/10/10 for training, validation, and testing

Key Finding:

- **Few-shot** classification with LLMs outperforms **zero-shot** across all metrics
- **GPT-4o-mini** achieves highest score across most metrics
- **BERT** remains competitive with a **91%** precision and **88%** F1-score

Evaluation - User Intent Classification

Model	Accuracy	Precision	Recall	F1-score
BERT Sequence Classifier	68 %	**91 %**	85 %	88 %
BART-large-MNLI (zero-shot)	3 %	32 %	67 %	43 %
GPT-4o-mini (zero-shot)	35 %	67 %	69 %	68 %
GPT-4o-mini (few-shot)	**74 %**	87 %	**96 %**	**91 %**

Evaluation - RAG Pipeline - Q&A

Goal: Evaluate the quality of retrieval and generation.

Framework: RAGAS (**LLM judge:** GPT-4o-mini | **Response LLM:** Llama-3.1-8B-Instruct | **top_k:** 5)

Metrics:

- Context Recall: *Did we retrieve enough relevant chunks?*
- Context Precision: *Were retrieved chunks actually relevant?*
- Faithfulness: *Is the LLM output supported by retrieved context (no hallucination)?*
- Answer Relevancy: *Is the LLM output relevant to the query?*

Evaluation - RAG Pipeline - Q&A

Experiment

- Compared different retrieval strategies (Dense, Sparse, Hybrid) with/without reranking.
    - 50 synthetically generated single-hop Q&A pairs for 5 European cities (Wikivoyage articles)

Key Findings

- **Sparse vector search** with **reranking** yields highest context recall (**77%**) and precision (**83%**)
- **Hybrid vector search** outperforms other approaches for **generation metrics**
- Reranking shows modest improvements in context precision

Evaluation - RAG Pipeline - Q&A

Vector Search Type	Context Recall	Context Precision	Faithfulness	Answer Relevancy
Dense Search	62 %	66 %	79 %	83 %
Dense Search + Rerank	62 %	68 %	75 %	81 %
Sparse Search	76 %	82 %	77 %	83 %
Sparse Search + Rerank	**77 %**	**83 %**	76 %	82 %
Hybrid Search	73 %	73 %	**81 %**	89 %
Hybrid Search + Rerank	68 %	75 %	79 %	**90 %**

Future Work

Conduct user study to expand the user intent and recommender action taxonomy

Conduct an ablation study to understand the contribution of intent-driven retrieval

Expand evaluation to larger datasets and multi-hop queries

Explore advanced prompting (e.g., CoT) or fine-tune smaller models for domain tasks

Thank You! Time for Q&A!

Akshat Tandon*
akshat.tandon@tum.de

Ashmi Banerjee*
ashmi.banerjee@tum.de

*Equal Contributions