

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)



ML Google Developer Expert & WTM
Ambassador
Loves traveling & outdoor activities 🏖️ 🚴 🏃

Introduction

LLMs in Tourism Recommender Systems (TRS)



Ability to generate explanations - improved interpretability through language generation








Often used as conversational question-answering system for recommendations

... bla bla bla !

Introduction

Challenges of using LLMs in TRS

-  Dynamic nature of tourism domain - data keeps changing
-  Fine-tuning LLMs on changing data is resource-intensive & computationally expensive
-  LLMs tend to hallucinate leading to misleading or factually inaccurate responses
-  Publicly available travel datasets are scarce
-  Recommendations often do not prioritize the environmental sustainability



Solution:

1. Retrieval Augmented Generation with Sustainability Metrics
2. Use a synthetically generated, factually grounded dataset tailored for sustainable TRS

Part 1: Retrieval Augmented Generation with Sustainability Metrics

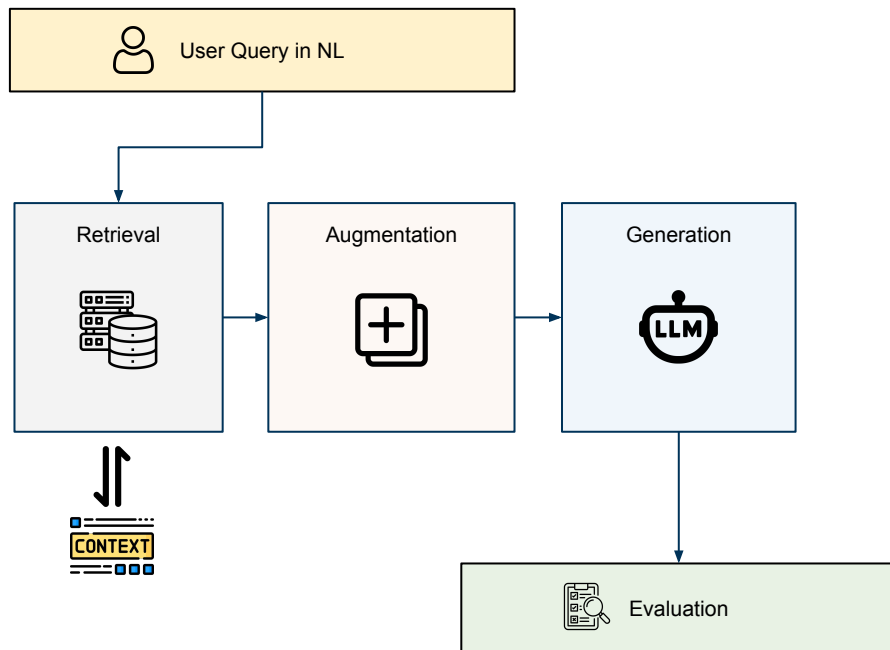


Based on the work:

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) Naïve RAG



RAG - Retrieval Augmented Generation^[1]

