# Automatic Generation of a Compositional QA Benchmark for Geospatial Reasoning under Spatial and Entity Constraints

**Tetsuhisa Suizu**[♠] **Shohei Higashiyama**[♡,♠] **Hiroyuki Shindo**[♢]
**Hiroki Ouchi**[♠] **Sakriani Sakti**[♠]

[♠]Nara Institute of Science and Technology [♢]MatBrain, Inc.
[♡]National Institute of Information and Communications Technology

suizu.tetsuhisa.st8@naist.ac.jp,{hiroki.ouchi,ssakti}@is.naist.jp
shohei.higashiyama@nict.go.jp,hshindo@matbrain.jp

## Abstract

Despite their recent success, the geospatial reasoning capabilities of large language models (LLMs)—which require understanding spatial relationships among real-world geo-entities—remain underexplored. We propose an automatic method for constructing compositional geographic question answering datasets that jointly consider spatial and entity constraints. The generated dataset serves as a principled benchmark for evaluating how LLMs coordinate spatial computation with entity-level understanding under diverse compositional settings. We evaluate two state-of-the-art LLMs, GPT-5.2 and Gemini 3 Flash, on our dataset. Experimental results show that while the models perform relatively well on questions involving rich entity grounding, their accuracy drops substantially on questions requiring precise quantitative spatial reasoning, such as distance estimation and containment judgment. Our dataset is publicly available for research and reproduction.

## 1 Introduction

Recent advances in large language models (LLMs) have greatly expanded their capability to perform reasoning tasks that integrate linguistic, visual, and factual information. Despite this progress, LLMs still struggle with **geospatial reasoning**, which requires understanding spatial relationships such as distances, containment, and direction among real-world **geographic entities** (geo-entities), e.g., places and facilities. This type of reasoning goes beyond purely geometric interpretation, requiring the capability to link spatial structures with *entity constraints* including historical, cultural, or functional characteristics that define specific locations and capture their unique social and semantic contexts. Evaluating how well LLMs can perform such integrated reasoning through natural language is therefore an essential step toward understanding their capacity for geographically grounded intelligence.

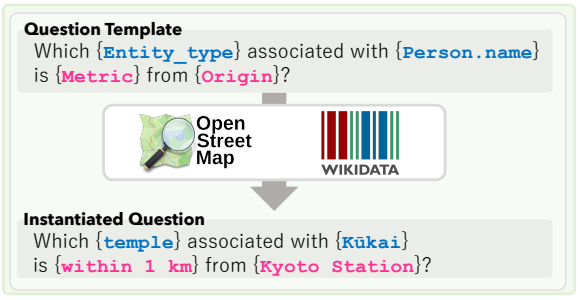


Figure 1: Overview of our approach.

Recent studies have introduced geospatial question answering (QA) tasks (Li et al., 2025a; Feng et al., 2025) to evaluate how LLMs reason over language that describes spatial and entity relationships. These tasks highlight the importance of enabling LLMs to perform *geospatial reasoning through natural language*, where spatial and entity cues are conveyed textually rather than through coordinates. While these benchmarks have advanced the evaluation of geographically grounded reasoning, most existing datasets do not explicitly specify the reasoning skills each question requires. Consequently, it remains unclear how *spatial constraints* and *entity constraints* are represented or interact within questions, making it difficult to interpret what kind of reasoning skills they actually test.

To fill this gap, we propose a method for automatically constructing a **compositional geospatial QA dataset** that jointly considers spatial and entity constraints. We first systematize the elements that comprise spatial- and entity-constrained questions, such as distance conditions and entity types. These elements are then combined in a controlled manner to generate diverse geospatial questions with varying reasoning complexity and interpretable structure. *Spatial constraints* capture geometric relations between geo-entities, such as distance, rela-



*Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*
*Volume 4: Student Research Workshop*, pages 818–830

tive position, direction, and containment, whereas *entity constraints* represent semantic and relational characteristics of geo-entities, including their associations with non-geo-entities such as people and events. To illustrate how these two types of constraints interact, consider the following example:

> *Which temple associated with Kūkai is located within 1 km of Kyoto Station?*

Answering this question requires reasoning about both spatial proximity (a 1 km radius) and entity association (the temple's relationship to Kūkai[1], a prominent Buddhist monk (774–835) and founder of the Shingon school of Japanese Buddhism). Despite the growing accuracy of commercial map services such as Google Maps[2], their query mechanisms are still limited to keyword- or coordinate-based retrieval. They can filter locations by spatial conditions like distance or travel time, but they cannot reason over semantic relations between entities, for example, identifying temples historically associated with a specific person within a given area. This gap highlights the need for systems that can jointly interpret spatial constraints and entity-level semantics through natural language.

Building on this framework, we automatically generate question-answer pairs by integrating geographic information from **OpenStreetMap (OSM)**[3] with the entity information from **Wikidata**[4]. This integration enables a wide spectrum of reasoning patterns, ranging from simple spatial queries to complex multi-level questions that jointly involve spatial and entity dimensions. The resulting dataset serves as a principled benchmark for evaluating how LLMs coordinate spatial computation and entity-level understanding under diverse compositional settings.

We evaluate two state-of-the-art LLMs, GPT-5.2 and Gemini 3 Flash on 5,309 automatically generated questions derived from 14 slot-based compositional templates. Our experiments reveal clear patterns. The models perform relatively well on questions involving rich entity grounding, such as those referring to people or cultural properties. In contrast, their accuracy drops substantially on questions requiring precise quantitative spatial reasoning such as distance estimation or containment judgment. These results indicate that current LLMs rely heavily on associative knowledge rather than on structured spatial computation, highlighting the need for a deeper integration of geographic knowledge representations.

