1. Domain

City digital twins, Urban AI, and spatial intelligence for decision support in planning and governance.

2. Main question

How do city digital twins enable evidence-backed decision support, and what types of empirical evidence, evaluation designs, and uncertainty practices make their claims verifiable?

3. Sub-questions

1. How do the selected sources define a city digital twin?
2. What system architectures and data pipelines are commonly used in city digital twins, and which components are critical for operational decision support?
3. What benchmark datasets, benchmark tasks, and evaluation protocols exist for city digital twins and spatial intelligence components, and what are the most common validity threats in these benchmarks?
4. Which real-world use cases are most frequently reported for city digital twins, and what evidence is provided for decision impact in planning, operations, resilience, mobility, energy, and public safety?
5. How is new data integrated over time, including streaming sensors, administrative records, remote sensing, and citizen-generated data, and what update mechanisms support model recalibration, drift detection, and versioning?
6. What human-centered design practices are used in digital twin interfaces and workflows, and what evidence shows improved usability, trust, interpretability, and adoption among planners, operators, and community stakeholders?

4. Scope

Included:

- Peer-reviewed papers, reputable technical reports, and standards that describe city digital twins, spatial intelligence methods in urban contexts, and evaluation methodologies.
- Evidence that can be traced to specific text passages (chunked excerpts) and supports a claim about methods, results, limitations, or design tradeoffs.

Excluded:

- Purely visionary essays without verifiable claims.
- Marketing materials without technical detail or evaluable results.
- Claims that are common knowledge but not stated in the provided excerpts.

5. Tasks

Task 1: Claim–evidence extraction
Required output:
5 rows with columns Claim | Direct quote or snippet | Citation (source_id, chunk_id)

Task 2: Cross-source synthesis
Required output:
a table with columns Agreement | Disagreement | What evidence supports each side
Every major statement in the table must include citations to the provided chunks.

6. Test cases and evidence packets

Phase 1 uses four fixed test cases. Each test case is a self-contained packet of chunked excerpts. Chunks are labeled as (source_id, chunk_01 ... chunk_N) and preserve the extracted text so quotes can be verified exactly.

Where the packets live (repo paths):
- source/test case/CE-1_UDT_UseCases_A.md
- source/test case/CE-2_UDT_Benchmarks_B.md
- source/test case/CS-1_UDT_Participatory_C_vs_UDT_Participatory_D.md
- source/test case/CS-2_UDT_Definition_E_vs_UDT_Definition_F.md

Test case mapping:
- CE-1 (Claim-evidence extraction): single-source packet from UDT_UseCases_A (use cases and decision-impact evidence).
- CE-2 (Claim-evidence extraction): single-source packet from UDT_Benchmarks_B (evaluation and benchmarking practices).
- CS-1 (Cross-source synthesis): two-source packet comparing UDT_Participatory_C vs UDT_Participatory_D (participation and workflow evidence).
- CS-2 (Cross-source synthesis): two-source packet comparing UDT_Definition_E vs UDT_Definition_F (definitions, scope, and uncertainty practices).

Citation resolution rule used in Phase 1:
- A citation is valid if (source_id, chunk_id) maps to a unique chunk header in the corresponding packet and the quoted snippet appears verbatim in that chunk.
- If the packet does not contain evidence for a claim, outputs must state that no evidence is found in the provided excerpts and should cite the closest relevant chunk where possible.

7. Prompt variants

For each task:
- Prompt A: baseline structured instruction with required format.

- Prompt B: improved prompt with guardrails, citation checks, and missing-evidence handling.

8. Models to compare

Model_A: GPT5.2 Thinking
Model_B: Gemini 3.0 Pro

9. Evaluation plan

Total MVP runs: 16
Combinations: 2 tasks × 2 test cases per task × 2 prompt variants × 2 models.

Rubric dimensions (1 to 4):

- Groundedness or faithfulness: claims are supported by the provided excerpts.
- Citation correctness: citations resolve to the right source and chunk, and the cited chunk contains the quoted snippet.
- Usefulness: output meets the required format and helps a researcher.

1. Patterns observed across runs

1.1 Citation traceability improved with explicit audits, but content quality did not always improve

Prompt B generally increased traceability when it required a citation audit. When the model explicitly restated citations and confirmed quotes, outputs were easier to verify and less likely to include fabricated support. However, stronger guardrails sometimes reduced informativeness. In CE runs, Prompt B occasionally encouraged the model to select the safest, easiest-to-verify sentences, increasing redundancy and decreasing coverage across the document.

A recurring tradeoff emerged: **better verifiability can lead to lower information density** when the prompt does not also constrain claim usefulness and diversity.

1.2 Chunk coverage was a strong predictor of usefulness

Outputs that drew evidence from a wider range of chunks produced more useful research artifacts. For example, in CE-1 and CE-2, responses that cited 4 to 5 distinct chunks typically captured a more complete picture of methods, scope, and implications. Responses that relied heavily on a single “intro” chunk tended to produce repetitive or overly generic claims. This suggests an actionable heuristic: **enforce minimum chunk diversity** as a quality control rule.

