Experts

Generating Sustainable Travel Recommendations using LLMs

Ashmi Banerjee

PhD Candidate @ TU Munich,
ML GDE, WTM Ambassador
@ashmi_banerjee

About Me

Photo from IWD Munich 2023

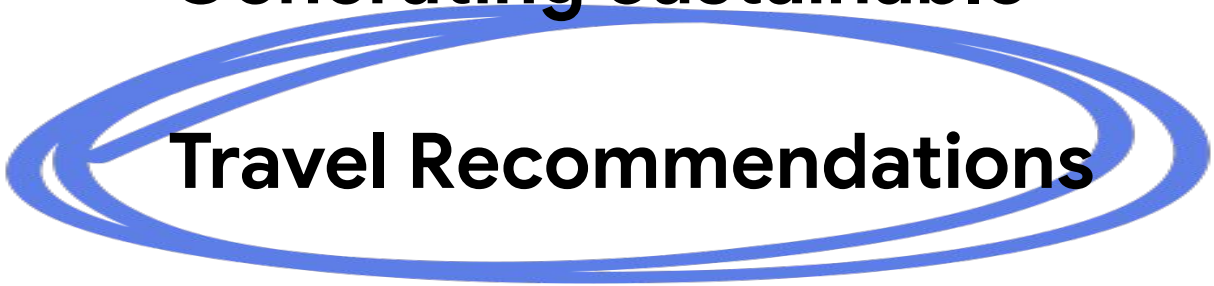
PhD Candidate at TU Munich (since 2022)
- Researching Tourism Recommender Systems

Previously, Data Scientist @ Deutsche Telekom,
Intern @ MPI-SWS (2018-19)

MSc. TU Munich (2019)

AI/ML Google Developer Expert (GDE) &
WTM Ambassador
Loves traveling & outdoor activities 🏊🚴🏃

Generating Sustainable Travel Recommendations using LLMs

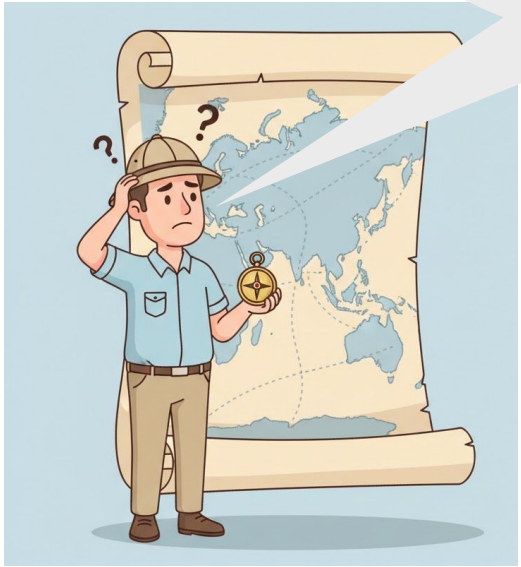


Wait but...

Generating Sustainable

Travel Recommendations

using LLMs



But data or even ground truth?

Generating Sustainable

Travel Recommendations

using LLMs

In this talk …

Step 1: Generate synthetic queries for sustainable travel recommendations

Step 2: Generate sustainable travel recommendations using these & RAG

Keeping it high-level today – feel free to read our papers or chat with me afterward!

