The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume L-4/W1-2026
Free and Open Source Software for Geospatial (FOSS4G) 2026 – Academic Track "Bridging Geospatial Technology and Humanity",
30 August–5 September 2026, Hiroshima, Japan

A Systematic Comparison of RAG Architectures for Geographic POI Question Answering Using OpenStreetMap Data

Noboru Otsuka[1]

[1] Geolonia Inc., 1-6-16 Fukagawa, Koto-ku, Tokyo 135-0033, Japan – noboru.otsuka@geolonia.com

Keywords: OpenStreetMap, Retrieval-Augmented Generation, Geospatial Question Answering, Point of Interest, Structured Spatial Processing, Large Language Models.

Abstract

Retrieval-Augmented Generation (RAG) grounds Large Language Models in external knowledge, yet geospatial question answering presents a distinctive challenge: spatial relationships such as distance and direction are directly computable from coordinate data, blurring the role of vector or graph-based retrieval that prevails in general-domain RAG. We systematically compare five enhanced RAG architectures—Structured, GraphRAG, Hybrid, Adaptive, and Agentic—for geographic Point of Interest (POI) question answering, all built on a shared vector-retrieval substrate over 1,047 OpenStreetMap POIs in Shibuya, Tokyo, with multi-area generalization tested across four Tokyo districts (about 3,600 POIs). Evaluation employs a hierarchical five-level prompt framework (L1–L5, 90–130 cases per phase) with multi-dimensional scoring covering keyword success, reasoning quality, evidence citation, constraint satisfaction, and uncertainty acknowledgement. In Phase 1 (90 cases), Structured RAG attained 89.1% versus GraphRAG's 76.7% and Adaptive RAG's 86.1% (Wilcoxon, Bonferroni-corrected, $p < 0.001$); per-category analysis identified two query types (directional comparison, competitor density) where GraphRAG remained superior. In Phase 2 (130 cases, four areas), Hybrid RAG achieved the best balance of composite quality (67.1/100) and cross-level stability, though pairwise differences with Adaptive and Graph RAG were not statistically significant. Findings suggest that, in dense-urban POI settings where coordinates are reliable, the marginal benefit of explicit graph edges shrinks for coordinate-computable relationships, while structured spatial processing complements vector retrieval. All software (ChromaDB, NetworkX, Hugging Face Transformers) and data (OpenStreetMap) are open-source, ensuring FOSS4G-community reproducibility.

1. Introduction

Large Language Models (LLMs) have demonstrated remarkable natural language understanding capabilities, yet they struggle with tasks requiring access to up-to-date, domain-specific knowledge. Retrieval-Augmented Generation (RAG) addresses this limitation by retrieving relevant external documents and incorporating them into the LLM's generation context (Lewis et al., 2020). While RAG has been widely adopted across domains, its application to geospatial question answering (GeoQA) presents distinctive challenges: geographic queries often require spatial computation (e.g., distance calculation, directional reasoning) that purely semantic retrieval cannot provide.

Recent work has proposed specialized frameworks for spatial RAG. Spatial-RAG (Yu et al., 2025) introduces a hybrid retriever combining sparse spatial filtering with dense semantic matching, while MapQA (Li et al., 2025) constructs a QA dataset from OpenStreetMap using SQL templates and evaluates LLM-based text-to-SQL approaches. WalkRAG (Amendola et al., 2025) targets walkable urban itinerary recommendation through spatially-enhanced retrieval, and Geode (Gupta et al., 2024) proposes a zero-shot geospatial QA agent with explicit reasoning. In parallel, GraphRAG approaches (Edge et al., 2024) have gained traction for their ability to encode explicit entity relationships in knowledge graphs, and recent comparative studies (Han et al., 2025) have begun systematically evaluating RAG against GraphRAG on general-purpose tasks.

A critical gap remains, however: existing studies propose individual frameworks without systematically comparing architectural alternatives under controlled conditions, particularly in the geospatial domain where spatial relationships have the distinctive property of being directly computable from coordinate data. Practitioners building location-based services, urban planning tools, or tourism applications lack empirical guidance on which RAG architecture best suits geographic Point of Interest (POI) queries. The relative contribution of deterministic spatial processing, semantic vector search, and graph-based relational retrieval also remains unquantified in this domain.

This paper presents a systematic comparison of five enhanced RAG architectures for geographic POI question answering, all built on a shared vector-retrieval substrate but differing in how structured spatial knowledge is incorporated. We evaluate them under identical test prompts and scoring functions using OpenStreetMap data, in two complementary phases (Section 4.2.3). Specifically, we investigate the relative effectiveness of deterministic structured processing, graph-based retrieval, and their combinations—and whether these components function as complementary rather than competing strategies.

Scope. This study evaluates RAG architectures on dense-urban POI settings within the Tokyo metropolitan area, using Japanese-language queries. The four evaluation districts (Shibuya, Shinjuku, Ikebukuro, Tokyo Station) share characteristics atypical of POI environments worldwide: high category diversity within compact spatial extents, dominant restaurant share, and predominantly Japanese-language signage and naming conventions. Findings should be interpreted within this scope; generalization to rural or sparse POI distributions, multilingual environments, or domains with distinct relational structures (e.g., social networks, citation graphs) requires further investigation.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-L-4-W1-2026-219-2026

Our contributions are:

1. A controlled comparison of five enhanced RAG architectures (Structured, GraphRAG, Hybrid, Adaptive, Agentic) for dense-urban POI question answering, all built on a shared vector-retrieval substrate but differing in how structured spatial knowledge is incorporated.
2. A hierarchical five-level evaluation framework (L1–L5) with multi-dimensional scoring designed for geospatial QA, capturing not only entity retrieval accuracy but reasoning quality, evidence, and uncertainty.
3. Empirical evidence that, in dense-urban POI settings where coordinates are reliable, deterministic spatial processing outperforms graph-based retrieval, with quantified per-category analysis identifying the specific query types where graph-based retrieval retains an advantage.
4. A multi-area generalization study across four Tokyo districts demonstrating a persistent gap between keyword retrieval accuracy and composite reasoning quality, motivating multi-dimensional evaluation for geospatial QA.