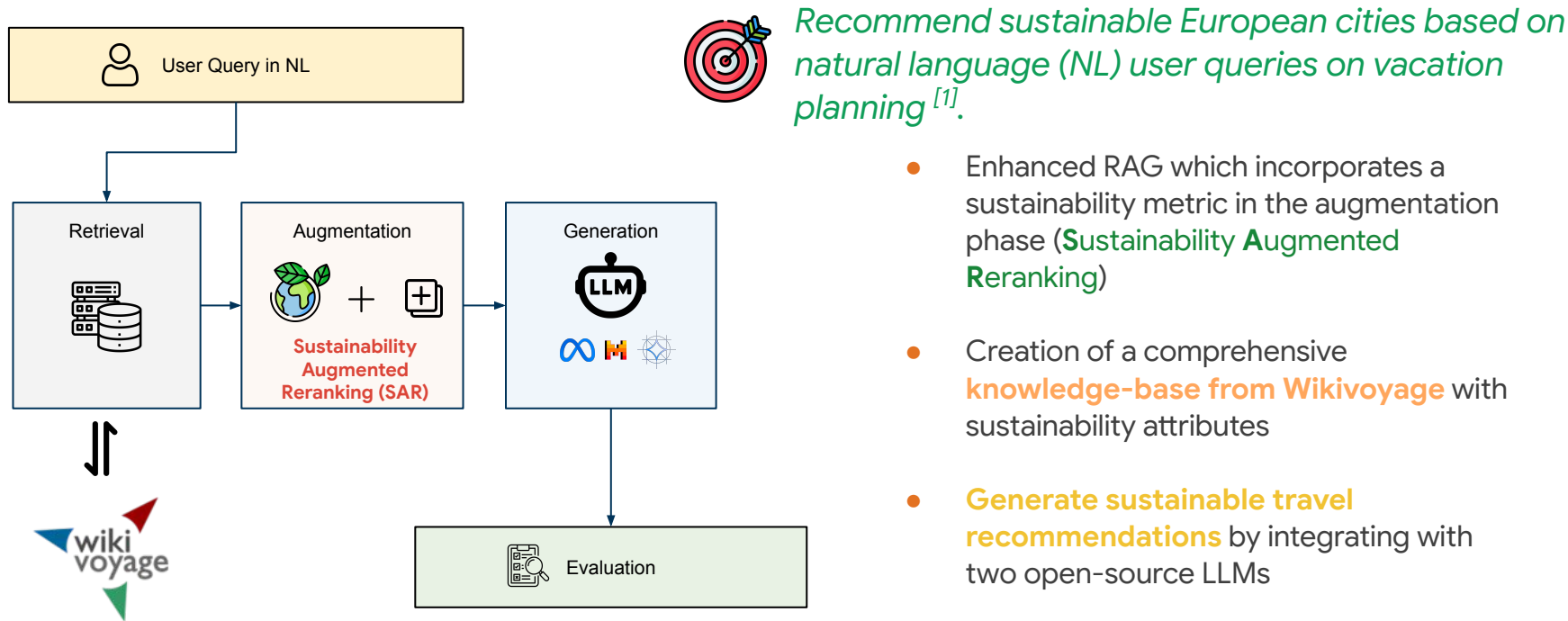
Idea: Provide the LLM with **relevant** context from external sources along with the prompt

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

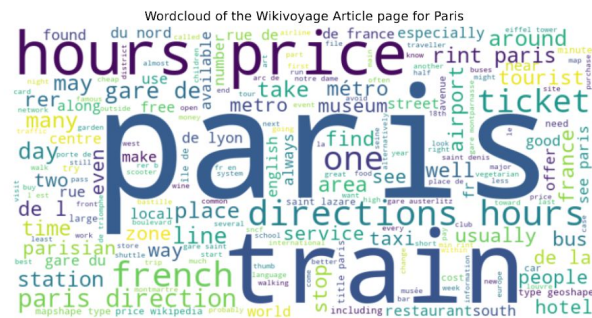
[1] Lewis et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.

Our Approach: Overview



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

-
- A detailed map of Europe and its surrounding regions, including North Africa, the Middle East, Greenland, and Iceland. The map is densely populated with blue location pins, indicating the sites of 100+ research projects. The pins are distributed across the entire continent, with a higher concentration in Western and Central Europe. The map also shows major cities, rivers, and geographical features.