The contributions of this paper are threefold:

- We propose a **compositional question generation framework** that systematically combines spatial and entity elements. This framework provides a principled approach to constructing geographically grounded questions with controllable reasoning complexity.
- We construct a **benchmark dataset** based on this framework by integrating OpenStreetMap and Wikidata. The dataset covers diverse reasoning types, from purely spatial to hybrid spatial-entity questions, and supports interpretable evaluation. Our dataset is publicly available for research and reproduction.[5]
- We conduct a **systematic evaluation** of multiple LLMs across diverse geospatial constraints and present key findings that reveal their strengths, weaknesses, and current limitations in integrated geographic reasoning.

Through this work, we aim to provide a foundation for compositional, interpretable, and geographically grounded evaluation of reasoning in LLM.

## 2 Related Work

This section reviews three lines of related research: formal models of spatial relations (§2.1), geographically grounded QA (§2.2), and compositional evaluation of linguistic and multimodal reasoning (§2.3). Our work integrates these perspectives into a unified framework that provides an interpretable, compositional, and geographically realistic benchmark for evaluating spatial reasoning in LLMs.

### 2.1 Spatial Reasoning and Representation

Spatial reasoning has long been a central topic in both cognitive science and spatial information theory. Early formal frameworks such as the Region Connection Calculus (RCC8) (Randell et al., 1992) and the 9-Intersection Model (Egenhofer, 1991) formalized qualitative spatial relations including containment, overlap, and adjacency. From a cognitive perspective, research on spatial representation has

[1] https://en.wikipedia.org/wiki/K%C5%ABkai
[2] https://www.google.com/maps
[3] https://www.openstreetmap.org
[4] https://www.wikidata.org/wiki/Wikidata:Main_Page
[5] We will release our dataset at https://github.com/NAIST-geo-and-lang/Compositional_GeoQA_Benchmark

emphasized how humans perceive and organize spatial knowledge (Mark, 1999). Their view bridged cognitive models of spatial understanding with geographic information systems (GIS), establishing a foundation for later formalizations of spatial relations. Cognitive linguistic studies further explored how humans conceptualize space through frames of reference (absolute, relative, intrinsic) (Levinson, 2003) and spatial schemas (Talmy, 2000). In natural language processing (NLP), ISO-Space (Pustejovsky et al., 2010; Pustejovsky and Yocum, 2013) has provided a unified framework for annotating spatial expressions, integrating linguistic, temporal, and geometric representations.

### 2.2 Geospatial QA Datasets

A growing number of QA datasets have been developed to evaluate models' capability to perform reasoning over spatial and geographic information. **GeoQA** (Chen et al., 2021) introduced one of the earliest large-scale benchmarks for geospatial QA, combining map-based data with natural language questions that require spatial reasoning such as distance, containment, and direction. **MapQA** (Chang et al., 2022) further extended this approach by generating questions grounded in cartographic layouts, testing the capability to interpret symbolic map elements and spatial relationships. Building on this line of work, the recent **MapQA (Open-domain)** (Li et al., 2025b) explores open-domain geospatial QA over large-scale map data, integrating geographic facts from multiple sources beyond static cartographic contexts. **STBench** (Li et al., 2025a) incorporated both spatial and temporal dimensions, evaluating reasoning over movements, trajectories, and spatio-temporal relations. **CityBench** (Feng et al., 2025) introduced a city-scale QA dataset that integrates urban geographic data and supports multi-hop reasoning about places and events within real metropolitan environments.

While these datasets have significantly advanced the study of geographic QA, most of them treat question sets as homogeneous collections without explicit decomposition of reasoning types. These studies evaluate overall QA accuracy but do not distinguish which aspects of reasoning—such as distance estimation, containment, direction, or entity-level association—pose greater challenges for LLMs. Moreover, existing QA datasets tend to emphasize either **spatial geometry** (coordinate-based reasoning) or **factual retrieval** (semantic and entity-based knowledge), but rarely integrate both within a single compositional structure.

### 2.3 Compositional Dataset Design

Compositionality has become a central concept in evaluating systematic generalization in language and reasoning models. Early benchmarks such as SCAN (Lake and Baroni, 2018), CFQ (Keysers et al., 2020), and CLOSURE (Dasgupta et al., 2022) were designed to test whether models can generalize to novel combinations of known primitives. These datasets revealed that sequence-to-sequence models often rely on shallow pattern matching rather than learning underlying compositional rules.

In the domain of visual and multimodal reasoning, **GQA** (Hudson and Manning, 2019) extended this principle by introducing compositional QA over real-world images. GQA questions are automatically generated from scene graphs that encode objects, attributes, and relations, allowing fine-grained control over reasoning steps and compositional complexity. Their work demonstrated that structured, interpretable question design can diagnose model reasoning behavior far more effectively than surface-level accuracy measures.

However, most existing spatial or geographic QA datasets lack such compositional control. The question templates in these datasets are often designed heuristically, without an explicit semantic decomposition that clarifies which reasoning skills are being tested. Building on the compositional perspective established in previous studies such as SCAN and GQA, our work extends this idea to the geospatial domain. We adopt a **compositional dataset design** that decomposes geographic questions into interpretable elements representing spatial and entity semantics.

## 3 Dataset

### 3.1 Dataset Design Principles

We design our dataset to evaluate how well LLMs can understand and reason about spatially grounded language, which refers to real-world locations and spatial relationships between them. Our goal is to examine how effectively a model can integrate spatial, entity, and contextual information to answer questions correctly. To achieve this goal, the dataset is constructed according to two main design principles.

1. **Structured Composition.** Each question is constructed by combining multiple spatial and

entity elements, enabling precise control over the type of knowledge or reasoning skill being tested, as well as the capability to adjust question difficulty and linguistic diversity.

2. **Geographic Realism.** All questions are grounded on real-world geo-entities such as train stations, temples, and museums. Each entity is associated with consistent coordinates and attributes, ensuring that spatial reasoning occurs in realistic situations rather than artificial or purely abstract contexts.

These principles ensure that the dataset can systematically test spatial reasoning in realistic contexts while maintaining clear control over question complexity and knowledge types.