2. Related Work

2.1 Geospatial Question Answering

Geospatial question answering has evolved from template-based approaches over linked data (Punjani et al., 2020) to LLM-powered systems. GeoLLM (Manvi et al., 2024) demonstrated that auxiliary map data from OpenStreetMap can significantly enhance LLM geospatial predictions. MapQA (Li et al., 2025) constructed 3,154 QA pairs from OSM for two US regions, finding that retrieval-based methods handle spatial relationships intuitively while LLM text-to-SQL approaches excel at single-hop reasoning. GeoAnalystBench (Zhang et al., 2025) and SpatialBench (Xu et al., 2025) provide benchmarks for evaluating LLM spatial cognition, though they focus on general spatial reasoning rather than POI-specific queries. Most existing systems target a single architecture; few studies systematically compare alternative retrieval strategies for POI-centric queries under identical conditions.

2.2 RAG Architectures and Comparative Studies

Standard RAG (Lewis et al., 2020) retrieves documents via embedding similarity. GraphRAG (Edge et al., 2024) extends this by constructing knowledge graphs that encode entity relationships, enabling structured traversal for context retrieval. A recent systematic evaluation of RAG versus GraphRAG on general-purpose tasks (Han et al., 2025) found that GraphRAG frequently underperforms vanilla RAG despite higher computational costs—a result we test in the geospatial setting. HybridRAG (Sarmah et al., 2024) combines vector and graph retrieval, reporting improvements in factual correctness for financial document QA. None of these comparative studies, however, address the geospatial domain, where spatial relationships have a distinctive property that distinguishes them from general knowledge: distance, direction, and proximity are directly computable from coordinate data, raising the question of whether explicit graph edges remain necessary.

2.3 Spatial RAG

Spatial-RAG (Yu et al., 2025) is the most closely related work, proposing a hybrid spatial retriever that combines sparse spatial filtering with dense semantic matching and formulating the answering process as multi-objective optimization. Real-time Spatial RAG (Campo et al., 2025) extends digital twin architectures for urban environments, while WalkRAG (Amendola et al., 2025) targets walkable itinerary recommendation. Our work differs in two key aspects: (1) we compare five enhanced architectures under controlled conditions rather than proposing a single framework, and (2) we introduce a hierarchical evaluation framework that captures reasoning quality beyond keyword accuracy, allowing us to characterize not only which architecture wins overall but in which query types each excels.

3. Data

3.1 POI Dataset Construction

We extracted POI data from OpenStreetMap (OSM) using the Overpass API on 2026-01-15, targeting four major Tokyo metropolitan districts: Shibuya, Shinjuku, Ikebukuro, and Tokyo Station areas. The primary dataset comprises 1,047 POIs within a 1 km radius of Shibuya Station (35.658034°N, 139.701636°E), with three additional areas contributing approximately 2,550 POIs for a total of approximately 3,600 POIs.

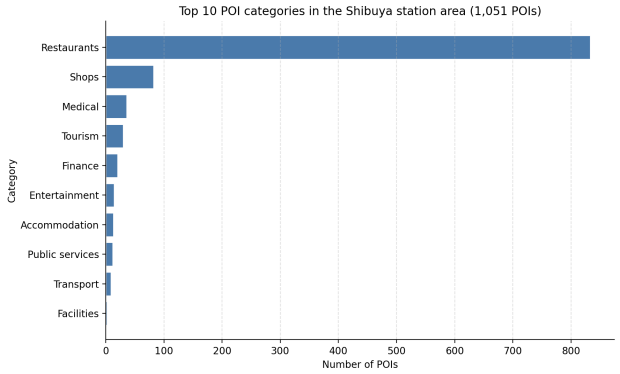

Figure 1. POI category distribution for the Shibuya station area (1,047 POIs). The dataset exhibits significant class imbalance, with restaurants comprising 78% of all POIs.

Each POI record contains: name, category, latitude, longitude, address, and opening hours extracted from OSM tags. We enriched each POI with computed spatial metadata: Haversine distance from the area reference point (in metres); compass direction classified into eight cardinal/intercardinal directions; and a relative position descriptor.

Dataset characteristics relevant to scope. Three properties of this dataset shape the interpretation of our results: (i) all four areas are dense urban districts within Tokyo with comparable POI density (approximately 800–1,050 POIs per 1 km radius), so geographic generalization claims (Section 6.3) are restricted to similar urban contexts; (ii) the category distribution is heavily skewed toward restaurants (78% in Shibuya, Figure 1), reflecting the entertainment-district character of central Tokyo and limiting our findings on category-aware reasoning to environments with similar mix; (iii) all evaluation queries are posed in Japanese, with the structured-RAG keyword classifier and the Japanese-aware embedding model (multilingual-e5-base) tuned for that language—multilingual or English-only settings remain untested.

3.2 Vector Store

All POI documents were embedded using the `intfloat/multilingual-e5-base` model (768 dimensions, max sequence length 512) and stored in ChromaDB 0.4.x, an open-source vector database with persistent local storage configured for cosine similarity. POI documents are formatted as "[name] ([category]) at [address], [hours]" before embedding, with the E5 prefix `passage:` for indexing and `query:` at retrieval time. The multilingual model was selected to handle Japanese text while maintaining compatibility with English query terms present in OSM data. Top-k retrieval ($k = 5$) is used as the default in all architectures.

