The Visa Wall: Benchmarking LLM Bias and Hallucinated Barriers Against Non-EU Applicants in the German Labor Market

Anh-Nhat Nguyen[2034311]

University of Mannheim, Germany

1 Introduction

The UNESCO World Heritage List documents the world's most treasured cultural and natural sites. However, systematic analysis of this domain remains difficult for researchers seeking to explore the collection programmatically. Information is often scattered across heterogeneous sources and lacks consistent metadata for provenance and geography, which makes data visualization and statistical analysis challenging.

This project addresses these limitations by developing an interactive Semantic Web application that provides flexible and reproducible access to World Heritage data. By putting Semantic Web ideas into practice, the application utilizes SPARQL and Linked Data principles to transform scattered, multilingual information into a coherent dataset. Ultimately, the objective is to offer a transparent resource that allows users to query sites by attributes such as country, type, or inscription century while retrieving clean datasets with summary statistics, supporting the tourism

2 Background

2.1 Dataset

To construct a robust Knowledge Graph (KG), this project integrates two principal open knowledge sources from the Linked Open Data cloud: Wikidata and DBpedia.

- **Wikidata (Primary Source):** [1] Wikidata serves as the structural backbone of the graph. Data is retrieved via the official Wikidata SPARQL endpoint to ensure verifiable access to structured attributes. This source provides the core entity definitions, including site names, heritage criteria, coordinates, and inscription years.
- **DBpedia (Supplementary Source):** [2] [Auer et al.(2007)Auer, Bizer, Kobilarov, Lehmann, Cyganiak, and Ives] DBpedia is employed to enrich the

[1] https://query.wikidata.org/

[2] https://dbpedia.org/sparql

dataset with unstructured textual context. Specifically, the project retrieves abstracts and descriptive comments to augment the structured data found in Wikidata.

The scope of the integrated dataset is designed to cover approximately 800 to 1,200 UNESCO sites worldwide. All data extraction is performed via official SPARQL endpoints, with results stored in standard RDF formats or CSV/JSON for downstream integration.

2.2 Data Retrieval

The final dataset is constructed using a four-step workflow: data retrieval, resource mapping, knowledge integration, and enrichment.

Extraction and Transformation Data is programmatically retrieved from SPARQL endpoints using Python libraries such as `SPARQLWrapper` and `rdflib`. Extracted RDF triples are transformed into an in-memory Pandas DataFrame using an ETL workflow, enabling vectorized queries and preprocessing including English-only label filtering and removal of null values.

Entity Alignment and Quality Control While `owl:sameAs` relations link many Wikidata and DBpedia resources, coverage is inconsistent. Therefore, a label-based entity alignment strategy is applied, matching normalized site names (e.g., "Historic Sanctuary of Machu Picchu") to ensure broader and more reliable integration.

During EDA, numeric attributes (e.g., area, inscription year) were found to default to zero when missing. To address this "0-Value Problem," the application suppresses unverified numeric fields in favor of trusted descriptive and geospatial data.

3 Knowledge Graph Integration (UNESCO: Wikidata ↔ DBpedia)

This task integrates UNESCO World Heritage Sites from Wikidata and DBpedia into a unified knowledge graph. Site entities are aligned across sources, high-confidence `owl:sameAs` links are created, and complementary attributes are consolidated into both a tabular view for data-science workflows and an RDF/-Turtle graph for graph-native consumption. The deliverables enable downstream querying, analytics, and UI/front-end development without revisiting the integration step.

3.1 Consolidated Source Tables from gathered data

From the Wikidata side, a site-level table (`wd_merged`) is constructed by starting with `unesco_basic` and left-joining comprehensive attributes (description, area, website, admin location, inscription date, Commons category), project attributes (heritage criteria & labels, inscription year), regional context (continent, continent label) and conservation fields (buffer zone, endangered status). From the DBpedia side, a site-level table (`db_merged`) is formed by joining the base UNESCO list with location/media (latitude, longitude, thumbnail) and criteria/region/area; the consolidated inscription year is `year_final`, preferring the criteria table's year if present. Result sizes and the final integrated view are summarized below.