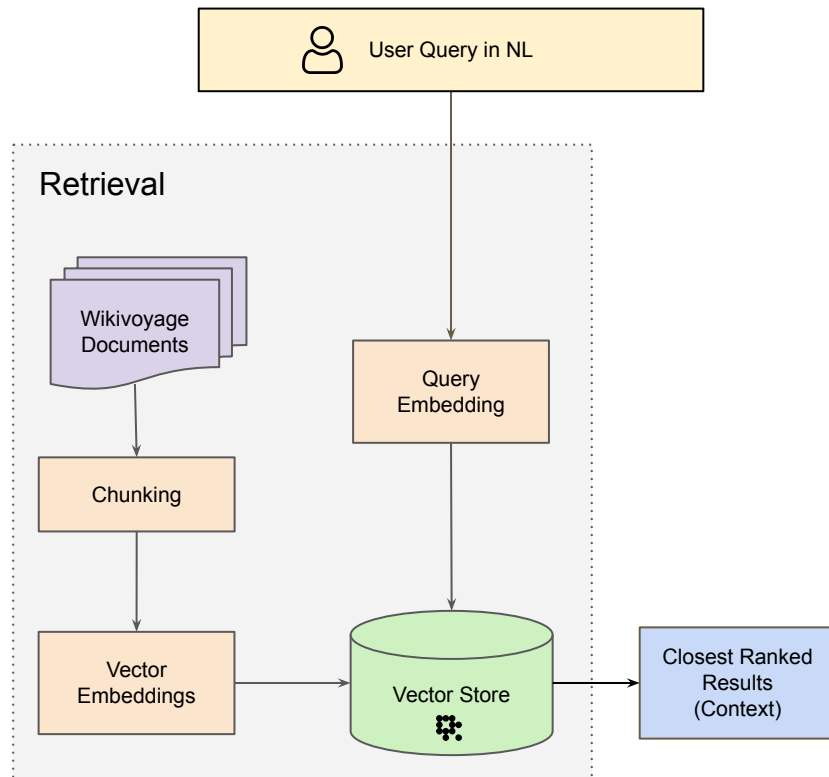


Approach: Information Retrieval

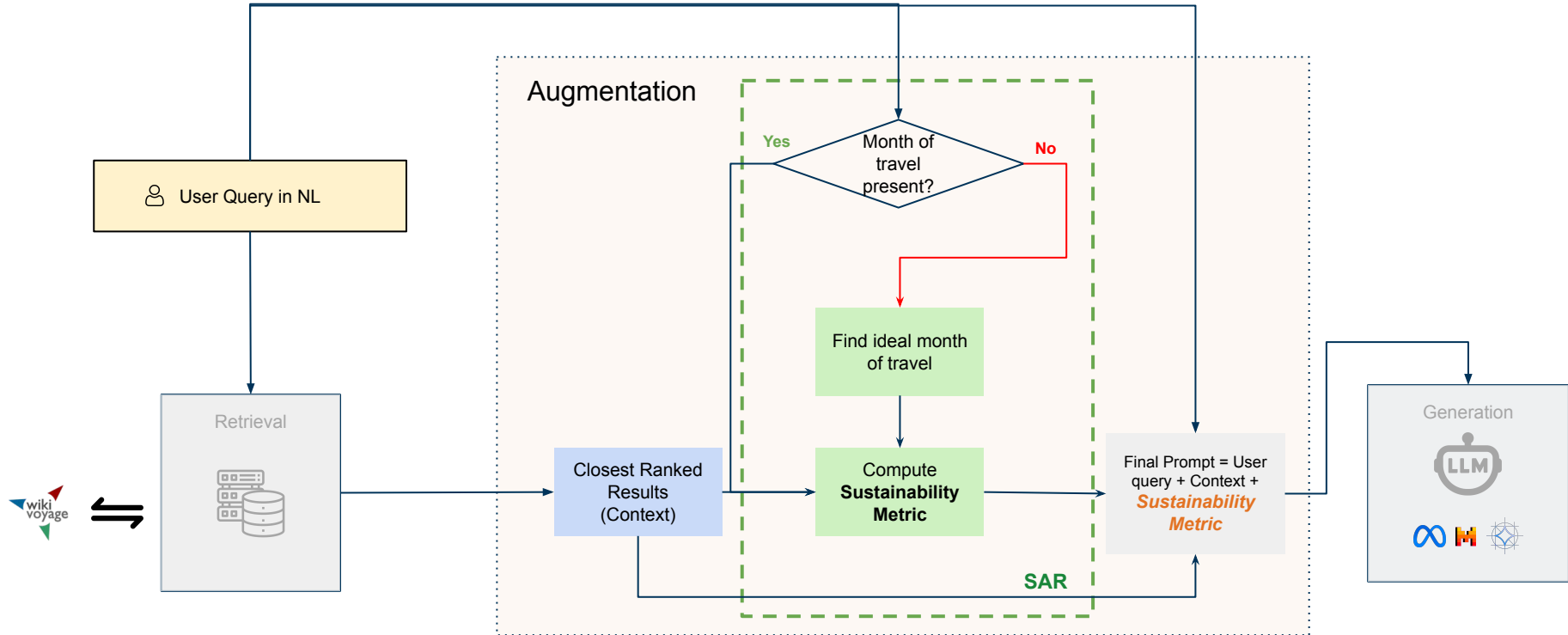
- **VectorDB:** LanceDB
- **Distance Measure:** Cosine Similarity
- **Embedding Model:** all-MiniLM-L6-v2
- Top-k-retrieval, where **k=10**