3.3 Knowledge Graph

For the GraphRAG architecture, we constructed a knowledge graph using NetworkX with 1,090 nodes (1,046 POI nodes, 12 category nodes, 32 area subdivision nodes) and approximately 82,000 edges across seven relationship types (Table 1).

Edge Type	Count	Description
NEAR_TO	66,248	POIs within 100 m
COMPETITOR	∼8,000	Same category, within 100 m
COMPLEMENTARY	∼3,000	Complementary services
SAME_CATEGORY	2,086	Shared POI category
SAME_HOURS	∼2,000	Similar operating patterns
SAME_CUISINE	∼500	Shared cuisine type
SAME_BRAND	∼200	Chain store affiliations

Table 1. Knowledge graph edge types and approximate counts.

4. Method

4.1 Architectures

All architectures share two common components: (i) the same underlying LLM (`Qwen/Qwen2.5-7B-Instruct`, 4-bit NF4 quantized via `bitsandbytes` 0.43.x with `double_quant=True` and `compute_dtype=bfloat16`), and (ii) a vector retrieval substrate using cosine similarity over POI documents embedded with `intfloat/multilingual-e5-base` (768 dimensions; top-k=5 by default). Generation uses temperature 0.7, top-p 0.9, and a maximum of 2,048 new tokens; deterministic seed 42 is set for both PyTorch and Python `random` where applicable. This controlled design isolates the effect of structured spatial processing—what each architecture *adds on top of* the shared vector substrate—from the underlying retrieval and language modelling capability. The five enhanced architectures differ as follows.

Structured RAG. Augments the vector substrate with four deterministic spatial processing modules, each triggered by Japanese keyword matching: a *proximity module* (e.g., "nearest") computing Haversine distances and returning distance-sorted results; a *sensitivity module* (e.g., "range") comparing POI sets at default and half-radius; a *comparison module* (e.g., "east", "west") computing distributions by compass direction; and an *aggregation module* (e.g., "how many", "category") generating counts and summary statistics. These modules are applied in combination rather than as mutually exclusive alternatives, with vector retrieval always running as a complementary source.

GraphRAG. Retrieves context by traversing the knowledge graph (Section 3.3). Given a query, relevant POI nodes are identified via embedding similarity, then the graph is traversed to gather neighbouring nodes and edge relationships as additional context.

Hybrid RAG. Combines Structured RAG's spatial processing modules with vector retrieval in a fixed pipeline. Structured context is generated first, followed by vector results, providing both computed spatial data and semantically similar documents.

Adaptive RAG. Dynamically selects between Structured RAG and GraphRAG for each query based on rule-based question classification. Queries identified as relationship-oriented (competitor, complementary) route to GraphRAG; all others route to Structured RAG.

Agentic RAG. Employs a ReAct-style reasoning loop (Yao et al., 2023) where the LLM autonomously decides which retrieval tools to invoke, including spatial computation tools, vector search, and graph traversal. The agent iterates through Thought → Action → Observation cycles until it determines sufficient context has been gathered.

4.2 Evaluation Framework

4.2.1 Hierarchical Test Prompt Design. We designed a hierarchical five-level test prompt framework (Table 2) comprising 55 prompts spanning 12 subcategories, with difficulty increasing from factual retrieval (L1) to multi-step reasoning under uncertainty (L5).

This design ensures evaluation goes beyond simple retrieval accuracy to assess spatial computation (L2), logical constraint handling (L3), multi-criteria reasoning (L4), and judgement under uncertainty (L5)—capabilities that distinguish geospatial QA from general-purpose QA tasks.

4.2.2 Multi-Dimensional Scoring. All systems were evaluated using a multi-dimensional scoring function with six metrics: (i) keyword success rate, (ii) reasoning score (0–5; reasoning indicators and numerical evidence), (iii) evidence score (0–5; citation of POI names, coordinates, distances), (iv) constraint score (0–5; satisfaction of query constraints), (v) uncertainty score (0–5; appropriate hedging), and (vi) composite score (0–100; weighted combination). Composite weights are level-dependent and progressively shift from keyword/coordinate matching at lower levels toward reasoning/evidence/uncertainty at higher levels: L1 weights keyword (0.4), coordinate (0.3), and POI name (0.3); L2 weights keyword (0.3), coordinate (0.2), and reasoning (0.5); L3 distributes across keyword (0.2), POI name (0.2), evidence (0.2), and constraint (0.4); L4 splits among reasoning (0.3), evidence (0.3), constraint (0.2), and uncertainty (0.2); L5 splits among reasoning (0.4), evidence (0.3), and uncertainty (0.3). All metrics were computed via rule-based automatic evaluation using regular expression and keyword matching—not LLM-as-judge—ensuring deterministic and reproducible scoring.

4.2.3 Evaluation Phases. Our evaluation comprises two complementary phases. Phase 1 (Architecture Comparison) isolates architectural differences using 90 test cases (55 base prompts plus 35 graph-specific prompts targeting competitor analysis, complementary POI discovery, and brand affiliation) on Shibuya data, evaluated across Structured RAG, GraphRAG, and Adaptive RAG. Phase 2 (Multi-Area Generalization) validates generalization with Hybrid, Graph, Adaptive, and Agentic RAG across the four Tokyo districts, using 130 prompts (L1–L5 across four areas plus cross-area prompts), yielding

Level	Name	Description	Prompts
L1	Basic Retrieval	Factual POI lookup by name or category	10
L2	Spatial Reasoning	Proximity, density, directional queries	15
L3	Constraint Satisfaction	Spatial + categorical combined filters	10
L4	Decision Support	Multi-criteria location recommendations	10
L5	Advanced Reasoning	Sensitivity analysis, uncertainty handling	10

Table 2. Hierarchical five-level test prompt framework (55 prompts total).

