空間・エンティティ制約下における地理空間推論のための構成的質問応答ベンチマークの自動生成

水津 徹久 [1]　東山 翔平 [2,1]　進藤 裕之 [3]　大内 啓樹 [1]　サクティ サクリアニ [1]
[1] 奈良先端科学技術大学院大学　[2] 情報通信研究機構　[3]MatBrain 株式会社
suizu.tetsuhisa.st8@naist.ac.jp　shohei.higashiyama@nict.go.jp
hshindo@matbrain.jp　{hiroki.ouchi,ssakti}@is.naist.jp

概要

本研究では，空間制約とエンティティ制約の双方を考慮した，構成的な地理空間質問応答ベンチマークを自動生成するための枠組みを提案する．異なる種類の知識・推論能力を要求する 14 の質問テンプレートを設計し，実世界の地理的エンティティに関する合計 5,306 件の質問を生成した．提案ベンチマークを用いた最先端 LLM の評価の結果，エンティティの属性情報や関連人物が紐づいた質問を得意とする傾向と，空間計算を要する質問において課題があることが明らかになった．

1　はじめに

大規模言語モデル（LLM）が進展し，言語・視覚情報を統合的に処理するための推論能力が大幅に向上したが，**地理空間推論**においては依然として課題を抱えている．地理空間推論とは，現実世界の場所や施設などの**地理的エンティティ**が関わる，地理空間的な問題の解決に必要な処理である．地理空間推論においては，地理的エンティティ間の距離，方向，地理的包含関係といった空間的関係（**空間制約**）を捉える能力が求められるとともに，地理的エンティティの文化的性質・社会的機能といった属性的情報や，他の非地理的エンティティとの関係性についての知識（**エンティティ制約**）も問われる．例として，図 1 の質問文に答える状況を想定すると，実用的な問題解決においてこの 2 種の情報を扱うことの必要性を理解しやすい．

従来研究で，空間制約やエンティティ制約が関わる推論を必要とする地理空間質問応答（QA）データセットを構築し，LLM の推論能力を評価する取り組み [1, 2] が行われている．しかし，既存データセットでは，各質問が要求する具体的スキルが明示的に定義されていないため，評価結果から LLM 間の総合的な性能の優劣を比較できても，各 LLM が有する具体的な能力・知識の程度を解釈することが困難である．この課題に対し，本研究では，空間制約とエンティティ制約の両方を取り入れた地理空間 QA ベンチマークを，**構成的**な方法で自動構築するための枠組みを提案する．

図 1　提案手法の概要

提案する枠組みでは，第一に，質問の構成要素となる制約条件を，空間制約とエンティティ制約の 2 種の観点から体系的に整理した分類体系（表 4）を用いる．第二に，分類体系上の制約条件のタイプとその組合せを選択し，それら制約条件をスロットとして含む質問テンプレートを設計・定義する．第三に，Wikidata[1)] および OpenStreetMap（OSM）[2)] 上のエンティティ情報を利用し，質問テンプレートの条件を満たすエンティティおよび条件値を当てはめた質問を生成する．この枠組みにより，単純な空間計算処理の質問から，空間的関係とエンティティ属性・関係的要素を複合的に扱う質問まで，実在するエンティティについての質問セットを柔軟かつ体系的に生成しつつ，質問テンプレートごとの回答精度から，要求されている知識・推論能力の程度を解釈することが可能になる．

本枠組みに基づいて，5,306 件の質問からなるベ

1）https://www.wikidata.org/
2）https://www.openstreetmap.org

ンチマークを生成し，2 つの最先端 LLM（GPT-5.2, Gemini 3 Flash）の性能評価を行った．実験結果から，両モデルとも，関連人物や文化的属性のようなエンティティ制約を含む質問に対して比較的良好な精度を示した一方，距離推定や包含関係判定など，精密な空間的推論を必要とする質問に対しては，大幅に精度が低下する傾向が認められた．

2 関連研究

地理空間情報を対象とした LLM の推論能力を評価するため，近年，多数の地理空間 QA ベンチマークが提案されている．**GeoQA** [1] は，地図データに基づく大規模 QA ベンチマークであり，距離・包含関係・方向などの空間推論を扱う．**MapQA** [3] はこのアプローチをさらに発展させ，地図レイアウトに基づく質問生成により，記号的要素と空間関係の解釈能力を評価する．さらに，**MapQA（Open-domain）** [4] は，大規模な地図データを対象に，複数情報源から地理的事実を統合するオープンドメイン QA を扱っている．また，**STBench** [5] は空間・時間の両次元を取り込み，移動や軌跡，時空間関係の推論を評価する．**CityBench** [2] は，都市地理データと統合した都市規模 QA により，都市環境での場所・事象に関する多段階推論を扱っている．