Part 1: Generating synthetic queries, using a factually grounded dataset

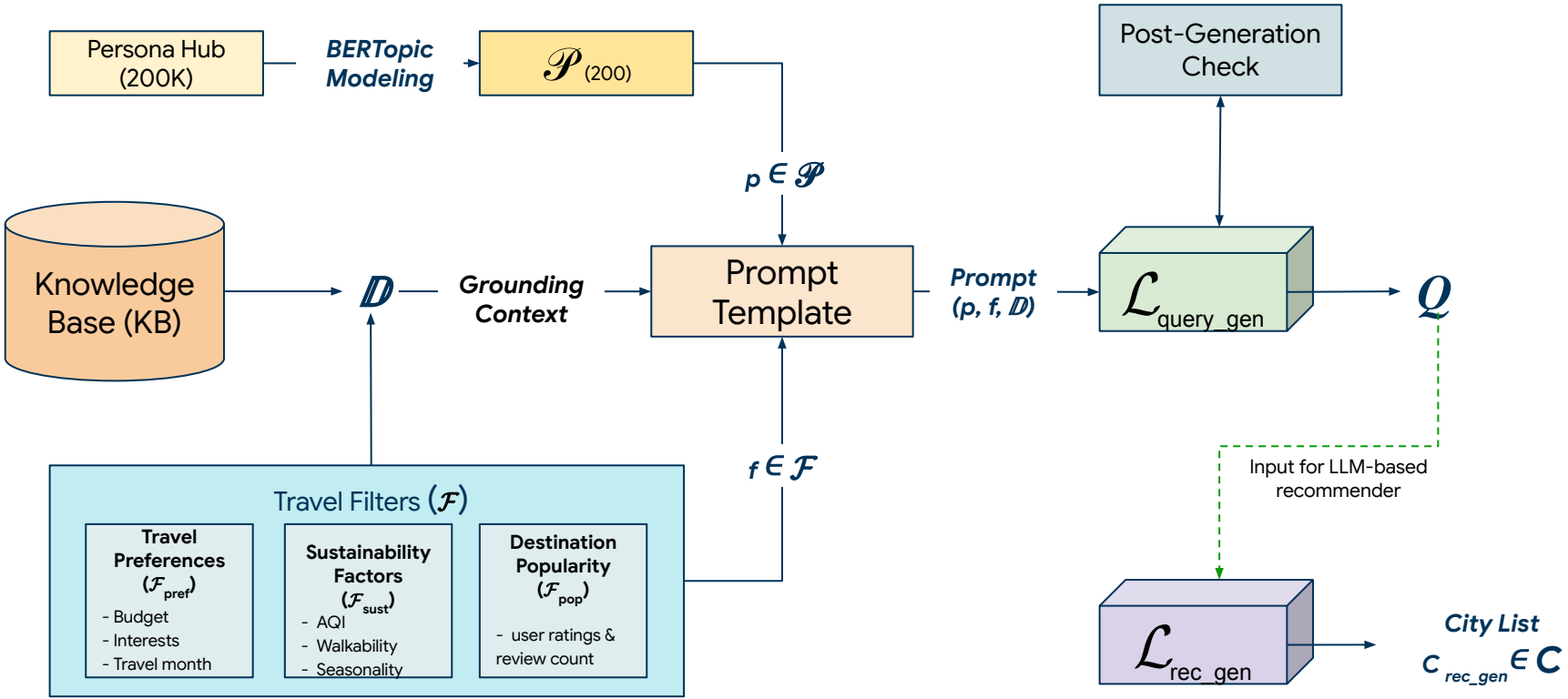
Completely open-source

Based on the work:

SynthTRIPs: A Knowledge-Grounded Framework for Benchmark Query Generation for Personalized Tourism Recommenders

Ashmi Banerjee, Adithi Satish, Fitri Nur Aisyah, Wolfgang Wörndl, and Yashar Deldjoo, [To appear] In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25) July 2025.

Generation Methodology



SynthTRIPS - Examples

Persona (p): *"A wanderlust-filled trader who appreciates and sells the artisan's creations in different corners of the world"*

Travel Filters (f): –
Popularity: Low
Interests: Nightlife Spot

Given this configuration, the pipeline generates a variety of queries:

- **Vanilla Query (qv):** "*Recommend off-the-beaten-path European cities with low popularity for a nightlife-focused trip with a mix of bars, clubs, and live music venues.*"
- **Persona-Specific Query 1 (qp0)**: "*Unique nightlife and cultural experiences in off-the-beaten-path European cities for a budget-conscious traveler interested in local artisans.*"
- **Persona-Specific Query 2 (qp1):** *"Which European cities offer a rich cultural heritage, historic centers, and local artisan markets to explore?"*

SynthTRIPS Dataset

Recent work published in SIGIR 2025!