520 total queries. Phase 1 thus characterizes the core trade-off between deterministic spatial processing and graph-based retrieval, while Phase 2 evaluates how the more elaborate hybrid and agentic strategies generalize to multiple geographic areas.

4.2.4 Statistical Testing. To assess the significance of pairwise system differences, we apply two complementary tests within each phase. McNemar's exact binomial test is applied to binary success outcomes (composite score ≥ 50 in Phase 1, success-rate flag in Phase 2), and Wilcoxon's signed-rank test is applied to continuous per-question composite scores (0–100 scale). Within each phase and metric, the family of pairwise p-values is corrected using the Bonferroni method at $\alpha = 0.05$. Tests are implemented with `scipy.stats.wilcoxon` and `statsmodels.stats.contingency_tables.mcnemar`; the script is released alongside this paper.

5. Results

5.1 Architecture Comparison: Structured RAG vs. GraphRAG

Vector-only baseline. As a reference point, an earlier evaluation of vector retrieval alone (no structured processing, no graph traversal) on the 55 L1–L5 base prompts yielded a mean composite score of 60.3, against which all enhanced architectures evaluated below deliver substantial absolute gains.

Overall and graph-specific performance. In the controlled architecture comparison using 90 test cases (55 base prompts plus 35 graph-specific prompts), Structured RAG achieved 89.1% overall composite score, while GraphRAG and Adaptive RAG attained 76.7% and 86.1% respectively (Figure 2). Notably, on the 35 graph-specific test cases designed to favour relational retrieval, Structured RAG still scored 86.9% versus GraphRAG's 80.2%.

Where each architecture excels. Per-category analysis (Figure 3, Table 3) reveals a more nuanced picture than the overall scores suggest. GraphRAG retains a clear advantage on two query types whose answers depend on relational structure rather than coordinate geometry: **comparison** (east–west directional analysis: 100.0% vs. 50.0%, +50.0 pt) and **competitor** (same-category density analysis: 88.9% vs. 66.7%, +22.2 pt). Conversely, Structured RAG reaches 100% accuracy on four of the categories shown in Table 3 plus the hours category, where deterministic spatial or attribute matching suffices (aggregation, basic_location, constraint_single, cuisine, hours); the first three reduce to coordinate-derivable computations, while cuisine and hours are resolved through attribute matching that vector retrieval supplies competitively. This bifurcation—GraphRAG winning where the answer requires walking explicit relational edges, Structured RAG winning where the answer is reducible to deterministic spatial functions on coordinates—underpins the architectural interpretation in Section 6.1.

Category	GraphRAG	Structured	Advantage
comparison	**100.0**%	50.0%	GraphRAG (+50.0)
competitor	**88.9**%	66.7%	GraphRAG (+22.2)
cuisine	66.7%	**100.0**%	Structured (+33.3)
constraint_single	75.0%	**100.0**%	Structured (+25.0)
basic_location	78.0%	**100.0**%	Structured (+22.0)
aggregation	80.6%	**100.0**%	Structured (+19.4)

Table 3. Categories where GraphRAG and Structured RAG differ most. GraphRAG dominates on relational queries (comparison, competitor); Structured RAG dominates on coordinate-derivable queries.

Statistical significance. Wilcoxon signed-rank tests on per-question composite scores confirm that Structured RAG significantly outperforms GraphRAG (Bonferroni-corrected $p < 0.001$), Adaptive RAG also exceeds GraphRAG ($p = 0.005$), and Structured RAG modestly but significantly exceeds Adaptive RAG ($p = 0.041$). McNemar's exact tests on binary outcomes (composite ≥ 50) yielded no significant pairwise differences after Bonferroni correction (all $p > 0.19$), reflecting that all three Phase 1 systems achieve high baseline pass rates and that architectural differences manifest primarily in answer quality rather than entity retrieval coverage (all reported p-values are Bonferroni-corrected).

5.2 Multi-Area Generalization

The multi-area evaluation across four Tokyo districts extends our analysis to multi-area settings, focusing on hybrid and agentic variants that build on the structured/vector combination (Figure 4). Hybrid RAG achieved the highest keyword success rate at 96.2%, followed by Graph RAG and Adaptive RAG (both 92.3%), and Agentic RAG (90.8%). Agentic RAG required 54.4 seconds per query on average—6× slower than the other systems—due to its multi-step reasoning loop with the 7B-parameter model.

Performance was consistent across geographic areas (Figure 5), with all systems maintaining 85–100% keyword success rates across Shibuya, Shinjuku, Ikebukuro, and Tokyo Station. By difficulty level (Figure 6), all systems perform well at L2 and L4 for keyword success, but Agentic RAG drops to 73% at L2, suggesting that autonomous tool selection is unreliable for spatial reasoning tasks with a 7B-parameter model.