これらのベンチマークは地理空間 QA 研究を推進してきた一方で，多くの研究では質問に含まれる推論パターンの違いを区別せず，推論タイプを明示的に分解して分析していない．また，QA システムの全体的な精度は評価するものの，距離推定・地理的包含関係・方向・エンティティ間の関係性など，どの推論要素が LLM にとってより難しいかを切り分けて示せていない．さらに，既存ベンチマークは一般的に，**空間幾何学的推論**（座標ベースの推論）か**事実検索**（意味・エンティティ知識）のいずれかに偏りがちで，両者を同一の質問構造で統合的に扱う試みは限定的である．

そこで本稿では，地理空間 QA を，距離・方向といった空間的制約と，文化的性質・社会的機能といったエンティティの属性的情報や他エンティティとの関係性についての知識に基づく推論要素に分解する．この分解に基づき，**構成的ベンチマーク設計手法**を採用することで，空間推論とエンティティ知識を統合的に用いる地理空間 QA 能力を，要素レベルで評価できる枠組みを提案する．

3 ベンチマークの構築手法

ベンチマークの構築手法として，質問の複雑さや，要求される知識の種類，推論パターンを柔軟に制御可能としつつ，多様な質問・解答ペアの自動生成を可能にする枠組みを提案する．

3.1 構成的な質問テンプレートの構築

本研究では，モデルに要求する知識・推論内容の種類に対応する構成要素を定義し，構成要素から質問を組み立てる**構成的**な方法を採用する．構成要素は，回答対象のエンティティが満たすべき制約条件に相当し，本研究で用いる制約は，距離・方向・地理的包含関係といった空間制約と，文化財指定などの属性情報やエンティティに付随する関連人物といったエンティティ制約である．構成要素すなわち制約の全体像は，**分類体系**として定義した（付録 A の表 4）．

続いて，分類体系から制約条件の組合せを選択し，自然言語文として記述することで，質問生成の雛型となる質問テンプレートが定義可能になる．質問テンプレートは，制約条件のスロットを持つ質問文として，次の例のように定義できる[3)]．

> {出発地}から{距離値}km 圏内にある{施設カテゴリ}はどこ？

さらに，テンプレートの各スロットに適切な値を代入することで，次のような具体的な質問が生成される．

> 奈良駅から3km 圏内にある仏教寺院はどこ？

質問テンプレートの制約を満たす条件値の情報は，以下の構造化表現で保持し，質問自動生成に用いた．

```
{
 "施設カテゴリ": "仏教寺院",
 "出発地": "奈良駅",
 "計量的": {
    "距離値": 3,
    "距離単位": "km",
    "閾値タイプ": "圏内"
  },
}
```

3) 質問テンプレートの作成方法として，制約条件の選択および自然言語文への変換を LLM 等で自動化する方法も考えられる．本研究では，自然かつバランスの取れた質問セットを設計する目的で，両者とも著者が人手で選択・作成した．

構成要素	質問文	質問数
	空間制約重視の質問	
S0, S2, E1	{S0}から{S2}km 圏内にある{E1}はどこ？	400
S0, S2, S3, E1	{S0}から{S2}km 圏内にあり{S3}に位置する{E1}はどこ？	400
S0, S1, S2, E1	{S1}で，{S0}から{S2}km圏内にある{E1}はどこ？	400
S0, S1, S2, S3, E1	{S1}で，{S0}から{S2}km 圏内にあり{S3}に位置する{E1}はどこ？	400
	エンティティ制約重視の質問	
S1, E1, E2	{S1}で，{E2}に関係のある{E1}はどこ？	257
S1, E1, E4	{S1}で，{E4}の{E1}はどこ？	326
	複合的な質問	
S0, S2, E1, E2	{S0}から{S2}km 圏内にある，{E2}に関係のある{E1}はどこ？	383
S0, S2, S3, E1, E2	{S0}から{S2}km 圏内にあり{S3}に位置し，{E2}に関係のある{E1}はどこ？	382
S0, S2, E1, E4	{S0}から{S2}km 圏内にある，{E4}として登録されている{E1}はどこ？	400
S0, S2, S3, E1, E4	{S0}から{S2}km 圏内にあり{S3}に位置し，{E4}として登録されている{E1}はどこ？	398
S0, S1, S2, E1, E2	{S1}で，{S0}から{S2}km圏内にある，{E2}に関係のある{E1}はどこ？	383
S0, S1, S2, E1, E4	{S1}で，{S0}から{S2}km圏内にある，{E4}として登録されている{E1}はどこ？	396
S0, S1, S2, S3, E1, E2	{S1}で，{S0}から{S2}km圏内で{S3}に位置し，{E2}に関係のある{E1}はどこ？	381
S0, S1, S2, S3, E1, E4	{S1}で，{S0}から{S2}km圏内で{S3}に位置し，{E4}として登録されている{E1}はどこ？	400
		5,306