Table 1. Site-level consolidation results (one row per site).

Table	Rows	Columns
Wikidata (`wd_merged`)	1,200	21
DBpedia (`db_merged`)	952	13
Integrated view (`unesco_integrated.csv`)	227	30

3.2 Entity Mapping Methodology

Labels are normalized to maximize exact matches:

$$\text{norm}(s) = \text{lowercase}(s) \text{ with } \{(), \_, , \} \text{ removed and whitespace normalized.}$$

Direct alignments are obtained by inner-joining `wd_merged.siteLabel` and `db_merged.name` on the normalized label. For Wikidata sites without a direct match, we compute TF–IDF vectors (1–2 grams) over concatenated texts: *WD* uses `siteLabel+description`; *DB* uses `name+criteria+region+whs`. Cosine similarity

$$\text{cosine}(x, y) = \frac{x \cdot y}{\|x\| \, \|y\|}$$

is evaluated between remaining WD and DB candidates; the best candidate is accepted if the score is at least 0.9. This strict threshold favors precision.

Table 2. Wikidata ↔ DBpedia alignment summary.

Match Type	Count	Notes
Direct (normalized label equality)	222	exact after normalization
Cosine TF–IDF (≥ 0.9)	5	high-confidence fallback
Total aligned pairs	**227**	retains `match_type`, `similarity_score`

3.3 Integrated Tabular View

Using the alignments, we assemble a unified table with identification and linking (`site_wd`, `site_db`, `wd_siteLabel`, `db_name`, `match_type`, `similarity_score`), geography and coordinates (`wd_country`, `wd_countryLabel`, `db_country`, `wd_continent`, `wd_continentLa bel`, `wd_coordinates`, `db_lat`, `db_long`), identifiers and years (`wd_unescoId`, `db_whs`, `wd_inscri ption Year`, `db_year_final`), criteria and extent (`wd_heritageCriteria`, `wd_heritageCri teriaLabel`, `db_criteria`, `wd_area`, `db_area`), protection and region (`wd_bufferZone`, `wd_endangred Status`, `db_region`), and presentation fields (`wd_descri ption`, `wd_image`, `db_thumbnail`, `wd_website`). The integrated table is saved as csv with 227 x 30 dimensions.

3.4 RDF/Turtle Export

The integrated table is exported as an RDF graph with Wikidata URIs as subjects, a custom UNESCO namespace for predicates, and explicit `owl:sameAs` links to DBpedia. The exported graph contains **4,230 triples** over **227** site subjects.

3.5 Validation and Parity Checks

We validated parity between the integrated CSV and the RDF export using `rdflib`. Subject coverage is perfect: the number of distinct `site_wd` in the CSV equals the number of subjects with `u:label` in the TTL (227/227), with no omissions. Attribute coverage parity is confirmed by comparing the number of non-null CSV values with the number of TTL subjects having each predicate. Representative predicates are shown below.

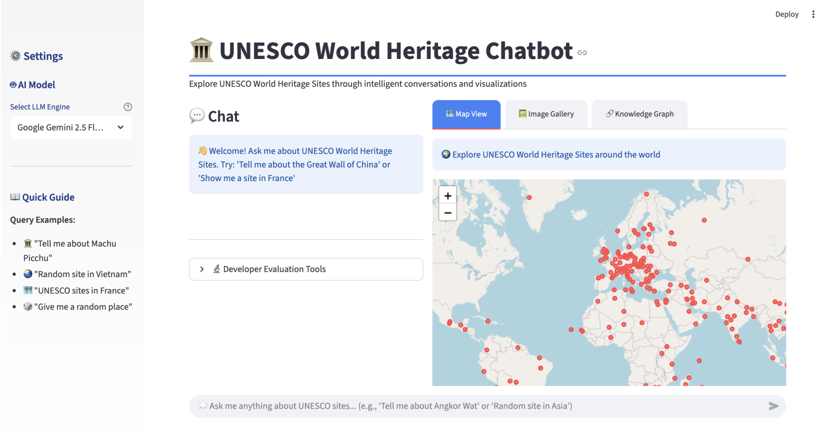
Predicate	CSV Non-Null	TTL Subjects
`u:label`	227	227
`u:countryLabel`	227	227
`u:unescoId`	226	226
`u:dbWHS`	115	115
`u:wdArea`	70	70
`u:dbArea`	37	37
`owl:sameAs`	227	227

Table 3. CSV ↔ TTL coverage parity for representative predicates.