5.3 Multi-Dimensional Quality Analysis

While keyword success rates exceeded 90% for most systems, multi-dimensional composite scoring revealed a critical qual-

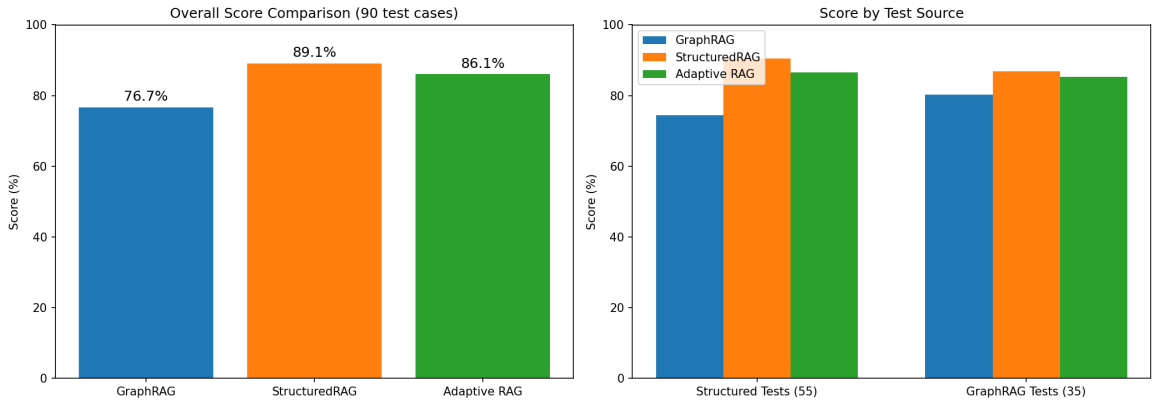

Figure 2. Architecture comparison across 90 test cases. Left: Overall scores. Right: Scores by test source—Structured RAG leads on both, while GraphRAG narrows the gap on its own graph-specific cases (86.9% vs. 80.2%).

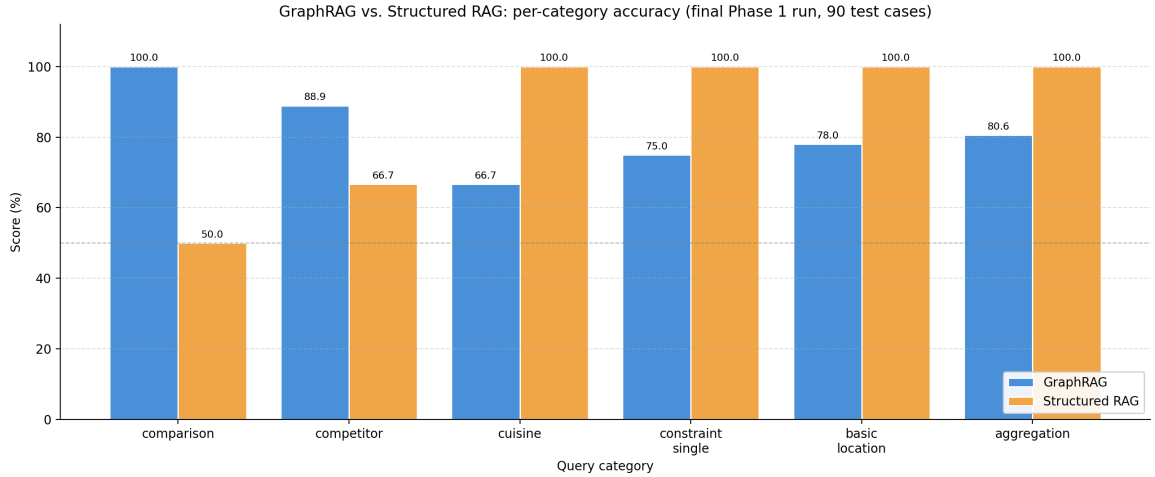

Figure 3. Per-category accuracy comparison between GraphRAG and Structured RAG (90 test cases).

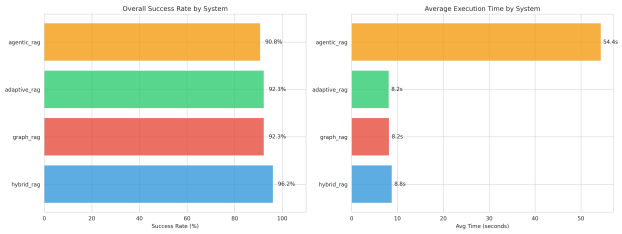

Figure 4. Multi-area evaluation: overall success rate (left) and execution time (right) across four RAG systems.

ity gap (Figure 7). Hybrid RAG scored 67.1/100 in composite quality despite near-perfect keyword accuracy.

The per-level composite analysis (Figure 8) reveals that the quality gap was most pronounced at higher difficulty levels. At L4 and L5, composite scores dropped sharply compared to L1–L3, with the gap between keyword success and composite quality widening (Table 5).

5.4 Generalization Stability

To assess cross-system generalization robustness, we computed the standard deviation of composite scores across difficulty levels and geographic areas (Table 6).

Hybrid RAG achieved the best overall balance among Phase 2 systems: highest composite quality (67.1) with the lowest level-

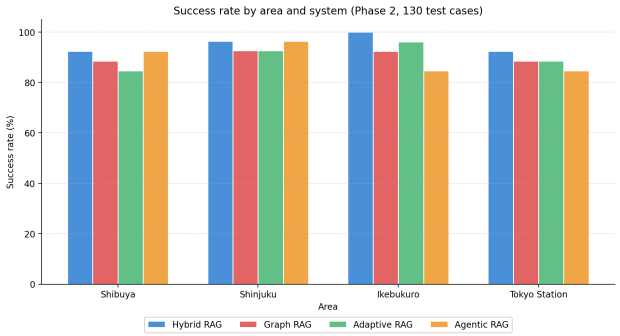

Figure 5. Keyword success rate by geographic area. All systems generalize well across areas, with Hybrid RAG achieving the highest or tied-highest rates in every area.

wise standard deviation (6.85). Graph RAG showed the lowest area variance (1.26), suggesting stable geographic generalization despite lower overall quality. Agentic RAG exhibited the highest instability (level SD = 16.44), with performance varying dramatically between simple retrieval and complex reasoning tasks.