### 3.2 Compositional Elements of Spatial and Entity Semantics

To implement these design principles, we construct a taxonomy of *Compositional Elements* that defines the semantic structure of each question, shown in Table 4 in Appendix C. These elements jointly encode both spatial conditions and entity semantics, linking natural language expressions with structured representations that describe reference points, distances, and entity types. This serves as the foundation for generating spatial questions that require interpretable and compositional reasoning.

#### 3.2.1 Example of Structured Composition

The following is a question template:

> *Which* `{entity_type}` *is* `{metric}` *from* `{origin}`*?*

By filling the slots in the template, we can generate concrete questions as follows:

> *Which temple is within 3 km from Nara Station?*

To fill the slots, we use the information stored in a structured representation as follows:

```
{
 "entity_type": "temple"
 "origin": "Nara Station",
 "metric": {
    "distance": 3,
    "unit": "km",
    "relation": "within"
  },
}
```

This process defines a transparent mapping between linguistic form and structured semantics, enabling systematic generation and analysis of spatial questions.

#### 3.2.2 Taxonomy of Compositional Elements

The design of the Compositional Elements follows cognitive and formal grounding; Each element corresponds to a cognitively motivated or formally defined aspect of spatial semantics. The design draws on well-known theories such as spatial frames of reference (Levinson, 2003), the RCC8 model of topological relations (Randell et al., 1992), and the ISO-Space framework (Pustejovsky et al., 2010; Pustejovsky and Yocum, 2013). This grounding ensures that the taxonomy is compatible with both linguistic theory and spatial reasoning.

The taxonomy consists of two complementary hierarchies:

**Space elements** encode spatial constraints that define how entities are located or related in space. They include topological relations, metric constraints, directional orientation, and route-based connectivity.

**Entity elements** describe the semantic characteristics of entities. They specify the entity's type, historical or personal associations, and descriptive attributes such as institutional designation, physical form, or perceptual quality. The structure aligns with existing ontological resources such as Wikidata.

Table 4 in Appendix 4 presents the taxonomy of the Compositional Elements, covering both spatial and entity dimensions. Each element is organized into a category (*S*-series for spatial and *E*-series for entity elements), and each facet represents a distinct and interpretable aspect of meaning such as spatial relation, direction, functional role, or institutional status. The table serves as a unified reference for question generation, showing how natural language templates correspond to structured representations. Spatial elements define the *locative and geometric constraints* of a question, while entity elements specify its *semantic identity and descriptive attributes*.

#### 3.2.3 Conceptual and Practical Contributions

The proposed decomposition offers both semantic clarity and practical utility for analyzing, generating, and evaluating spatial questions. Its main contributions can be summarized as follows:

| Elements | Question Template | Count |
|---|---|---|
| | *Spatial-Constraints-Focued Questions* | |
| S0, S2, E1 | Which {entity_type} is within {distance} from {origin}? | 400 |
| S0, S2, S3, E1 | Which {entity_type} is within {distance} from {origin} to the {direction}? | 400 |
| S0, S1, S2, E1 | Which {entity_type} in {region} is within {distance} from {origin}? | 400 |
| S0, S1, S2, S3, E1 | Which {entity_type} in {region} is within {distance} from {origin} to the {direction}? | 400 |
| | *Entity-Constraints-Focused Questions* | |
| S1, E1, E2 | Which {entity_type} in {region} is related to {person}? | 257 |
| S1, E1, E4 | Which {entity_type} in {region} is designated as {cultural_property}? | 326 |
| | *Spatial and Entity Constraints Highly Mixed Questions* | |
| S0, S2, E1, E2 | Which {entity_type} related to {person} is within {distance} from {origin}? | 383 |
| S0, S2, S3, E1, E2 | Which {entity_type} related to {person} is within {distance} from {origin} to the {direction}? | 382 |
| S0, S2, E1, E4 | Which {entity_type} designated as {cultural_property} is within {distance} from {origin}? | 400 |
| S0, S2, S3, E1, E4 | Which {entity_type} designated as {cultural_property} is within {distance} from {origin} to the {direction}? | 399 |
| S0, S1, S2, E1, E2 | Which {entity_type} in {region} related to {person} is within {distance} from {origin}? | 383 |
| S0, S1, S2, E1, E4 | Which {entity_type} in {region} designated as {cultural_property} is within {distance} from {origin}? | 398 |
| S0, S1, S2, S3, E1, E2 | Which {entity_type} in {region} related to {person} is within {distance} from {origin} to the {direction}? | 381 |
| S0, S1, S2, S3, E1, E4 | Which {entity_type} in {region} designated as {cultural_property} is within {distance} from {origin} to the {direction}? | 400 |
| **Total** | | **5,309** |

Table 1: Statistics of our dataset with question templates.

**Semantic clarity.** Each element corresponds to a cognitively grounded, linguistically transparent component of spatial meaning. This design makes it easier for both humans and models to understand how spatial relations and entity semantics are represented in a question.

**Compositional control.** Complex spatial questions can be systematically constructed, modified, or decomposed by combining a small number of atomic elements. This compositional structure provides fine-grained control over the types of knowledge and reasoning being tested.

By aligning natural language with structured spatial semantics, the Compositional Elements framework establishes a systematic and transparent foundation for structured dataset design and for comparative evaluation of spatial reasoning performance across models and datasets.

### 3.3 Dataset Construction Flow

To evaluate basic geospatial reasoning capabilities of LLMs, we designed 14 distinct question templates, shown in Table 1 based on the compositional elements defined in Table 4 in Appendix 4. These templates integrate multiple constraints, including spatial factors (e.g., distance limits $S_2$ and directional reasoning $S_3$) and entity attributes (e.g., entity types $E_1$ and historical associations $E_2$). The dataset is constructed through a semi-automatic pipeline that links structured geographic data with these natural language question templates. In this section, we explain the process in detail.