Random spot checks compared CSV values and TTL objects for labels, country labels, inscription years, areas, descriptions, and `owl:sameAs` targets; all matched as expected, noting the intended normalization from full datetime to year for `u:inscriptionYear` and the numeric cast for areas and lat/long.

4 Application Development and Technical Implementation

This section details the technical architecture of the interactive application, focusing on the seamless integration of the frontend interface, local knowledge retrieval mechanisms, and generative AI logic.



🏛 UNESCO World Heritage Chatbot

Explore UNESCO World Heritage Sites through intelligent conversations and visualizations

⚙ Settings

AI Model

Select LLM Engine: Google Gemini 2.5 Fl...

📖 Quick Guide

Query Examples:

- 🏛 "Tell me about Machu Picchu"
- 🌏 "Random site in Vietnam"
- 🗺 "UNESCO sites in France"
- 🎲 "Give me a random place"

💬 Chat

👋 Welcome! Ask me about UNESCO World Heritage Sites. Try: 'Tell me about the Great Wall of China' or 'Show me a site in France'

🌍 Explore UNESCO World Heritage Sites around the world

Ask me anything about UNESCO sites... (e.g., 'Tell me about Angkor Wat' or 'Random site in Asia')

Fig. 1. Interactive application interface with AI and geospatial visualization

4.1 Frontend Architecture and Layout

The user interface was developed using `Streamlit`, enabling a responsive web interface in pure Python. As illustrated in Figure 1, the layout comprises three components:

Configuration Sidebar. The sidebar allows users to switch between inference engines (**Google Gemini 2.5 Flash** vs. **Google T5**) via dropdown, adapting to cloud or local environments without disrupting the chat workflow.

Conversational Interface. The chat window uses `st.session_state` to maintain dialogue context, enabling multi-turn conversations despite the framework's stateless nature.

Multi-View Visualization Panel. Adjacent to the chat, a dynamic dashboard visualizes retrieved data through a `Folium` map (auto-zooming to queried sites), an image gallery, and an interactive knowledge graph viewer.

4.2 Search Logic: ETL and Fuzzy Matching

To optimize for performance and user experience, the system utilizes a local retrieval strategy rather than relying on direct, high-latency SPARQL queries to DBpedia during runtime.

ETL Pipeline While the source data is maintained in **Turtle (.ttl)** format to preserve semantic standards, an Extract-Transform-Load (ETL) process is triggered at runtime. This process transforms the graph data into an in-memory `pandas` DataFrame, allowing for millisecond-level retrieval speeds for interactive queries.

Fuzzy Search Logic To address potential user input errors (e.g., typographic errors like "Angkor Watt"), the system implements the `rapidfuzz` library. By calculating token-set similarity ratios, the application identifies the intended site from the local dataset with high robustness, superior to strict database string matching.

4.3 Intelligence Layer: Intent Understanding and Generative Logic

The core intelligence logic leverages a **context-aware generative workflow**, enhanced by a **Dual-LLM strategy**. The process follows three steps:

- **Intent Recognition:** An LLM first classifies the user's intent (e.g., distinguishing between a request for a *Specific Site* versus a *Random Recommendation*).
- **Data Grounding:** Instead of relying solely on the model's training memory, structured data (Name, Location, Description) is dynamically retrieved from the local DataFrame and injected into the system prompt. This ensures the generated response is factually accurate and strictly based on the provided dataset.
- **Model Flexibility:** The architecture supports toggling between **Google Gemini 2.5 Flash** (via `google-genai` for high-performance cloud processing) and **Google T5** [Raffel et al.(2020)Raffel, Shazeer, Roberts, Lee, Narang, Matena, Zhou, Li, and Liu] (via `transformers` for commodity processing).