Pairwise significance (Phase 2). McNemar's exact tests on binary success outcomes yielded no significant pairwise differences after Bonferroni correction (all corrected $p > 0.5$), confirming that the four Phase 2 systems are nearly indistinguish-

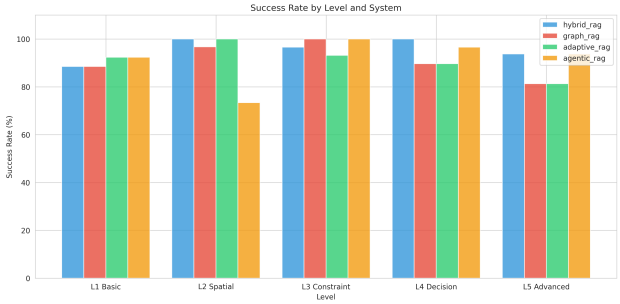

Figure 6. Keyword success rate by difficulty level. Agentic RAG shows a notable drop at L2 (Spatial Reasoning), indicating unreliable autonomous tool selection for spatial queries.

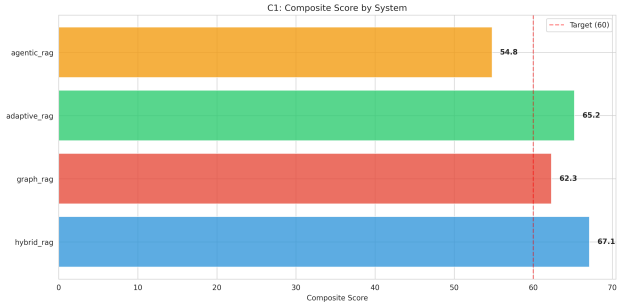

Figure 7. Composite score by system. All systems except Agentic RAG surpass the 60-point threshold, with Hybrid RAG achieving the highest score at 67.1.

able on entity-retrieval coverage. Wilcoxon signed-rank tests on continuous composite scores reveal a more nuanced picture: Hybrid RAG significantly outperforms Agentic RAG (corrected $p < 0.001$), as do Graph RAG ($p = 0.016$) and Adaptive RAG ($p < 0.001$); the differences between Hybrid, Graph, and Adaptive do not reach significance after correction (Hybrid vs. Graph corrected $p = 0.056$; Hybrid vs. Adaptive $p = 0.353$; Graph vs. Adaptive $p = 0.216$). These results indicate that, within the dense-urban POI setting, the principal separable dimension is the cost of autonomous tool selection by the agentic loop, not the choice among structured/graph/hybrid strategies.

6. Discussion

6.1 Why Deterministic Processing Outperforms Graph-Based Retrieval (in Coordinate-Computable Cases)

Coordinate-based computability as the boundary. The Phase 1 advantage of Structured RAG over GraphRAG (89.1% vs. 76.7%, Wilcoxon $p < 0.001$ after Bonferroni correction) is best understood through the lens of *coordinate-based computability*. Many geographic relationships—distance, compass direction, proximity, density within radius—are directly calculable from (lat, lon) pairs using well-established formulas (e.g., the Haversine distance). For such relationships, an explicit graph edge is essentially a pre-computed cache of a function that the system can also evaluate on demand. The principal value proposition of knowledge graphs in general-domain RAG—making implicit relationships explicit—thus offers reduced marginal benefit when the relationships in question are derivable from coordinate primitives.

Where graph edges retain value. This boundary is not absolute. Per-category results (Table 3) show that GraphRAG retains a clear advantage on **comparison** queries (directional distribution: +50.0 pt) and **competitor** queries (same-category density: +22.2 pt). Both involve operating over a pre-aggregated relational structure that even coordinate-based computation must reconstruct on the fly. Furthermore, several relevant POI relationships are not coordinate-derivable at all: brand affiliation, cuisine taxonomy, complementary services (e.g., hotel–restaurant pairing), and similarity of operating hours all depend on attribute matching or domain knowledge rather than geometry. Where such relationships dominate the query workload, graph-based retrieval remains an appropriate—and likely necessary—design choice.

System	Kw SR	Composite	Reasoning	Evidence
Hybrid RAG	**96.2%**	**67.1**	**2.65**	**3.71**
Adaptive RAG	92.3%	65.2	2.62	3.30
Graph RAG	92.3%	62.3	2.53	3.03
Agentic RAG	90.8%	54.8	2.11	2.79

Table 4. Multi-dimensional evaluation results across four RAG systems. While all systems achieve 90%+ keyword success, composite quality varies substantially. (Keyword success rates from the multi-area run; composite/reasoning/evidence scores from the C1 prompt evaluation; see Section 4.2.)

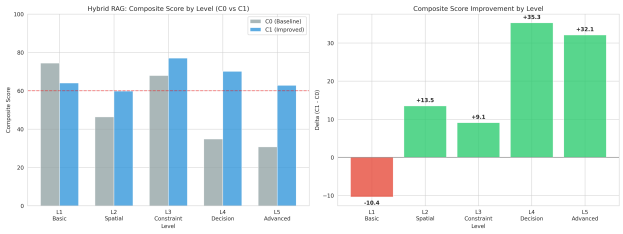

Figure 8. Composite score by difficulty level for Hybrid RAG. L4 and L5 show the largest quality gaps.

Generalization caveat. Our findings should not be read as a general claim that GraphRAG offers limited value. They are specific to dense-urban POI settings where coordinates are reliable and the dominant query types target spatial proximity, direction, and aggregation. In domains where relationships are implicit by nature—citation networks, social graphs, biological pathways, ontological hierarchies—the trade-off can favour graph-based retrieval, since coordinate primitives are unavailable. Even within POI domains, sufficient scale (e.g., 10^6 POIs) makes spatial indices such as R-tree, S2, or H3 act as coordinate-derived structures with similar pre-computation benefits to a knowledge graph. Future hybrid designs should treat graph edges as a selective tool for relationships that resist direct coordinate-based computation, rather than a default substrate for all spatial queries.