**Overview.** In this study, we defined 14 templates that combine four spatial elements (Origin $S_0$, Topological $S_1$, Metric $S_2$, Directional $S_3$) and three entity elements (Type $E_1$, Person $E_2$, Attributes $E_4$). For each template, we (i) enumerate valid slot instantiations, (ii) compute the corresponding gold answer set by deterministic filtering, and (iii) sample a balanced subset to form the final benchmark dataset. We focus on entities primarily located in Japan, where both Wikidata and OSM provide dense coverage.

#### Step 1: Origin and Target

We design all questions to explicitly ask for entities belonging to a target category $E_1$, e.g., "Which `{entity_type}` is ...?". Accordingly, the first step constructs an entity inventory that supports two roles: (i) **origins** ($S_0$) used as spatial anchors, and

(ii) **targets** ($E_1$) to be returned as answer entities.[6]

**For origin constraints ($S_0$).** We consistently use railway stations as origins because they are ubiquitous, well-defined, and naturally serve as reference points for human mobility. We retrieve station entities from Wikidata (items whose type corresponds to the tag `railway_station`), and store for each station its identifier and latitude/longitude.

**For target (entity type) constraints ($E_1$).** We retrieve answer entity candidates from Wikidata and OpenStreetMap, restricting the entity type $E_1$ to five categories: `temple`, `shrine`, `castle`, `art_museum`, and `museum`. For each target, we extract its identifier(s) and latitude/longitude.

**Step 2: Spatial Computation**

**For distance constraints ($S_2$).** To conduct spatial computation efficiently, we use a two-stage procedure: fast retrieval by coordinates, followed by an exact distance check. For each target entity type $E_1$, we build a spatial index (BallTree). Given an origin station $S_0$ and a distance threshold $\{0.5, 1.0, 5.0, 10.0, 50.0\}$ km, we first retrieve candidate entities within the radius using the index (fast filtering). We then recompute the *exact* geodesic distance on the Earth's surface using the Haversine formula, and keep only entities that satisfy the distance constraint $S_2$.

**For direction constraints ($S_3$).** We compute the azimuth, i.e., the direction angle measured clockwise from North, from the origin to each candidate and assign it to one of four cardinal directions: North ($\theta \in [315°, 45°)$), East ($[45°, 135°)$), South ($[135°, 225°)$), and West ($[225°, 315°)$). We then keep only candidates whose direction matches the template's required label.

**Step 3: Topological and Entity Constraint**

**For topological constraints ($S_1$).** Some templates restrict answers to a specific administrative area. To support this, we normalize the address information of each entity to the `Prefecture + Municipality` level. Concretely, we parse the address strings provided by Wikidata/OSM, keep only valid (prefecture, municipality) pairs, and store the resulting region label for each entity.

[6]In this study, we adopted the answer type ($E_1$) to enable comparable evaluation across templates. Our slot-based formulation, however, is not limited to type-conditioned retrieval and can be extended to generate other question families; we plan to explore such extensions in future work.

**For person association constraints ($E_2$).** We use Wikidata relations that connect a place to a person. Depending on the available metadata, these relations include links such as `founded_by`, `dedicated_to`, or `associated_with`. We treat an entity as satisfying $E_2$ if it has an explicit Wikidata link to the target person.

**For attribute constraints ($E_4$).** We filter entities using attribute metadata from Wikidata, including person-related attributes (e.g., architect, founder, owner) and cultural_property designations (e.g., National Treasure, Important Cultural Property).

**Step 4: Structured Representation**

For each answer candidate entity, we store a JSON record that contains the instantiated spatial and entity fields used in question generation, such as `origin`, `entity_type`, `distance`, `direction`, `metric`, `person`, and `attribute` (see Section 3.2.1).

**Step 5: Question Realization**

Using the JSON records, we generate a natural-language question by filling the placeholders in the corresponding template with instantiated values (e.g., origin name, distance value/unit, direction, region, and entity type). Each generated question is stored together with (i) its JSON record and (ii) the gold answer set computed in Step 4.

**Step 6: Answer Set Construction**

For each instantiated question, we construct its gold answer set by collecting all entities of the target `entity_type` ($E_1$) that satisfy every active constraint in the template, including spatial constraints ($S_0/S_1/S_2/S_3$) and entity constraints ($E_2/E_4$). We keep only questions whose gold answer set size falls within $[1, 5]$, removing unanswerable cases (0 answers) and overly ambiguous cases (more than 5 answers).

**Step 7: Balanced Sampling for Geographic Diversity**

We apply a sampling step to avoid geographic bias. In particular, we prevent questions from being concentrated in a small number of stations or regions, and we keep the distribution of spatial conditions as balanced as possible.

We sample valid question instances using the following rules: (1) **Distance balance:** we sample an equal number of questions from each of the five distance thresholds; (2) **Origin balance:** within

each distance group, we limit how many questions come from the same origin station, and for templates with $S_3$ we also balance the four directions (North/East/South/West); (3) **Region balance:** we spread questions across regions at the `Prefecture + Municipality` level using round-robin selection; (4) **Backfilling:** if a group does not have enough valid questions, we add more questions from other regions within the same distance threshold while keeping the overall balance. When possible, we sample 80 questions per entity type and template.

### 3.4 Characteristics of Our Dataset

As a result, each question in our dataset is paired with (i) a machine-readable slot record and (ii) a answer set. This makes evaluation interpretable and reproducible under controlled combinations of constraints. Table 1 summarizes the statistics of our dataset. The dataset contains a total of 5,309 questions generated from 14 question templates drawn from OpenStreetMap and Wikidata.

**Structural complexity.** The structural complexity of each question is determined by the number and type of instantiated Compositional Elements. Note that while a larger number of elements generally increases the structural richness of a question, it does not necessarily correspond to higher reasoning difficulty, as some elements may be independent or redundant. In our experiments, we further analyze how reasoning difficulty varies across questions with different structural configurations.