Addressing the data scarcity in RecSys & LLM hallucination problem

We provide a structured database with factually grounded responses for a personalized user query written in natural language

Our data can be used for (offline) evaluation of RecSys models.

Paper available on arxiv: https://arxiv.org/pdf/2504.09277

SynthTRIPS

SynthTRIPs: A Knowledge-Grounded Framework for Benchmark Query Generation for Personalized Tourism Recommenders

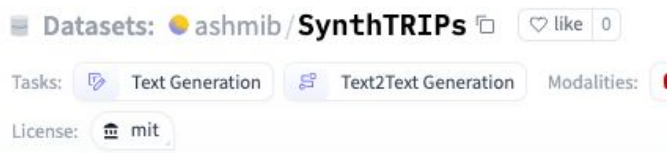
Ashmi Banerjee
ashmi.banerjee@tum.de
Technical University of Munich
Munich, Germany

Adithi Satish
adithi.satish@tum.de
Technical University of Munich
Munich, Germany

Fitri Nur Aisyah
fitri.aisyah@tum.de
Technical University of Munich
Munich, Germany

Wolfgang Wörndl
woerndl@in.tum.de
Technical University of Munich
Munich, Germany

Yashar Deldjoo
yashar.deldjoo@poliba.it
Polytechnic University of Bari
Bari, Italy



bitly

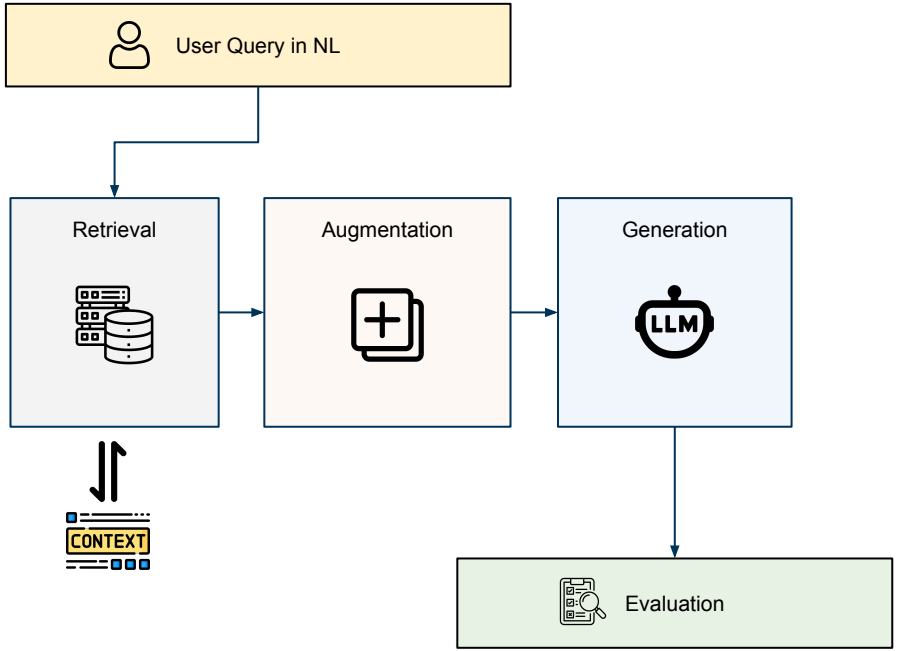
Part 2: Generating Recommendations using: Retrieval Augmented Generation with Sustainability Metrics

Based on the work:

Enhancing Tourism Recommender Systems for Sustainable City Trips Using Retrieval-Augmented Generation

Ashmi Banerjee, Adithi Satish, Wolfgang Wörndl, In Proceedings of Recommender Systems for Sustainability and Social Good (RecSoGood 2024). Springer Communications in Computer and Information Science, vol 2470 (co-located with ACM RecSys 2024), October 14-18, 2024 Bari, Italy.

