RAG System Evaluation Report

Date: 2026-02-15 15:21
Total Queries Evaluated: 22

Query Set Design

This evaluation uses a fixed query set stored at `src/eval/query_set.json`.

- Query type counts: {'direct': 11, 'edge_case': 5, 'synthesis': 6}
- Difficulty counts: {'easy': 2, 'hard': 9, 'medium': 11}

The set includes direct questions, cross-source synthesis questions, and edge cases that should trigger explicit missing-evidence behavior.

System Configuration (This Run)

- Generator model: gemini-2.5-pro
- Retrieval configuration snapshot: {'k': 10, 'k_raw': 60, 'use_hybrid': True, 'use_reranking': True, 'top_k_after_rerank': 10, 'vector_weight': 0.5, 'bm25_weight': 0.5}

Metrics

Groundedness is judged against retrieved evidence. Citation precision checks whether citations in the answer match retrieved chunks. Answer relevance rates whether the answer addresses the question.

Implementation Notes

Retrieval uses a hybrid of semantic vector search and lexical BM25. Answers are generated with strict citation constraints that only allow citing retrieved chunks, and a repair pass runs when the model outputs citations outside the allowed set.

Summary Metrics (This Run)

Metric	Average Score	Notes
Groundedness	3.95/4	LLM-judged faithfulness to retrieved evidence
Citation Precision	100.00%	Fraction of citations that match retrieved chunks
Answer Relevance	3.59/4	LLM-judged relevance to the question

Enhancement Impact (Baseline vs Enhanced)

Baseline settings: vector-only retrieval, no LLM reranking.

Enhanced settings: hybrid retrieval (BM25 + vector) plus LLM reranking.

Metric	Baseline	Enhanced	Delta
Groundedness (avg /4)	4.00	3.95	-0.05

Metric	Baseline	Enhanced	Delta
Citation Precision (avg)	100.00%	100.00%	+0.00%
Answer Relevance (avg /4)	3.32	3.59	+0.27

Enhancement Impact by Query Type

```
| query_type    |   groundedness_baseline | citation_precision_baseline   |   relevance_baseline |   groun
```

|:-------------|------------------------:|:------------------------------|---------------------:|------------------------:|:------------------------------|-------------------
---:|---------------------:|:---------------------------|------------------:|
| direct | 4 | 100.00% | 3.27 | 3.91 | 100.00% | 3.55 | -0.09 | 0.00% | 0.27 |
| edge_case | 4 | 100.00% | 2.8 | 4 | 100.00% | 3.2 | 0 | 0.00% | 0.4 |
| synthesis | 4 | 100.00% | 3.83 | 4 | 100.00% | 4 | 0 | 0.00% | 0.17 |

```
## Breakdown by Query Type
```

query_type	groundedness_avg	citation_precision_avg	relevance_avg	count
direct	3.90909	1	3.54545	11
edge_case	4	1	3.2	5
synthesis	4	1	4	6

Per-Query Summary

query_id	query_type	groundedness_score	citation_precision_value	answer_relevance_score
Q01	direct	4	100.00%	4
Q02	direct	4	100.00%	3
Q03	direct	4	100.00%	4
Q04	direct	4	100.00%	3
Q05	direct	4	100.00%	4
Q06	direct	4	100.00%	4
Q07	direct	3	100.00%	4
Q08	direct	4	100.00%	4
Q09	direct	4	100.00%	3
Q10	direct	4	100.00%	2
Q11	synthesis	4	100.00%	4
Q12	synthesis	4	100.00%	4

query_id	query_type	groundedness_score	citation_precision_value	answer_relevance_score
Q13	synthesis	4	100.00%	4
Q14	synthesis	4	100.00%	4
Q15	synthesis	4	100.00%	4
Q16	edge_case	4	100.00%	4
Q17	edge_case	4	100.00%	4
Q18	edge_case	4	100.00%	4
Q19	edge_case	4	100.00%	2
Q20	edge_case	4	100.00%	2
Q21	direct	4	100.00%	4
Q22	synthesis	4	100.00%	4

Best Performing Queries

Q01: What architectural components are critical for operational city digital twins?

- Groundedness: 4 | Citation precision: 100.00% | Relevance: 4

Q03: How is real-time sensor data integrated into digital twin platforms?

- Groundedness: 4 | Citation precision: 100.00% | Relevance: 4

Q05: What are the most common use cases for city digital twins in urban planning?

- Groundedness: 4 | Citation precision: 100.00% | Relevance: 4

Failure Cases (Representative)

Q07: What evaluation protocols are used to validate digital twin predictions?