## 4 Experiments

### 4.1 Experimental Setup

**Models.** We evaluate the geospatial reasoning capability of two proprietary LLMs—GPT-5.2 (OpenAI, 2025) and Gemini 3 Flash (Google, 2025)—accessed via public APIs under the same prompt. We configured both models to rely solely on internal knowledge by enabling chain-of-thought reasoning (set to "medium") and explicitly disabling external search tools. We leave more comprehensive evaluations, including LLMs equipped with external search tools and other open LLMs, for future work. Detailed hyperparameters are listed in Table 8 in the Appendix B.2.

**Prompt format.** As shown in Figure 2 in Appendix B.1, we designed the structured prompt that requires two outputs. Each model was instructed to first produce the predicted location name beginning with *"Answer:"*, followed by a textual explanation beginning with *"Reason for Answer:"*. Shown in Figure 2 in Appendix B.1, we designed the structured prompt that requires two outputs. We explicitly included this *"Reason for Answer:"* field to visualize the model's inference process, allowing us to inspect the underlying logic behind its spatial deductions.

**Evaluation.** The evaluation focuses on the accuracy of the answer. In our task setup, **models are explicitly instructed to provide exactly one location**, even though multiple entities may satisfy the specified conditions. Consequently, if the LLM's single answer matches any of these entities in the gold-standard set, it is considered correct. Model predictions are evaluated by checking whether the predicted location name appears in this set using exact string matching. Quantitative results are reported as accuracy (%).

**Evaluation data size.** We used the complete set of 5,309 generated questions and answers for model evaluation. The dataset consists of an approximately equal number of questions from each of the 14 question templates. However, some templates fell short of the target number because valid questions could not be generated due to the lack of entities satisfying the specific conditions.

### 4.2 Results

Table 2 presents the accuracy (%) of the two models, GPT-5.2 and Gemini 3 Flash, evaluated across all 14 question templates. Looking at the specific categories, both models exhibited their lowest performance in the *Spatial Constraints* category. For these four templates (e.g., "S0,S2,E1"), accuracies ranged from approximately 17% to 35%, indicating that questions relying solely on geometric constraints were the most challenging for both models.

In contrast, the *Entity Constraints* category yielded significantly higher scores. Gemini achieved approximately 65% accuracy on these templates, and GPT also showed improved performance compared to the spatial category.

For the *Composite* category, which combines spatial and entity elements, both models generally maintained higher accuracy than in the pure *Spatial Constraints* category. Notably, Gemini exceeded 65% in four out of eight templates, with three of them achieving close to 70% accuracy.

| Template Elements | GPT-5.2 | Gemini 3 Flash |
|---|---|---|
| *Spatial Constraints* | | |
| S0,S2,E1 | 26.75 | 36.75 |
| S0,S2,S3,E1 | 23.25 | 33.75 |
| S0,S1,S2,E1 | 25.25 | 35.25 |
| S0,S1,S2,S3,E1 | 17.25 | 25.50 |
| *Entity Constraints* | | |
| S1,E1,E2 | 43.97 | 67.32 |
| S1,E1,E4 | 55.52 | 65.34 |
| *Composite (Spatial Constraints + Entity Constraints)* | | |
| S0,S2,E1,E2 | 38.12 | 61.36 |
| S0,S2,S3,E1,E2 | 37.96 | 59.95 |
| S0,S2,E1,E4 | 43.75 | 60.25 |
| S0,S2,S3,E1,E4 | 39.10 | 60.40 |
| S0,S1,S2,E1,E2 | 40.73 | 69.45 |
| S0,S1,S2,E1,E4 | 53.02 | 65.58 |
| S0,S1,S2,S3,E1,E2 | 37.80 | 68.24 |
| S0,S1,S2,S3,E1,E4 | 51.75 | 69.00 |
| **Overall** | **37.75** | **55.00** |

Table 2: Accuracy (%) by question template. Each template is represented by the IDs of its constituent spatial (**S**) and entity (**E**) elements (Table 4).

### 4.3 Qualitative Analysis

To investigate the performance gap shown in Table 2, we analyze representative cases in Table 3. We hypothesize that entity attributes (e.g., historical figures or cultural designations) act as **"Entity Anchors,"** which guide the model to the correct answer even when spatial reasoning is imperfect.

**Failures in Pure Spatial Reasoning.** **Question 1** illustrates the difficulty of pure spatial constraints. Lacking entity cues, both models failed fundamentally: GPT satisfied the distance but misidentified the direction, while Gemini hallucinated a non-existent location. This confirms that without an entity anchor, the search space remains too vast for reliable geometric reasoning.

**Effectiveness of Entity Anchors.** **Questions 2 and 3** demonstrate how anchors mitigate these spatial limitations. In Question 2, the semantic cue ($E_2$: "related to Tokugawa Mitsukuni") allowed GPT to pinpoint the correct shrine. Notably, even GPT's incorrect answer in Question 3 ("Hyokeikan") supports our hypothesis. Unlike the complete hallucination in Question 1, "Hyokeikan" is a specific building *within* the correct facility (Tokyo National Museum). This implies that the anchor successfully narrowed the search to the correct location, resulting merely in a granularity error rather than a spatial deviation.

**Question1 (S0,S2,S3,E1)**: Which temple is within 50.0 km from Owani Onsen Station to the East?
**GPT-5.2:** Seiryū-ji Temple
**Gemini 3 Flash:** Hōjū-ji Temple
**Gold answer(s):** {Daien-ji Temple, Daien-ji}

**Question2 (S0,S2,E1,E2)**: Which shrine related to Tokugawa Mitsukuni is within 10.0 km from Sawa Station?
**GPT-5.2:** Tokiwa Shrine
**Gemini 3 Flash:** Sakatsura Isosaki Shrine
**Gold answer(s):** {Tokiwa Shrine}

**Question3 (S0,S1,S2,E1,E4)**: Which museum in Taito Ward, Tokyo designated as Important Cultural Property is within 0.5 km from Uguisudani Station?
**GPT-5.2:** Hyokeikan
**Gemini 3 Flash:** Tokyo National Museum
**Gold answer(s):** {Tokyo National Museum}

Table 3: Examples of model predictions.