Background: (Traditional) Naíve RAG



RAG - Retrieval Augmented Generation[1]

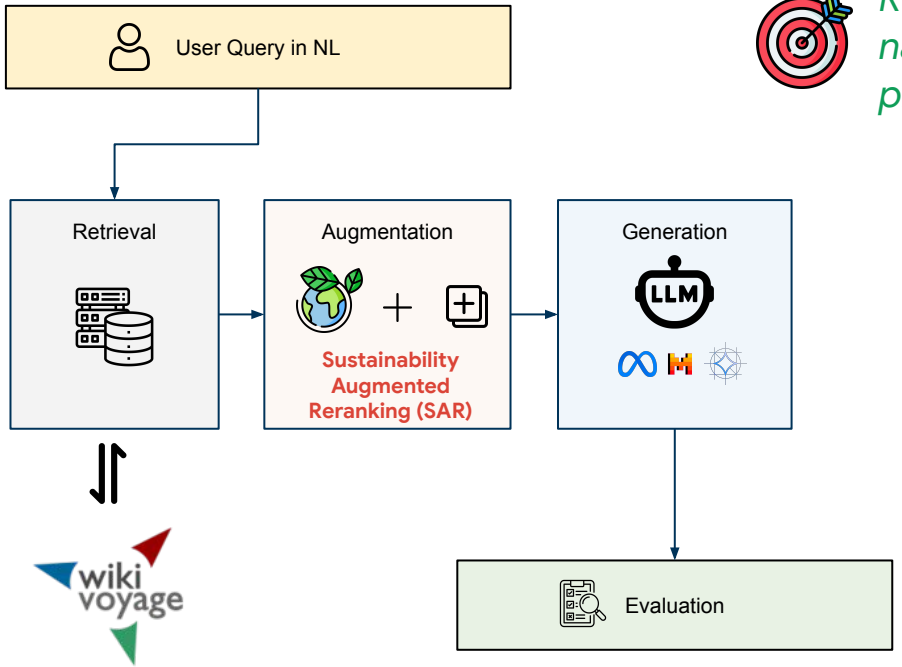
Idea: Provide the LLM with **relevant** context from external sources along with the prompt

Benefits:

- Frequent LLM knowledge updates
- No fine-tuning of weights
- Easier verification of information in the generated answer, enabling better detection of hallucination
- Not entirely black-box, leading to better interpretation

[1] Lewis et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474.

Our Approach: Overview



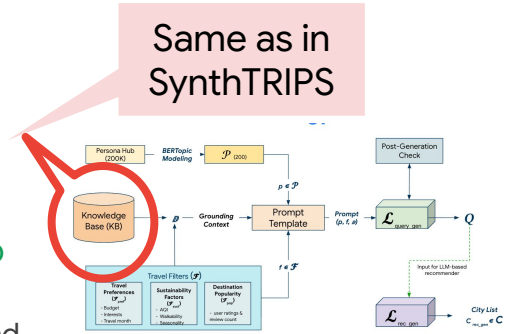
Recommend sustainable European cities based on natural language (NL) user queries on vacation planning [1].

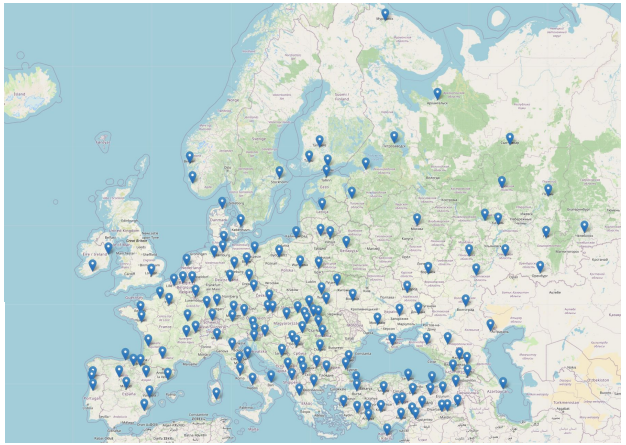
- Enhanced RAG which incorporates a sustainability metric in the augmentation phase (**S**ustainability **A**ugmented **R**eranking)
- Creation of a comprehensive **knowledge-base from Wikivoyage** with sustainability attributes
- **Generate sustainable travel recommendations** by integrating with two open-source LLMs