1.3 Task compliance failures dominated CS results when output constraints were not hard enough

The largest failure mode in Phase 1 appeared in **cross-source synthesis**. Several runs defaulted to claim extraction tables rather than producing the required agreement/disagreement structure. Even when claims and citations were grounded, the output failed the research operation: it did not align points across the two sources and did not surface disagreements with paired evidence.

This indicates that for synthesis tasks, **format compliance is not cosmetic**. It is the mechanism that forces comparison and prevents the model from reverting to easier single-source summarization behavior.

1.4 “Benchmark” scope mismatch can be driven by corpus selection rather than model behavior

In CE-2, the extracted content emphasized evaluation frameworks, expert-based indicators, and method descriptions. Even high-quality outputs did not strongly reflect classic benchmark components such as standardized datasets, shared tasks, and reproducible baselines. This is likely a **corpus-content effect**: the selected text segments contain more evaluation methodology than benchmark artifacts. The model mostly mirrors what is present in the chunks. The correct remedy is to adjust corpus composition and chunk selection in Phase 2 rather than expecting prompting alone to produce missing benchmark evidence.

2. Major failure modes and what they imply

2.1 Low-information claims and redundancy in CE

Some CE outputs used a fixed 5-row budget inefficiently, including meta claims that add little research value. This problem appears when prompts request “5 claims” without specifying what makes a claim valuable. It becomes more pronounced under Prompt B because the model prefers low-risk statements that are easy to quote and audit.

Recommended fix: explicitly ban boilerplate framing claims and require a spread across claim types such as problem motivation, method/process, system capabilities, empirical results, and limitations.

2.2 Missing or unresolvable citations in CS

A synthesis table can look correct while still failing the assignment requirement if the evidence cell does not contain **resolvable (source_id, chunk_id)** citations. Using “Source 1” and “Source 2” without chunk IDs breaks traceability. This is a critical trust failure because graders cannot map claims back to evidence.

Recommended fix: add a hard constraint that every row must contain at least two citations in the required format, one per source, or explicitly state no evidence for a side while still citing the closest chunk.

2.3 Wrong output format for CS

Several CS runs produced claim extraction tables instead of agreement/disagreement tables. This is a prompt-compliance failure rather than a grounding failure. It implies that Phase 2 must include systematic checks that the model output matches the required schema.

Recommended fix: use an output validator that checks column headers, minimum row count, and per-row dual citations.

3. Phase 2 design choices grounded in Phase 1 findings

3.1 Retrieval and chunking strategy

Phase 1 showed that claim quality depends heavily on which passages are available and how they are chunked. For Phase 2, chunking should be **section-aware** for papers and aim to preserve evaluative statements, protocol details, and limitations. Because PDF extraction artifacts appeared in quotes, Phase 2 should include basic text cleaning while preserving exact snippet fidelity for citation.

Design choice: implement a chunking pipeline that records **source_id, chunk_id, section label, page range**, and stores the raw text snippet so citations remain resolvable.

3.2 Citation enforcement and trust behavior

Prompt B demonstrated that traceability improves when the model is forced to audit citations. In Phase 2, this should become a system-level trust policy, not just a prompt trick.

Design choice: enforce citation rules in generation:

- Every major claim requires a citation key
- No evidence means the system must explicitly say “not found” and suggest next retrieval steps
- Add lightweight post-generation checks to reject outputs with missing citations

3.3 Output validation and schema-first generation

The CS failures indicate a need for schema validation. A portal that returns the wrong artifact shape is not research-grade.

Design choice: implement structured output schemas for CE and CS and validate:

- Table headers match required schema
- Row count meets minimum
- Each row includes required citations, including cross-source paired citations for synthesis

If validation fails, the system will rerun with a correction instruction or return a structured error that is logged.

3.4 Evaluation set design and metrics

Phase 1 suggests two metrics are especially diagnostic:

- **Citation correctness**: whether cited chunk contains the quoted snippet

- **Groundedness**: whether claims are supported by retrieved evidence
  Additionally, Phase 2 should add at least one retrieval metric aligned with observed failure modes:
- **Citation coverage diversity**: number of distinct chunks cited per answer
- Or standard retrieval metrics like context precision/recall

Design choice: log and evaluate not only answer quality but also evidence coverage patterns, since narrow chunk reliance correlates with low usefulness.

4. Summary of actionable takeaways

Phase 1 demonstrated that grounding can be achieved reliably when chunked evidence is provided and citations are required, but research usefulness depends on constraints that promote claim diversity and prohibit low-information claims. The largest gap was cross-source synthesis compliance, where models frequently reverted to simpler claim extraction unless the prompt and validation forced the agreement/disagreement structure with dual-source citations. Phase 2 will therefore prioritize section-aware chunking, strict citation policies, schema validation, and evidence coverage metrics, ensuring that the RAG system is traceable, evaluable, and robust to missing evidence.