- Groundedness: 3 | Citation precision: 100.00% | Relevance: 4
- Grounding note: The answer is comprehensive and very well-structured, with nearly all claims directly supported by the cited evidence. It accurately synthesizes information from multiple sources. However, there is one minor unsupported detail where the answer characterizes datasets as "real-world and synthetic" when the source document does not provide this specific characterization.
- Invalid citations: []
- Evidence snippets:
    - (digital_twin_citiesframework_and_global_practices_2022, chunk_91): standardization projects Standards and evaluation Standards and evaluation are the regulations followed in the construction of digital twin cities; they are also important references for the evaluation of digital twin city project operations. In terms of standards, digital twins have received attention from the ISO (International Organization for Standardization); the standardization of digital twin cities in China has officially begun. The National Information Technology Standardization Technical Committee, China's Digital Twin Alliance, the Digital Twin Technology Application Working Committee of the Internet Society of China, the CAICT and other organizations have proposed application scenario-based targets for the standardization of digital twin technologies, taking into account their
    - (advancing_intra_and_inter_city_urban_digital_twins_2024, chunk_58): rks for evaluation and comparison. Such efforts include assessing the technological and societal readiness levels of UDTs based on the implementation phase and

public acceptance (Botín-Sanabria et al. 2022). However, these measures are often qualitative, leading to potential subjectivity and difficulty in inter-city comparison. Another work introduced a system of city service indicators named Digital Twin DNA, developed based on ISO standards and allowed city service comparison across different cities (Badawi, Laamarti, and El Saddik 2021). Yet, these indicators often lack real-time updates and may not provide sufficiently fine-grained measurements outside strategic planning. There are also efforts in creating standards for data security to verify the authenticity of digital twin data, suc

Q10: What methods are used for uncertainty quantification in digital twin predictions?

- Groundedness: 4 | Citation precision: 100.00% | Relevance: 2
- Grounding note: The answer accurately extracts all relevant information from the provided evidence regarding uncertainty in digital twins. Each claim is directly supported by the cited source, covering the use of entropy as a metric, uncertainty in LLMs, and uncertainty arising from data collection. The answer correctly concludes by noting the limited scope of the information available in the provided texts.
- Invalid citations: []
- Evidence snippets:
  - (arxiv_2501_02842v1, chunk_54): [109]. More advanced paradigms involve the analyze of knowledge boundary [110] and the estimation of prediction uncertainty in LLMs [106, 111]. Theoretically speaking, since the study of when to retrieve shares similar motivations and foundations with the study of hallucination detection, existing studies on LLM hallucination [112, 113] could provide important inspiration for research on this topic. Promising directions including better fact checking systems for LLMs [114] and more investigations on how to characterize the confidence and uncertainty of LLM predictions based on both external behavior and internal state analysis [111]. The question of what to retrieve focuses on analysis the intents and information needs of LLMs in inference. LLMs often need the help of different tools and s
  - (towards_human_centric_digital_twins_leveraging_com_2023, chunk_36): to measure, which brings uncertainty into the model development as well as in the output and analysis (more discussion will be provided in Sections 5 and 7). Another missing point of the data collection is micro-climate variables(e. g., wind, temperature, rainandhumidity).Thosemicro-climate variables commonly require professional-grade sensors (e. g., weather stations) for the collection. At the time of writing this article, we have not identified any cheap and portable devices that can be easily used to collect those data; therefore, we consider such data collection beyond the scope of promoting crowdsourced data and accessible research. Further discussion will be provided in Sections 6 and 8. As a designing 2 https://cozie-apple. app/ 3 https://meters. uni-trend. com/product/ut 353-ut 3

Q19: What is the typical computational cost of running a city-scale digital twin?

- Groundedness: 4 | Citation precision: 100.00% | Relevance: 2
- Grounding note: The answer correctly states that a specific computational cost is not provided in the evidence. It then accurately summarizes the factors that contribute to the cost, such as data processing and hardware, and provides correctly cited, grounded examples for each point. All claims made in the answer are directly supported by the provided text.
- Invalid citations: []
- Evidence snippets:
  - (challenges_of_urban_digital_twins_a_systematic_rev_2023, chunk_80): ution which currently does not exist'. It is still unknown 'where the users are' and 'who is willing to pay'. Moreover, one expert states that it is hard to process massive data with the high cost and limited manpower for simulation use cases. The last stage is updating. It is not only a phase to detect changes and make updates but also a step to complete and restart the life cycle of urban digital twins. Therefore, technical challenges are mostly regarding hardware, version management and reconstruction — what new data should be obtained, and what new infrastructure should be installed. For instance, it challenges the 'efficient update process to align the digital representation with the changes happening in the real world' and 'revise workflow to update digital twin models regularly'. Fr
  - (challenges_of_urban_digital_twins_a_systematic_rev_2023, chunk_37): eneous techniques – Incompatible system Data complexity [1,86,91,100,101] – Dynamic activities – Overloaded information Software [85,93,99,102] – License Hardware [103,104] – Connectivity Update and versioning [86,88,105–107] – Version management – Latency – Cost survey. Since digital twins are dynamic and interactive, we consider the lifecycle a suitable way to map challenges in different phases described in Section 2.2. Such mapping enables a clear representation of our results and helps practitioners better understand the state of the art of digital twins in the urban and geospatial domain. 4. Results 4.1. Reviewing challenges from the literature We analysed the 34 papers to provide a broader view of the trend by investigating the relation between the number of publications and the publ

Run Logs

Machine-readable runs: outputs/eval/eval_runs_enhanced_20260215_145315.jsonl

Limitations and Next Steps

The evaluation uses an LLM judge, so scores can be noisy. For follow-up, consider adding deterministic retrieval metrics, expanding edge cases for missing evidence, and validating citation alignment against longer evidence spans.