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

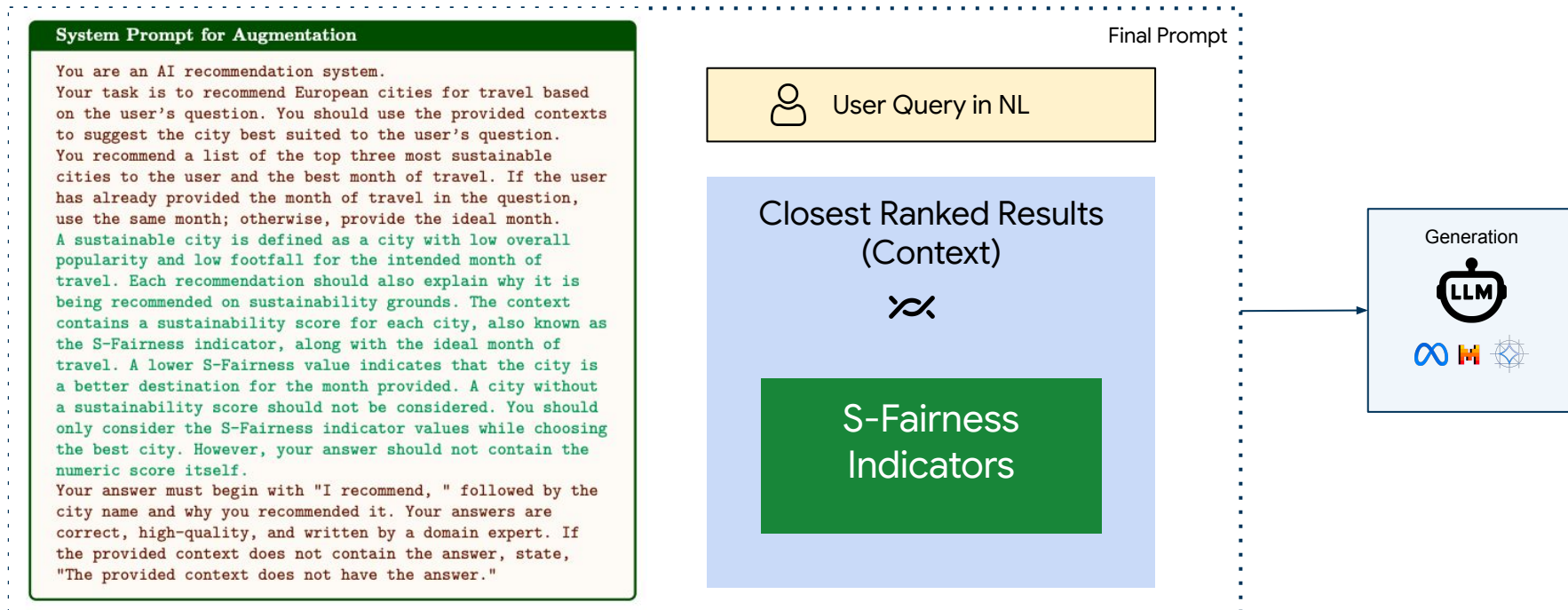
$$\psi(c_i^j) = 0.334 \cdot \rho(c_i) + 0.385 \cdot \sigma(c_i^j)$$

- 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

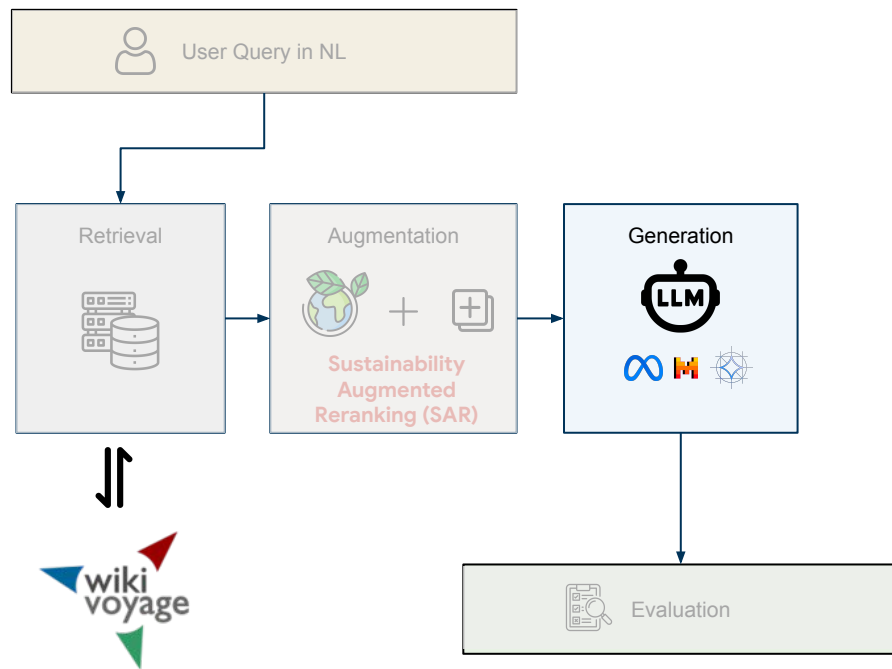
Approach: Sustainability Augmented Reranking (SAR)

Putting it all together!