## 5 Conclusion

This paper has presented a method for automatically generating a compositional benchmark, **compositional geospatial QA dataset**, for geospatial reasoning under spatial and entity constraints. We systematized the elements that comprise spatial- and entity-constrained questions, such as distance conditions and entity types. These elements are then combined in a controlled manner to generate diverse geospatial questions with varying reasoning complexity and interpretable structure.

Through extensive evaluation of GPT and Gemini, we found that both models perform well on questions involving rich entity grounding, such as those referring to historical figures or cultural properties, but show substantial performance degradation when quantitative spatial computation (e.g., distance estimation or containment) is required. These results suggest that current LLMs rely heavily on associative knowledge rather than structured spatial reasoning.

Our findings suggest the need for future models to incorporate explicit geographic representations and spatial computation modules, enabling deeper integration between linguistic and spatial reasoning.

We plan to extend this work by expanding the dataset to include multi-hop reasoning, temporal dimensions, and real-world map-based visualization interfaces, providing a more comprehensive benchmark for geographically grounded intelligence.

## Limitations

**Language.** Our dataset was constructed from Japanese OpenStreetMap and Wikidata entries, and therefore all experiments in this paper were conducted in Japanese. The dataset generation pipeline itself, however, is language-agnostic and can, in principle, be applied to other languages with sufficient OpenStreetMap and Wikidata coverage. Future work will explore multilingual generation and evaluation settings, enabling cross-lingual assessment of geospatial reasoning capabilities.

**Geographical coverage.** The current dataset covers geographic entities primarily located in Japan, as the initial construction focused on regions where OpenStreetMap and Wikidata entries provide dense and reliable coverage. However, the proposed generation pipeline itself is globally applicable, since both data sources include worldwide geographic information. Future extensions will incorporate a broader range of countries and cultural contexts, allowing cross-regional comparison of spatial reasoning performance and evaluation of geographic generalization across diverse environments.

**Source diversity and generalizability.** The dataset was constructed entirely from open data sources, namely OpenStreetMap for geographic information and Wikidata for entity information. While this ensures transparency and reproducibility, it also limits the diversity of underlying knowledge to what is represented in these platforms. For example, less-documented regions or entities with incomplete metadata may lead to gaps in spatial or entity coverage. Future work will explore the integration of additional open geographic and entity databases to enhance source diversity and improve the generalizability of geospatial reasoning evaluation across heterogeneous data sources.

**Prompt design.** In the current experiments, we evaluated model performance using a single prompt format for all question types in Figure 2 in Appendix B.1 While this setting allows controlled comparison between models, it may not fully capture the variability of model behavior under different instruction styles. Alternative prompt formulations, such as chain-of-thought guidance or few-shot exemplars, could lead to different reasoning strategies and accuracy patterns. Future work will systematically examine the impact of prompt phrasing and output structure on geospatial reasoning performance.

## Ethics Statement

**License of used resources.** All data sources used in this study are publicly available under open licenses. Geographic information was obtained from OpenStreetMap, which is released under the Open Database License (ODbL) 1.0. Entity information was collected from Wikidata, distributed under the Creative Commons CC0 1.0 Public Domain Dedication. The generated benchmark dataset itself consists only of automatically constructed question–answer pairs derived from these open resources and does not include any copyrighted or personally identifiable content. The LLMs used for evaluation, GPT-5.2 (OpenAI, 2025) and Gemini 3 Flash (Google, 2025), were accessed via their official APIs in compliance with the respective terms of service.

**Use of Logos.** The logos of OpenStreetMap and Wikidata are included in this paper (Figure 1) solely for illustrative and academic purposes to indicate the data sources integrated in our dataset construction process. The **OpenStreetMap logo with text**[7] is licensed under the *Creative Commons Attribution–ShareAlike 4.0 International License (CC BY-SA 4.0)*. © OpenStreetMap contributors. "OpenStreetMap" and the magnifying glass logo are trademarks of the OpenStreetMap Foundation, used in accordance with the *OSMF Trademark Policy*[8]. The **Wikidata logo**[9] is licensed under the *Creative Commons Attribution–ShareAlike 4.0 International License (CC BY-SA 4.0)*. © Wikimedia Foundation. "Wikidata" and the associated logo are trademarks of the Wikimedia Foundation, used under the terms of the *Wikimedia trademark policy* for informational and non-commercial academic use. No endorsement by either organization is implied.

## Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments. This study was partially supported by JSPS KAKENHI Grant Number JP23K24904, JP23K28148, and JST RISTEX Grant Number JPMJRX20B2.

[7] https://wiki.openstreetmap.org/wiki/File:OpenStreetMap_logo_with_text.svg

[8] https://wiki.osmfoundation.org/wiki/Trademark_Policy

[9] https://commons.wikimedia.org/wiki/File:Wikidata-logo.svg

## References


Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. 2022. MapQA: A dataset for question answering on choropleth maps. *Preprint*, arXiv:2211.08545.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. 2021. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523, Online. Association for Computational Linguistics.

Ishita Dasgupta, Danyal Guo, Andreas Stuhlmüller, Samuel J. Gershman, and Noah D. Goodman. 2022. Closure: Assessing systematic generalization of combinatorial structure. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*.

Max J Egenhofer. 1991. Categorizing binary topological relations between regions, lines, and points in geographic databases. *Technical Report, Department of Surveying Engineering, University of Maine*.

Jie Feng, Jun Zhang, Tianhui Liu, Xin Zhang, Tianjian Ouyang, Junbo Yan, Yuwei Du, Siqi Guo, and Yong Li. 2025. CityBench: Evaluating the capabilities of large language models for urban tasks. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 5413–5424, New York, NY, USA. Association for Computing Machinery.

Google. 2025. Gemini 3 Flash: frontier intelligence built for speed.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, pages 6700–6709.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. 2020. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations (ICLR 2020)*.