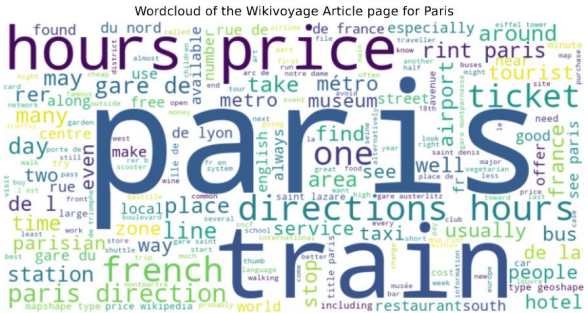
[1] Banerjee et al. (2024). Enhancing Tourism Recommender Systems for Sustainable City Trips Using Retrieval-Augmented Generation, RecSoGood Workshop co-located with RecSys 2024

Approach: Data Preparation

- **200 cities** spanning **41 countries** in Europe
  - Source: world cities database
- Knowledge Source: **Wikivoyage**
  - online travel guide hosted by Wikimedia Foundation
  - Wikivoyage Articles: XML dump of article pages for cities
    - Each page has an abstract and contains sections like “Get In”, “See”, “Do”, etc.
  - Wikivoyage Listings:
    - Contains specific names and descriptions of attractions, hotels and restaurants for each city