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



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
	<u>10.5</u> ↑	42.5
	7.5	36.5
	7.5	<u>71.5</u> ↑

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



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



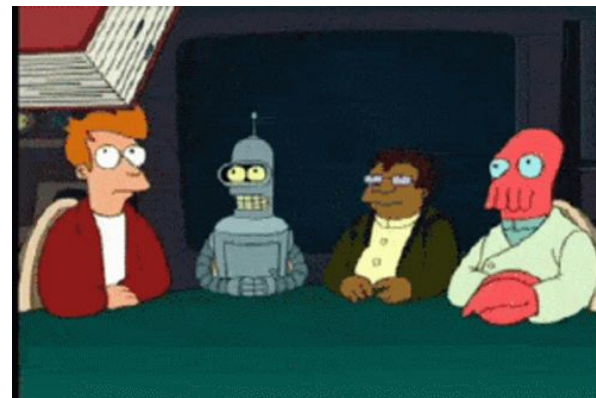
While Llama performs very well, Mistral and Gemma struggles with quality



Although “stronger” LLMs can be used as judges, at the moment, human evaluation is still necessary to ensure consistency and accuracy



LLM responses are also not very stable and can vary with time, so results are not very reproducible



Demo

Green City Finder

Note that this works best if you ask it for city recommendations.

Country

City
Select a city as your starting point.

Enter your preferences e.g. beaches, night life etc. and ask for your recommendation for European cities!

Ask for your city recommendation here!

Model
Select your model. The model will generate sustainable recommendations based on your query.

Your recommendations are sustainable with respect to the environment, your starting location, and month of travel.

Settings

Search

Clear

Cancel



<https://bit.ly/green-city-finder-ashmib>

Limitations & Future Work



Expanding the knowledge base to include more diverse, real-time datasets



Exploring additional sustainability metrics, such as carbon footprint and local economic impact



Adopting a conversational RecSys to address the cold-user problem & personalize the results



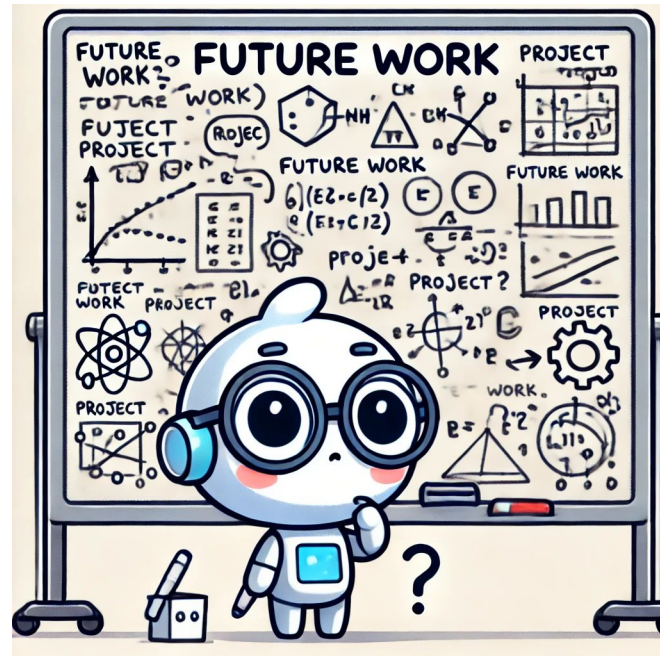
Examining the effects of various prompts and incorporating popularity and seasonality indices into the context



Although “stronger” LLMs can be used as judges, at the moment, human evaluation is still necessary to ensure consistency and accuracy



LLM responses are also not very stable and can vary with time, so results are not very reproducible



Part 2: Use a synthetically generated, factually grounded dataset



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.](#)

SynthTRIPS Dataset

Very recent work published in SIGIR 2025!