Brenden M. Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, pages 2879–2888. PMLR.

Stephen C. Levinson. 2003. *Space in Language and Cognition: Explorations in Cognitive Diversity*. Language Culture and Cognition. Cambridge University Press.

Wenbin Li, Di Yao, Ruibo Zhao, Wenjie Chen, Zijie Xu, Chengxue Luo, Chang Gong, Quanliang Jing, Haining Tan, and Jingping Bi. 2025a. Stbench: Assessing the ability of large language models in spatio-temporal analysis. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 749–752, New York, NY, USA. Association for Computing Machinery.

Zekun Li, Malcolm Grossman, Eric, Qasemi, Mihir Kulkarni, Muhao Chen, and Yao-Yi Chiang. 2025b. Mapqa: Open-domain geospatial question answering on map data. *Preprint*, arXiv:2503.07871.

David M Mark. 1999. Spatial representation: A cognitive view. *Geographical information systems: Principles and applications*, 1:81–89.

OpenAI. 2025. Update to GPT-5 system card: GPT-5.2.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An international standard for semantic annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

James Pustejovsky and Zachary Yocum. 2013. Capturing motion in ISO-SpaceBank. In *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 25–34, Potsdam, Germany. Association for Computational Linguistics.

David A. Randell, Zhan Cui, and Anthony G. Cohn. 1992. A spatial logic based on regions and connection. In *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*, KR'92, page 165–176, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Leonard Talmy. 2000. *Toward a Cognitive Semantics, Volume 1: Concept Structuring Systems*. MIT Press, Cambridge, MA.

# A Additional Experimental Results

## A.1 Accuracy across Distance Ranges

Table 5 and Table 6 show accuracy across distance ranges for each question template, with questions sharing the same tamplate and distance range aggregated, for GPT-5.2 and Gemini 3 Flash.

In contrast to distance ranges above 5 km, both models achieved relatively higher accuracy at distances $\leq$1.0 km.

| **Template Elements** | **Distance Range [km]** | | | | |
|---|---|---|---|---|---|
| | **≤0.5** | **≤1.0** | **≤5.0** | **≤10.0** | **≤50.0** |
| *Spatial Constraints* | | | | | |
| S0,S2,E1 | 13.41 | 26.83 | 24.39 | 27.16 | 43.84 |
| S0,S2,S3,E1 | 17.50 | 33.75 | 18.75 | 18.75 | 27.50 |
| S0,S1,S2,E1 | 25.00 | 22.50 | 26.25 | 23.75 | 28.75 |
| S0,S1,S2,S3,E1 | 20.00 | 26.25 | 17.50 | 13.75 | 8.75 |
| *Composite (Spatial Constraints + Entity Constraints)* | | | | | |
| S0,S2,E1,E2 | 46.81 | 49.12 | 40.00 | 35.42 | 28.57 |
| S0,S2,S3,E1,E2 | 40.43 | 43.86 | 39.29 | 35.42 | 34.69 |
| S0,S2,E1,E4 | 52.31 | 56.10 | 42.35 | 34.15 | 36.05 |
| S0,S2,S3,E1,E4 | 46.15 | 52.44 | 44.71 | 40.24 | 14.12 |
| S0,S1,S2,E1,E2 | 57.45 | 35.09 | 35.29 | 39.58 | 41.84 |
| S0,S1,S2,E1,E4 | 63.08 | 60.98 | 46.43 | 48.15 | 48.84 |
| S0,S1,S2,S3,E1,E2 | 51.06 | 38.60 | 31.33 | 35.42 | 38.78 |
| S0,S1,S2,S3,E1,E4 | 55.38 | 57.32 | 44.71 | 50.00 | 52.33 |
| **Overall** | **38.18** | **42.03** | **34.47** | **33.72** | **33.87** |

Table 5: Accuracy across distance ranges for each question template (GPT-5.2).

| **Template Elements** | **Distance Range [km]** | | | | |
|---|---|---|---|---|---|
| | **≤0.5** | **≤1.0** | **≤5.0** | **≤10.0** | **≤50.0** |
| *Spatial Constraints* | | | | | |
| S0,S2,E1 | 34.15 | 35.37 | 34.15 | 35.80 | 45.21 |
| S0,S2,S3,E1 | 38.75 | 42.50 | 27.50 | 25.00 | 35.00 |
| S0,S1,S2,E1 | 33.75 | 35.00 | 35.00 | 36.25 | 36.25 |
| S0,S1,S2,S3,E1 | 30.00 | 33.75 | 17.50 | 21.25 | 25.00 |
| *Composite (Spatial Constraints + Entity Constraints)* | | | | | |
| S0,S2,E1,E2 | 70.21 | 73.68 | 58.82 | 58.33 | 55.10 |
| S0,S2,S3,E1,E2 | 70.21 | 59.65 | 63.10 | 58.33 | 54.08 |
| S0,S2,E1,E4 | 80.00 | 74.39 | 58.82 | 43.90 | 48.84 |
| S0,S2,S3,E1,E4 | 83.08 | 76.83 | 56.47 | 54.88 | 36.47 |
| S0,S1,S2,E1,E2 | 72.34 | 80.70 | 67.06 | 67.71 | 65.31 |
| S0,S1,S2,E1,E4 | 76.92 | 74.39 | 55.95 | 62.96 | 60.47 |
| S0,S1,S2,S3,E1,E2 | 76.60 | 70.18 | 69.88 | 68.75 | 61.22 |
| S0,S1,S2,S3,E1,E4 | 80.00 | 74.39 | 69.41 | 63.41 | 60.47 |
| **Overall** | **58.96** | **59.91** | **51.50** | **50.58** | **49.43** |

Table 6: Accuracy across distance ranges for each question template (Gemini 3 Flash).

## A.2 Analysis of Refusal Rate

Some responses contained phrases indicating refusal (e.g., “Answer: Cannot determine”), or no answer text was produced (i.e., null responses). We therefore calculated the refusal rate, defined as the percentage of instances in which the LLM failed to produce a valid response. Table 7 presents the results.