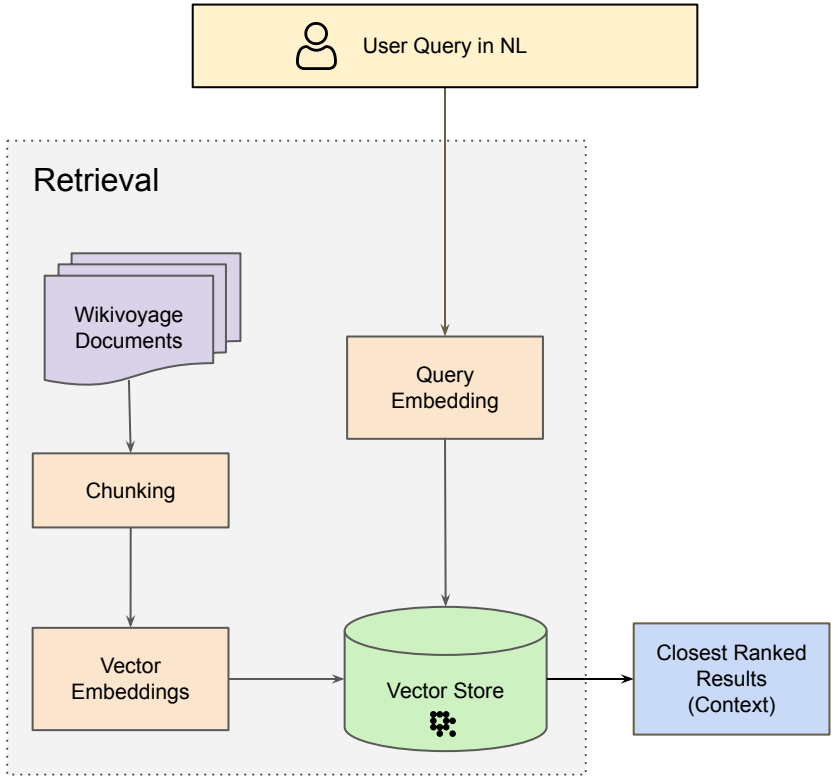


Approach: Information Retrieval

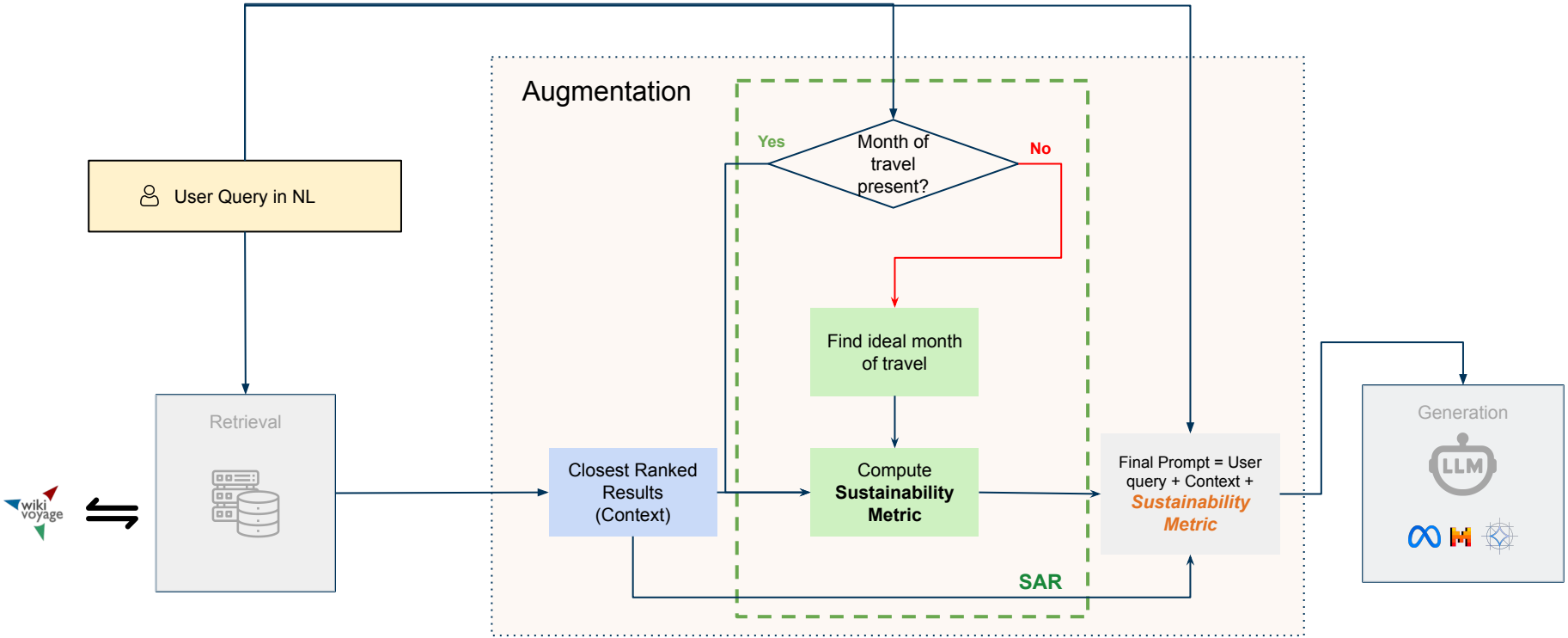
- **VectorDB:** LanceDB
- **Distance Measure:** Cosine Similarity
- **Embedding Model:** `all-MiniLM-L6-v2`
- Top-k-retrieval, where **k=10**

Chunking

- Done for efficient storage, retrieval, and to maintain context
- Each article has inherent structure (divided into sections)
  - Docs chunked according to section headers
  - Makes sure relevant information stored as part of the same chunk



Approach: Sustainability Augmented Reranking (SAR)



Approach: Sustainability Augmented Reranking (SAR)

Computing Sustainability Metric

$$\psi(c_i^j) = 0.334 \cdot \rho(c_i) + 0.385 \cdot \sigma(c_i^j)$$

- We use **S-Fairness Indicator**[1] as a measure of sustainability how sustainable the destination c_i is, for month *j*
  - Lower the value of $\Psi(c_i^j)$, the more sustainable the city
- Weighted combination of popularity (ρ) and seasonality (σ) scores for c_i for month *j*
  - ***Popularity Scores*** ➡️*Normalized POI count from Tripadvisor*
  - ***Seasonality Scores*** ➡️*Normalized monthly footfall from whereandwhen.net*
- Weights (computed through a user study) are taken from the original paper

[1] Banerjee et al.(2024). Modeling Sustainable City Trips: Integrating CO2 Emissions, Popularity, and Seasonality into Tourism Recommender Systems. Springer Information Technology & Tourism, pp.1-38.