表 1　構築したベンチマークデータセットの統計情報

3.2　ベンチマーク構築プロセス

本研究で定義した分類体系（付録 4 の表 4）に基づき，4 つの空間制約（起点 S_0，位相的関係 S_1，計量的関係 S_2，方向的関係 S_3）と，3 つのエンティティ制約（種別 E_1，関連人物 E_2，属性 E_4）を組み合わせることで，表 1 に示す質問テンプレートを設計した．作成した質問テンプレートは，空間制約重視の質問 4 種類，エンティティ制約重視の質問 2 種類，両方の制約を重視した複合的な質問 8 種類の，合計 14 種類である．

続いて，回答対象の候補となる日本国内のエンティティを Wikidata および OSM から取得し，質問テンプレートごとに，後述する方法で条件値の設定とその制約を満たすエンティティの判定を行い，個々の質問を生成した．各質問テンプレートにつき最大 400 問となるよう，生成された質問の中から，出発地が分散するようにサブセットをサンプリングし，合計 5,306 問を最終的なベンチマークデータセットとした[4]．

各質問テンプレートの制約を満たすエンティティの判定は，以下の手順で行った．

ステップ 1：施設カテゴリ

全ての質問は，特定の施設カテゴリ（E_1）に属するエンティティ[5]を求める形式とする．つまり「...{entity_type}はどこ？」（例：「...{仏教寺院}はどこ？」）という形式である[6]．

ステップ 2：空間的起点と空間計算処理

質問を構成する基本要素として**起点**（S_0）がある．これは，質問者と想定される人物の移動の起点を表すもので，本研究では，人間の移動における自然な基準点とみなせる鉄道駅[7]を一貫して採用する．起点を基に，距離制約（S_2）および方向制約（S_3）が決定可能となる．距離制約を提供する場合，起点からの距離を $\{0.5, 1.0, 5.0, 10.0, 50.0\}$ km の範囲から選択するものとする．方向制約を適用する場合，起点からの方向を { 東, 西, 南, 北 } の中から選択するものとする．

ステップ 3：位相幾何的計算

必要に応じ，位相幾何的制約（S_1）として，特定の行政区域（都道府県・市区町村）との包含関係を設定する．

ステップ 4：エンティティ制約

必要に応じ，エンティティの属性や他の（非地理的）エンティティとの関係についての制約を設定する．対象エンティティに関連する人物についての関連人物制約（E_2）では，Wikidata 上で，対象エンティティと他の人物エンティティの間に関係性を示すプロパティ[8]が付いている場合，制約を満たすと

4）　構築したデータセットは次の URL で公開する：https://github.com/NAIST-geo-and-lang/Compositional_GeoQA_Benchmark

5）　施設カテゴリは，寺院，神社，城，美術館，博物館の 5 カテゴリに限定し，該当するプロパティまたはタグを持つエンティティを Wikidata および OSM から取得した．

6）　この形式はタイプ条件付き検索に相当するが，本研究でのスロットベースの定式化はこれに限定されない．エンティティの属性や関係，件数を問うものなど他の質問タイプの質問生成への拡張は今後の課題とする．

7）　instance_of=railway_station を持つ Wikidata エンティティを取得し，そのラベルと緯度・経度を使用した．

8）　founded_by（創設者），dedicated_to（対象物を捧げる人や組織），associated_with（関連人物）を使用した．

判定する．属性制約（E_4）では，対象エンティティが，Wikidata 上でエンティティの属性を示すプロパティ値を持つ場合，制約を満たすと判定する．本研究では，国宝や重要文化財といった文化財指定を属性情報として用いる．

4 実験

4.1 実験設定

モデル 公開 API を通じてアクセス可能な商用 LLM である GPT-5.2 [6] および Gemini 3 Flash [7] の地理空間推論能力を評価する．両モデルとも，同等のハイパーパラメータ（付録 B）およびプロンプト条件を適用し，外部検索ツールは使用せず内部知識のみに依存するよう設定し，推論レベルは Medium とした．

プロンプト形式 回答にあたる場所 1 か所と，回答根拠にあたる説明文という 2 つの情報を，この順に出力することを求めるプロンプト（付録 C）を用いた．

評価方法 各質問は 1 個以上 5 個以下の正解エンティティを持つ．LLM が回答した場所が，正解エンティティのいずれかに，文字列として完全一致する場合に正解とみなした．評価指標として，質問セットに対する正解率を報告する．