4.4 Hybrid Knowledge Graph Visualization

A planning feature is **Hybrid KG Viewer** to demonstrates system interoperability. By default, the viewer displays triples from the local dataset. Users can activate "Live Mode" to trigger real-time **SPARQL queries** to the **DBpedia endpoint**, retrieving enriched properties (abstracts, area measurements) not present locally, showcasing the architecture's extensibility to external Linked Open Data sources.

5 Experimental Setup

Evaluation 1: End-to-End Country Retrieval. We evaluated the complete pipeline, from intent recognition to response generation, using a dataset of **13**

country-specific queries. Each query used an ISO-style code (for example, `C_DE` for Germany, `C_FR` for France, `C_VN` for Vietnam). The system returned 15 candidate heritage sites, which were manually checked against the official UNESCO World Heritage List.

Binary ground truth labels (`is_correct`) were assigned using strict UNESCO designation criteria. For example, the reconstructed Roman fortress *Limeskastell Pohl* is not part of the UNESCO list for Germany. The correct associated heritage site is *Frontiers of the Roman Empire*, which refers to the various Roman Limes.

Evaluation 2: Similarity-based Entity Matching. To evaluate the robustness of Knowledge Graph alignment, we constructed a dataset of **15 Wikidata and DBpedia label pairs**. The model predicted equivalence using a cosine similarity threshold of **0.9**.

The dataset also contained hard negatives, which are pairs with high textual similarity but different semantic scope. For example, the model must correctly separate the UNESCO site *Budapest, including the Banks of the Danube* from the general encyclopedic topic *History of Budapest*. We report Precision, Recall, and F1-score against manually assigned ground truth labels (`gt_same`).

6 Results & Discussion

Retrieval Performance (Eval 1) Across 13 country-level (15 entities) test queries, the system achieved an **Overall Precision of 0.7333** (see Appendix 4 for breakdown).

6.1 Evaluation Results

Country-level Retrieval Performance. The system demonstrated exceptional reliability for 10 out of 13 queries (e.g., `C_CA`, `C_UK`), achieving **1.0 (100%) precision**. Overall, 11 of the 15 retrieved candidates were validated as correct, confirming the effectiveness of the LLM-based intent extraction. However, performance degraded in specific edge cases. Precision dropped to **0.5** for Germany due to noise introduced by the sampling strategy for broad queries. Furthermore, queries for Australia and France yielded **0.0 precision**. This failure was traced to label mismatches in the knowledge graph; discrepancies between user inputs (e.g., “France”) and formal literals (e.g., “French Republic”) caused the substring matching logic in `_query_kg_by_intent` to fail.

Entity Matching Robustness. The similarity-based matcher proved highly effective. At a threshold of $\tau = 0.9$, the system identified **13 True Positives** and **2 False Positives**, resulting in a **Precision of 0.87** and an **F1-Score of 0.93**. Crucially, the system achieved a **Recall of 1.00** (zero False Negatives). The lower precision means that a few incorrect matches were included. However, this is acceptable because it is better to show extra candidates than to accidentally exclude correct heritage sites.

6.2 Limitations

This work faces several key limitations that affect system accuracy and completeness.

Static Dataset and Outdated Information. The knowledge graph relies on a static `unesco_integrated_full.ttl` snapshot rather than real-time UNESCO API data. Newly inscribed sites, delisted entries, or status changes are not reflected without manual re-ingestion, preventing the system from serving as a current reference.

Missing and Invalid Numeric Data. Many numeric fields such as `area` and `inscriptionYear` contain zero values when data is actually missing from Wikidata. For example, a site might show an area of 0 square kilometers, which is clearly incorrect. These zeros substitute for NULL values in the crowd-sourced extraction process. To avoid confusion, our application suppresses these unreliable fields and prioritizes textual descriptions. While "Live Mode" attempts to retrieve valid values from DBpedia, coverage remains inconsistent.