Addressing the data scarcity in RecSys & LLM hallucination problem

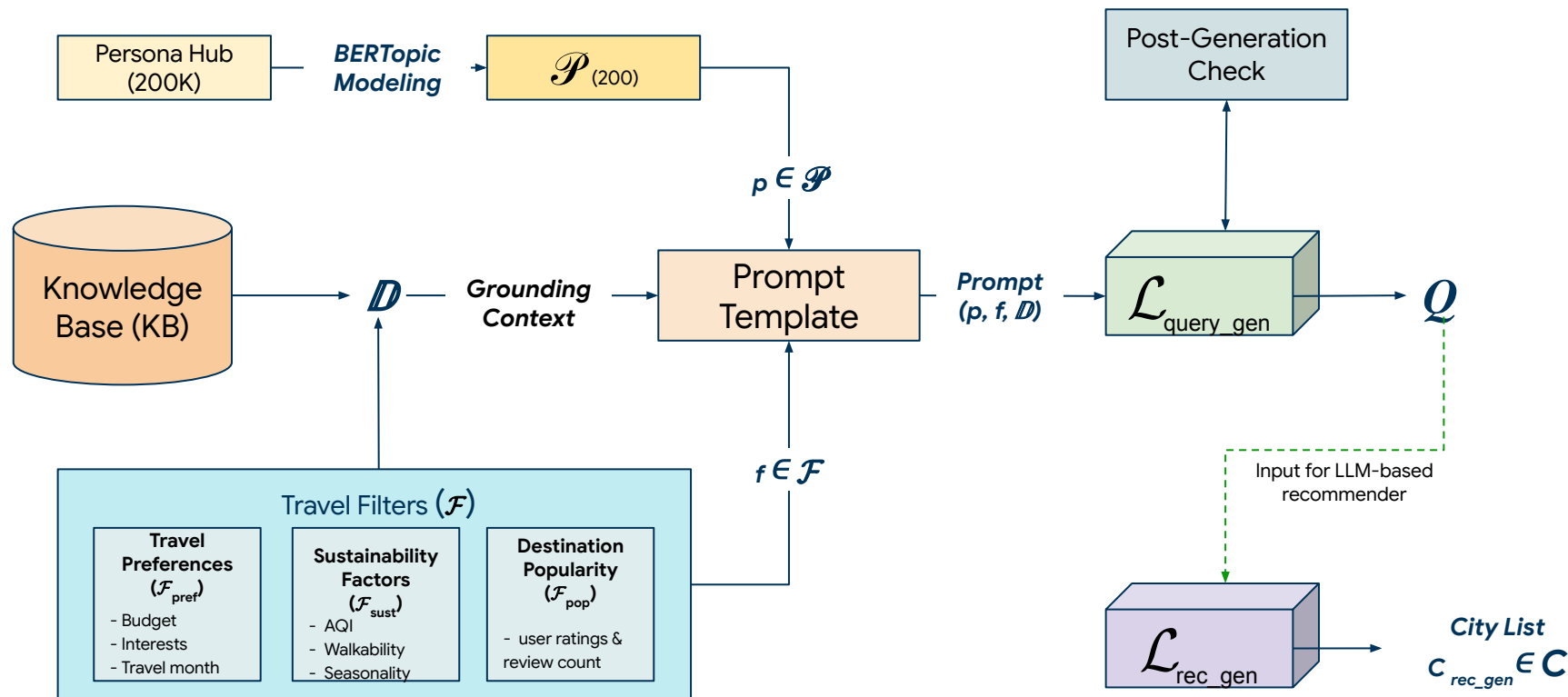
We provide a structured database with factually grounded responses for a personalized user query written in natural language

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

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



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 (q_v):** *"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 (q_{p0}):** *"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 (q_{p1}):** *"Which European cities offer a rich cultural heritage, historic centers, and local artisan markets to explore?"*

SynthTRIPS

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

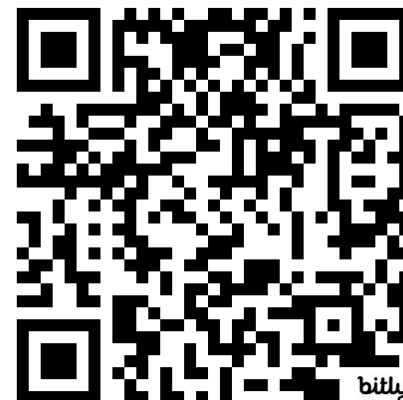
Ashmi Banerjee
ashmi.banerjee@tum.de
Technical University of Munich
Munich, Germany




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



Datasets:  ashmib / **SynthTRIPS**   like 0

Tasks:  Text Generation  Text2Text Generation Modalities: 

License:  mit



synthTRIPS-code

Public

Thank You! Time for Q&A!



Talk Feedback

<https://bit.ly/ashmib-feedback>



Ashmi Banerjee

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

Scan me for feedback!