Approach: Sustainability Augmented Reranking (SAR)

Putting it all together!

Text → default prompt for Baseline;
Text → prompt with additional sustainability information for SAR

System Prompt for Augmentation

You are an AI recommendation system.
Your task is to recommend European cities for travel based on the user's question. You should use the provided contexts to suggest the city best suited to the user's question.
You recommend a list of the top three most sustainable cities to the user and the best month of travel. If the user has already provided the month of travel in the question, use the same month; otherwise, provide the ideal month.
A sustainable city is defined as a city with low overall popularity and low footfall for the intended month of travel. Each recommendation should also explain why it is being recommended on sustainability grounds. The context contains a sustainability score for each city, also known as the S-Fairness indicator, along with the ideal month of travel. A lower S-Fairness value indicates that the city is a better destination for the month provided. A city without a sustainability score should not be considered. You should only consider the S-Fairness indicator values while choosing the best city. However, your answer should not contain the numeric score itself.
Your answer must begin with "I recommend, " followed by the city name and why you recommended it. Your answers are correct, high-quality, and written by a domain expert. If the provided context does not contain the answer, state, "The provided context does not have the answer."

Final Prompt

User Query in NL

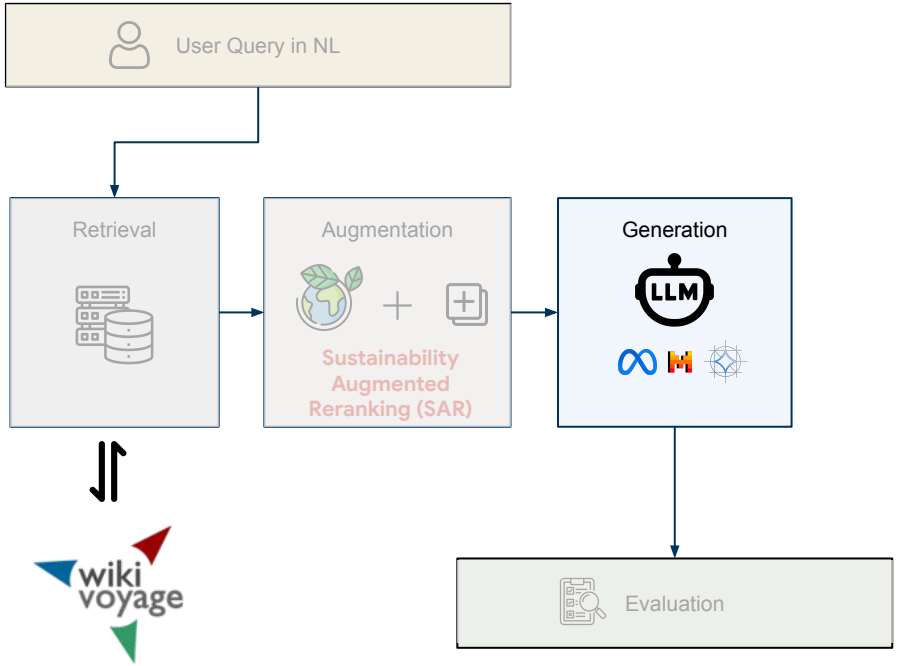
Closest Ranked Results (Context)

S-Fairness Indicators

Generation

LLM

Approach: Response Generation



- Three **open-source** LLMs of comparable size for reproducibility purposes
- All experiments run for **Baseline** & **SAR** on an Nvidia A40 GPU for 200 generated test prompts

	Llama-3.1-8B-it	Mistral-7B-it	Gemma-2-9B-it
# Parameters	8B	7.3B	9B
Context Length	128k tokens	8192 tokens	8192 tokens