Label Matching Failures. Entity alignment fails when user inputs do not match formal names in the graph. For instance, searching "France" fails if the graph stores "French Republic" instead, causing 0.0 precision for France and Australia queries. The TF-IDF similarity fallback handles spelling variations but cannot resolve systematic naming differences.

Limited Model Flexibility. While supporting both Gemini 2.5 Flash and Google T5, only Gemini provides acceptable accuracy. T5 lacks the reasoning capability for reliable intent classification, making cloud-based Gemini effectively mandatory and requiring stable internet connectivity.

Limited Evaluation Scope. Testing covered only 13 country queries and 15 entity pairs. This small sample may not represent performance on rare sites, multilingual queries, or challenging inputs. No user study assessed real-world usability.

6.3 Future Work

Future work should address three key improvements. First, implementing live synchronization with the official UNESCO API would eliminate reliance on static snapshots. Second, developing entity normalization with country name mappings (e.g., "France" and "French Republic") would improve retrieval robustness, as evidenced by our label mismatch failures. Third, integrating a hybrid retrieval-augmented generation architecture combining SPARQL queries with vector-based semantic search could overcome exact string matching. Our experience demonstrates that even high-quality Linked Open Data sources require substantial preprocessing and alignment logic, suggesting future semantic web systems should prioritize data quality validation and fuzzy matching from the outset.

7 Appendix

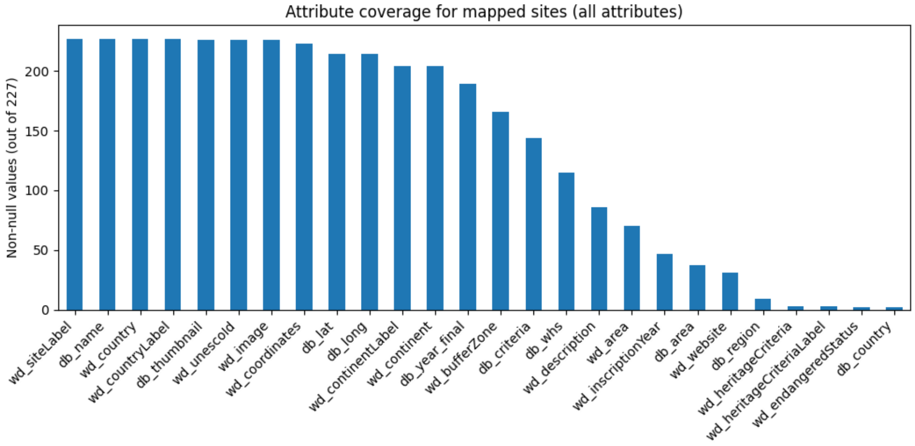


Fig. 2. Coverage for mapped sites. only approx. first 12 attributes might be relevant due to coverage

Table 4. Country-level retrieval precision for Evaluation 1.

Query	AU	CA	CH	CN	DE	ES	FR	GB	IS	JP	MX	VN	ZA	Overall
Precision	0	1	1	1	0.5	1	0	1	1	1	1	1	1	0.73

References

Auer et al.(2007)Auer, Bizer, Kobilarov, Lehmann, Cyganiak, and Ives. Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *international semantic web conference*, pages 722–735. Springer, 2007.

Raffel et al.(2020)Raffel, Shazeer, Roberts, Lee, Narang, Matena, Zhou, Li, and Liu. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Declaration of Authorship

I hereby declare that I have prepared this report for project alignment independently, without the assistance of third parties and without using any sources or aids other than those indicated.

Declaration of Used AI Tools

Tool	Purpose	Where?	Useful?
ChatGPT - 5	Academic rephrasing	Report writing	++
DeepL	Translation	Report writing	+++
Toggle Writefull	Grammar check in Overleaf	Report writing	+
Gemini	Code debugging support	Coding	+++

Signature
Mannheim, November 17, 2025