Adaptive routing as a boundary case. The Adaptive RAG system (86.1%) failed to match Structured RAG's overall performance despite dynamically selecting between Structured and GraphRAG. Analysis of its selection distribution revealed that it routed 68.9% of queries to Structured RAG and 31.1% to GraphRAG. The keyword-based classifier likely misroutes a subset of borderline queries—those answerable by either approach—preventing Adaptive RAG from reaching Structured RAG's overall score, suggesting that automatic architecture selection in the geospatial setting requires classifiers more sophisticated than keyword-based rules, ideally informed by the coordinate-computability boundary discussed above.

Level	Kw SR	Composite	Gap
L1 Basic	88.5%	64.0	24.5
L2 Spatial	100.0%	59.8	40.2
L3 Constraint	96.6%	77.0	19.6
L4 Decision	100.0%	70.1	29.9
L5 Advanced	93.8%	62.8	31.0

Table 5. Keyword success rate vs. composite score by difficulty level (Hybrid RAG), showing the persistent quality gap.

System	Composite	Lvl SD	Area SD
Hybrid RAG	**67.1**	**6.85**	2.85
Adaptive RAG	65.2	9.71	3.51
Graph RAG	62.3	9.91	**1.26**
Agentic RAG	54.8	16.44	4.33

Table 6. Generalization stability metrics. Hybrid RAG achieves the best balance of quality and consistency.

6.2 The Keyword–Composite Quality Gap

One of the most significant findings is the persistent gap between keyword success rate and composite quality score. All four Phase 2 systems achieved keyword success rates above 90% (Table 4), yet composite quality scores ranged widely from 54.8 (Agentic RAG) to 67.1 (Hybrid RAG). This demonstrates that **retrieving the correct entities is necessary but insufficient** for generating well-reasoned geospatial answers.

The gap was most pronounced at L4 (Decision Support) and L5 (Advanced Reasoning), where generating useful answers requires not merely mentioning relevant POIs but providing structured reasoning with evidence, satisfying multiple constraints simultaneously, and acknowledging uncertainty. Our hierarchical evaluation framework, with its level-dependent scoring weights (Section 4.2.2), was specifically designed to capture this distinction.

This finding has practical implications: evaluation of geospatial QA systems based solely on entity retrieval metrics (precision, recall, F1 over mentioned POIs) would significantly overestimate system quality for real-world applications where users expect reasoned recommendations.

6.3 Limitations

6.3.1 Geographic and Linguistic Scope. The four evaluation districts (Shibuya, Shinjuku, Ikebukuro, Tokyo Station) are all dense urban areas within Tokyo, exhibiting characteristics atypical of POI environments worldwide: high category diversity within compact spatial extents, restaurant-dominated category mix (78% in Shibuya, Section 3.1), reliable coordinate data, and predominantly Japanese-language signage and naming conventions. Findings should therefore be interpreted as specific to comparable dense-urban, coordinate-rich POI settings. Generalization to the following remains an open question requiring dedicated evaluation: (i) rural or sparse POI distributions, where the typical inter-POI distance is one to two orders of magnitude larger and density-based queries dominate less; (ii) multilingual or non-Japanese query environments, where the structured-RAG keyword classifier and the embedding model's language alignment behave differently; (iii) larger spatial scales (regional or national queries spanning 10^6+ POIs), at which spatial indices (R-tree, S2, H3) become essential and the comparison between graph-based and coordinate-based representations may be reframed; and (iv) POI domains with distinct categorical structure—natural landmarks, transit infrastructure, public services—where the relational vs. coordinate-computable balance differs from the entertainment-district mix studied here.

6.3.2 Other Limitations. Scale. Our evaluation used 90 (Phase 1) and 130 (Phase 2) test prompts. While we apply paired statistical tests (Section 5.4), larger-scale evaluations would be required to reliably detect smaller effect sizes within the 90–95% success rate band; in particular, Phase 2 pairwise differences between Hybrid, Graph, and Adaptive RAG do not reach statistical significance after Bonferroni correction and warrant re-examination at higher sample sizes.

Evaluation method. Our rule-based scoring, while deterministic and reproducible, may not capture all aspects of answer quality. Comparison with human judgements or LLM-as-judge evaluation would strengthen the validity of our metrics.

Model scale. All experiments used a 7B-parameter LLM (4-bit quantized). Larger models may exhibit different architectural preferences, particularly for Agentic RAG where autonomous tool selection demands stronger reasoning capabilities.

Temporal data. OSM data is a snapshot; POI opening hours, existence, and categories change over time. Our evaluation does not account for temporal validity.

7. Conclusions and Future Work

This paper presented a systematic comparison of five enhanced RAG architectures for geographic POI question answering using OpenStreetMap data. Our controlled evaluation—identical data, test prompts, and scoring functions across all systems—yielded three principal findings:

1. **Deterministic structured processing outperformed graph-based retrieval in our evaluation.** Structured RAG achieved 89.1% accuracy versus GraphRAG's 76.7% across 90 Phase 1 cases (Wilcoxon $p < 0.001$ after Bonferroni correction), including on graph-specific test cases designed to favour relational retrieval. We interpret this as evidence that, *for dense-urban POI queries where spatial relationships are directly computable from coordinates*, the marginal benefit of explicit graph edge representations is limited. This finding does not generalize to domains with implicit or non-computable relationships (e.g., citation networks, social graphs), where graph-based retrieval may remain the appropriate choice; rather, it suggests that future hybrid designs should treat graph edges as a selective tool for relationships that resist direct computation (Section 6.1).
2. **Structured processing and vector search are complementary strategies.** Phase 1 evidence shows that augmenting the vector substrate with deterministic spatial processing yields a $+28.8$ pt absolute improvement over a vector-only baseline ($60.3 \rightarrow 89.1$, $p < 0.001$). In Phase 2, Hybrid RAG, which extends this combination explicitly, achieved the highest numerical composite quality (67.1/100, under L4/L5-weighted composite scoring distinct from Phase 1; see Section 4.2) and the most stable

level-wise performance (SD = 6.85) *among the four evaluated systems*. We caution, however, that pairwise differences between Hybrid, Graph, and Adaptive RAG did not reach statistical significance after Bonferroni correction (Section 5.4). The architectural principle of combining structured and vector retrieval is well-supported, while specific hybrid implementations are largely interchangeable for dense-urban POI queries; the clearest separation is between the non-agentic systems and Agentic RAG.