Evaluation: Metrics

Answer Relevance: LLM-as-a-Judge

- Human evaluation of responses is not feasible
- LLM-as-a-Judge: helpful in assessing results in a human way, without incurring huge costs
- Grading scheme with instructions, user query and generated response passed to "*superior*" LLMs for score generation

Evaluation: Metrics

Answer Relevance: LLM-as-a-Judge

- Llama: higher scores vs. Mistral/Gemma
  - Mistral, particularly Gemma: struggles with quality
- SAR: consistent scores, maintains quality
- Mistral + GPT-4o-mini (SAR): marginal improvement
- SAR: preserves the overall quality & relevance of the answers to the question while considering sustainability
- Results with Claude-Sonnet-3.5 as judge - more comparable with GPT4

Models	Method	Answer Relevance (Mean ± SD)	
		GPT-4o-mini	Gemini-1.5-Pro
Llama3.1-8B-IT	Baseline*	8.16 ± 1.78	4.86 ± 2.56
	SAR	7.69 ± 1.73	4.85 ± 2.57
Mistral-7B-IT	Baseline*	3.85 ± 2.72	2.25 ± 2.50
	SAR	3.96 ± 2.65 ↑	2.23 ± 2.52
Gemma2-9B-IT	Baseline*	2.31± 1.42	0.57 ± 1.47
	SAR	2.88± 1.56 ↑	0.61 ± 1.57

*Baseline = without SAR

Evaluation: Metrics

Sustainability

Accuracy

How often does each model recommend the city with the lowest sustainability score as its top choice?

Frequency

How often is the top-choice the most sustainable option among the recommended cities?

- *While the models may not always prioritize the most sustainable cities, sustainability still plays a significant role in their reranking process.*

Models	Sustainability (%)	
	Accuracy	Frequency
	10.5 ↑	42.5
	7.5	36.5
	7.5	71.5 ↑

Evaluation: Metrics

Model Agreement & Faithfulness

Agreement

- Total agreement (Recommends same cities for same prompt)
  - Baseline: 25%
  - SAR: 22% ↓
- Partial agreement (Recommends at least *one* common city)
  - Baseline: 13.5%
  - SAR: 14.5% ↑

Faithfulness

Number of out-of-context responses

Models	Method	Faithfulness (%)
Llama3.1-8B-IT	Baseline*	0
	SAR	0
Mistral-7B-IT	Baseline*	14.0
	SAR	9.5
Gemma2-9B-IT	Baseline*	11.5
	SAR	11.5

*Baseline = without SAR

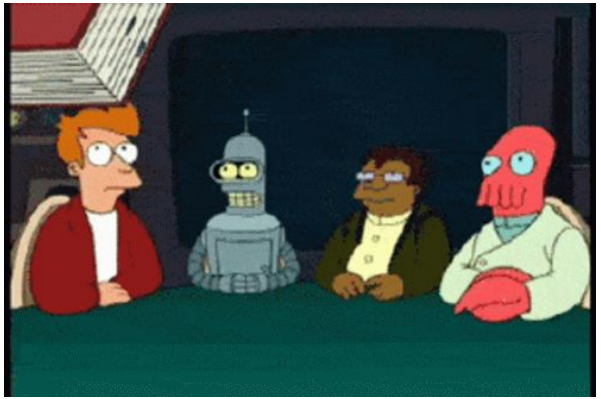
Final Takeaways

A verifiable knowledge-base of 200 European cities and their metadata which can be used as a KB for RAGs or other models

A list of 2302 synthetically generated travel queries that can be used for city recommendations.

Enhancing a naive RAG with SAR generally matches or enhances model performance, without compromising answer quality for TRS

However, that does not solve our lack of ground-truth problem entirely but is a step forward in that direction



Thank You! Time for Q&A!

Talk Feedback

https://bit.ly/ashmib-feedback

Scan me for feedback!

Ashmi Banerjee

PhD Candidate @ TU Munich,
ML GDE, WTM Ambassador
@ashmi_banerjee