While refusals accounted for approximately 25% of responses for GPT-5.2, Gemini 3 Flash maintained a refusal rate below 1%. These results suggest that GPT-5.2 tends to refrain from answering questions in uncertain cases, whereas Gemini 3 Flash tends to generate responses more aggressively.

| **Template Elements** | **GPT-5.2** | **Gemini 3 Flash** |
|---|---|---|
| *Spatial Constraints* | | |
| S0,S2,E1 | 20.14 | 0.40 |
| S0,S2,S3,E1 | 19.22 | 0.00 |
| S0,S1,S2,E1 | 28.43 | 0.39 |
| S0,S1,S2,S3,E1 | 34.44 | 0.67 |
| *Entity Constraints* | | |
| S1,E1,E2 | 15.28 | 0.00 |
| S1,E1,E4 | 9.66 | 1.77 |
| *Composite (Spatial Constraints + Entity Constraints)* | | |
| S0,S2,E1,E2 | 20.25 | 0.00 |
| S0,S2,S3,E1,E2 | 25.32 | 0.00 |
| S0,S2,E1,E4 | 32.89 | 0.63 |
| S0,S2,S3,E1,E4 | 25.51 | 0.00 |
| S0,S1,S2,E1,E2 | 27.75 | 0.00 |
| S0,S1,S2,E1,E4 | 27.27 | 1.46 |
| S0,S1,S2,S3,E1,E2 | 26.58 | 0.83 |
| S0,S1,S2,S3,E1,E4 | 26.94 | 0.81 |
| **Overall** | **24.99** | **0.46** |

Table 7: Refusal rates (%) across all questions for GPT-5.2 and Gemini 3 Flash.

# B Detailed Experimental Settings

## B.1 Prompt Format

Figure 2 and Figure 3 show the Japanese prompt used in our experiments and its English translation.

**Prompt**

次の質問の条件を満たす場所を1つ回答してください。回答の形式は、次のように、「回答:」に続けて場所の名前のみを一行で記述してください。
また、次の行に「回答根拠: 」に続けて回答根拠を記述してください。
`{question}`

Figure 2: The Japanese prompt used in the experiments.

Prompt (English Translation)

Please provide exactly one location that meets the conditions of the following question. Your response should be formatted as follows: “Answer” followed by only the name of the location on a single line. Then, on the next line, provide your reasoning for the answer starting with “Reason for Answer:”.
`{question}`

Figure 3: English Translation of the Japanese prompt used in the experiments.

### B.2 Model Hyperparameters

Table 8 shows the hyperparameter settings used for both models in our experiments.

| **Parameter** | **Value** |
|---|---|
| ***GPT-5.2*** | |
| max_output_tokens | 1536 |
| reasoning.effort | medium |
| reasoning.summary | detailed |
| text.verbosity | low |
| ***Gemini 3 Flash*** | |
| temperature | 1.0 |
| maxOutputTokens | 1536 |
| thinkingConfig.includeThoughts | True |
| thinkingConfig.thinkingLevel | medium |

Table 8: Hyperparameter values for the evaluated models.

## C Taxonomy of Compositional Elements

Table 4 shows our taxonomy of Compositional Elements. With the exception of S2, each sub-category appears independently in the question templates, whereas S2 sub-categories always co-occur within the same template.

| Category | Facet (Subtype) | Description | Example (Question fragment) |
|---|---|---|---|
| | | Spatial Compositional Elements | |
| S0 Origin | — | Reference or departure location | "from Kyoto Station …" |
| S1 Topological | disjoint | Spatially separate (no contact) | "far from A …" |
| | adjacent | Touching / sharing boundary | "adjacent to a park …" |
| | overlap | Partially overlapping or crossing | "overlapping A and B wards …" |
| | region | Administrative scope or address filtering (Prefecture, City) | "in Kyoto City, Kyoto Pref. …" |
| S2 Metric | distance_value | Numeric distance value | "3 km from the station …" |
| | distance_unit | Unit of distance (km, m, minutes) | "within 3 km / 10 min on foot …" |
| | threshold_type | Threshold type (within, beyond) | "within 3 km / beyond 3 km …" |
| S3 Directional | absolute_direction | Cardinal direction (north, northeast) | "north of the station …" |
| | relative_direction | Egocentric direction (left, right, front, back, uphill) | "to the right of the station …" |
| S4 Route | via | Intermediate places on a route | "via Tō-ji Temple …" |
| | along | Relation to linear landmarks (streets, rivers) | "along Horikawa Street / Kamo River …" |
| | | Entity Compositional Elements | |
| E1 Type | entity_type | Entity class (e.g., facility category) | "Where is the temple / bakery?" |
| E2 Person | name | Name of person associated with the place | "related to Kūkai …" |
| | relation_type | Relation type (founded_by / visited_by / associated_with / …) | "founded by Kūkai …" |
| E3 Event | name | Name of historical event | "related to the Ōnin War …" |
| | relation_type | Relation type (occurred_at / battle_site_of / meeting_site_of / …) | "the site where the Ōnin War occurred …" |
| E4 Attributes | institutional | Institutional or legal status (designation, zoning) | "a designated Important Cultural Property / in a scenic district" |
| | physical | Physical and morphological properties (shape, material, age) | "a wooden structure / built in 796 CE" |
| | functional | Functional role or use (purpose, capacity) | "used for worship / accommodation" |
| | operational | Operational conditions (hours, fees, accessibility) | "open 8:00–17:00 / wheelchair-accessible" |
| | perceptual | Perceptual or experiential qualities (scenic, quiet, historic) | "a scenic and quiet place" |
| | meta | Identifiers, aliases, and data provenance | "has Wikidata QID / also known as Kyōō-gokoku-ji" |

Table 4: Taxonomy of spatial and entity compositional elements with subtypes and examples.