3. **Single-metric evaluation is insufficient for geospatial QA.** All four Phase 2 systems achieved keyword success rates above 90% yet composite quality scores ranged from 54.8 to 67.1, demonstrating that entity retrieval accuracy alone cannot assess the reasoning quality of geospatial answers. Hierarchical, multi-dimensional evaluation frameworks are essential, particularly for higher-difficulty levels (L4 Decision Support and L5 Advanced Reasoning) where the gap is most pronounced.

All experiments were conducted using exclusively open-source software (ChromaDB, NetworkX, Hugging Face Transformers) and open data (OpenStreetMap), ensuring full reproducibility by the FOSS4G community.

Future work will pursue four directions: (1) extending the evaluation framework to larger-scale multilingual POI datasets beyond the Tokyo metropolitan area, incorporating spatial databases such as PostGIS for scalable spatial indexing; (2) investigating adaptive weighting and selective use of graph edges only for relationships not directly computable from coordinates (e.g., competitor density, complementary services, brand affiliation); (3) re-examining Phase 2 system differences at larger sample sizes, since the present 130-case study cannot statistically distinguish Hybrid from Adaptive or Graph RAG; and (4) conducting human evaluation studies to validate our rule-based scoring framework against expert judgements.

Acknowledgements

This research used OpenStreetMap data, available under the Open Database License (ODbL). Computational experiments were conducted on Google Colab with T4 and A100 GPU instances. We thank the anonymous reviewers of the FOSS4G 2026 abstract submission for their constructive feedback, which helped us refine the scope and framing of our claims, particularly regarding the generalizability of our findings beyond the Tokyo metropolitan POI setting.

References

Amendola, M., Pugliese, C., Perego, R., Renso, C., 2025. Spatially-Enhanced Retrieval-Augmented Generation for Walkability and Urban Discovery. *arXiv preprint arXiv:2512.04790*.

Campo, D. N., Conde, J., Alonso, Á., Huecas, G., Salvachúa, J., Revíriego, P., 2025. Real-time Spatial Retrieval Augmented Generation for Urban Environments. *arXiv preprint arXiv:2505.02271*.

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Larson, J., 2024. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv preprint arXiv:2404.16130*.

Gupta, D. V., Ishaqui, A. S. A., Kadiyala, D. K., 2024. Geode: A Zero-shot Geospatial Question-Answering Agent with Explicit Reasoning and Precise Spatio-Temporal Retrieval. *arXiv preprint arXiv:2407.11014*.

Han, H., Ma, L., Wang, Y., Shomer, H., Lei, Y., Qi, Z., Guo, K., Hua, Z., Long, B., Liu, H., Aggarwal, C. C., Tang, J., 2025. RAG vs. GraphRAG: A Systematic Evaluation and Key Insights. *arXiv preprint arXiv:2502.11371*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., Kiela, D., 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems (NeurIPS)*.

Li, Z., Grossman, M., Qasemi, E., Kulkarni, M., Chen, M., Chiang, Y.-Y., 2025. MapQA: Open-domain Geospatial Question Answering on Map Data. *arXiv preprint arXiv:2503.07871*.

Manvi, R., Khanna, S., Mai, G., Burke, M., Lobell, D., Ermon, S., 2024. GeoLLM: Extracting geospatial knowledge from large language models. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Punjani, D., Iliakis, M., Stefou, T., Singh, K., Both, A., Koubarakis, M., Angelidis, I., Bereta, K., Beris, T., Bilidas, D., Ioannidis, T., Karalis, N., Lange, C., Pantazi, D.-A., Papaloukas, C., Stamoulis, G., 2020. Template-Based Question Answering over Linked Geospatial Data. *arXiv preprint arXiv:2007.07060*.

Sarmah, B., Hall, B., Rao, R., Patel, S., Pasquali, S., Mehta, D., 2024. HybridRAG: Integrating Knowledge Graphs and Vector Retrieval Augmented Generation for Efficient Information Extraction. *arXiv preprint arXiv:2408.04948*.

Xu, P., Wang, S., Zhu, Y., Li, J., Qi, G., Zhang, Y., 2025. SpatialBench: Benchmarking Multimodal Large Language Models for Spatial Cognition. *arXiv preprint arXiv:2511.21471*.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y., 2023. ReAct: Synergizing reasoning and acting in language models. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yu, D., Bao, R., Ning, R., Peng, J., Mai, G., Zhao, L., 2025. Spatial-RAG: Spatial Retrieval Augmented Generation for Real-World Geospatial Reasoning Questions. *arXiv preprint arXiv:2502.18470*.

Zhang, Q., Gao, S., Wei, C., Zhao, Y., Nie, Y., Chen, Z., Chen, S., Su, Y., Sun, H., 2025. GeoAnalystBench: A GeoAI Benchmark for Assessing Large Language Models for Spatial Analysis Workflow and Code Generation. *arXiv preprint arXiv:2509.05881*.