4.2 実験結果

表 2 に実験結果を示す．両モデルとも，3 種類の質問カテゴリの中で，「空間制約重視の質問」に対して最も低い正解率（30%前後）を示した．「エンティティ制約重視の質問」では，それよりも高い正解率が得られ，Gemini は 65%を超える正解率を達成した．空間制約とエンティティ制約を組み合わせた「複合的な質問」では，両モデルとも「空間制約重視の質問」より高く，「エンティティ制約重視の質問」と同等またはそれより低い正解率を示した．

4.3 結果の考察

「空間制約重視の質問」の正解率が低い原因として，正解候補のエンティティを絞り込むために正確な空間計算が要請される点において，その空間計算が LLM にとって困難であることが考えられる．一方，「複合的な質問」では，正確な空間計算を行わなくても，エンティティ制約の情報が，正解候補を絞り込むための重要な手がかりとして機能し，正解へと辿り着けるケースが増えると考えられる．たとえば，表 3 に示す予測結果例では，GPT-5.2 は正解「常磐神社」を回答している．この質問では，関連人物（E_2）のエンティティ制約「徳川光圀に関連する」が手がかりとして有用であったと考えられる．

構成要素	GPT-5.2	Gemini 3 Flash
空間制約重視の質問		
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
エンティティ制約重視の質問		
S1,E1,E2	43.97	67.32
S1,E1,E4	55.52	65.34
複合的な質問		
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	39.19	60.25
S0,S1,S2,E1,E2	40.73	69.45
S0,S1,S2,E1,E4	53.28	65.25
S0,S1,S2,S3,E1,E2	37.80	68.24
S0,S1,S2,S3,E1,E4	51.75	69.00
全質問平均	36.20	53.75

表 2 質問テンプレート別の正解率（%）．各テンプレートは，その構成要素である空間要素（**S**）とエンティティ要素（**E**）の ID によって識別される（表 4 参照）．

質問 (S0,S2,E1,E2): 茨城県ひたちなか市にある佐和駅から 10.0km 圏内にある，徳川光圀に関係のある神社はどこ？

GPT-5.2（正解）：常磐神社
Gemini 3 Flash（不正解）： 酒列磯前神社
正解セット: 常磐神社, **常盤神社**

表 3 モデルの予測結果例．正解セットとの一致に基づき正解・不正解を判定している．

5 おわりに

本稿では，空間制約とエンティティ制約下における地理空間推論のための**構成的地理空間 QA ベンチマーク**の構築と，最先端 LLM の性能評価について報告した．今後の展開として，本ベンチマークを拡張し，多段階推論を要する質問，時間的要素を含む質問，ユーザが地図インターフェースを通じて問い合わせる状況を想定したマルチモーダルな質問などを追加し，より包括的な基盤ベンチマークを構築することを予定している．

謝辞

本研究は JSPS 科研費 JP23K24904，JP23K28148 の助成を受けたものです．

参考文献

[1] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pp. 513–523, Online, August 2021. Association for Computational Linguistics.

[2] Jie Feng, Jun Zhang, Tianhui Liu, Xin Zhang, Tianjian Ouyang, Junbo Yan, Yuwei Du, Siqi Guo, and Yong Li. 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, p. 5413–5424, New York, NY, USA, 2025. Association for Computing Machinery.

[3] Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. Mapqa: A dataset for question answering on choropleth maps, 2022.

[4] Zekun Li, Malcolm Grossman, Eric, Qasemi, Mihir Kulkarni, Muhao Chen, and Yao-Yi Chiang. Mapqa: Open-domain geospatial question answering on map data, 2025.

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

[6] OpenAI. Update to GPT-5 system card: GPT5.2, 2025. https://cdn.openai.com/pdf/3a4153c8-c748-4b71-8e31-aecbde944f8d/oai_5_2_system-card.pdf.

[7] Google. Gemini 3 Flash: Frontier intelligence built for speed, 2025. https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/.

A　構成要素

カテゴリ	サブタイプ	説明	例（質問の断片）
		空間構成要素	
S0 起点	—	基準点または出発点	「京都駅 から...」
S1 位相的	非連続	空間的に分離している（接触なし）	「駅から離れた 位置にある...」
	隣接	接している／境界を共有している	「公園に隣接する 場所...」
	重複	部分的に重なっている／交差している	「A 区と B 区が 重複している...」
	地域	行政区画または住所フィルタリング（都道府県，市区町村）	「京都府京都市...」
S2 計量的	距離値	数値で表される距離	「駅から 3 km の位置...」
	距離単位	距離の単位（km，m，分）	「駅から3km 圏内 / 徒歩10 分以内...」
	閾値タイプ	閾値の種類（以内，超過）	「3km 圏内 / 3km 超過...」
S3 方向性	絶対方向	基本方位（北，北東）	「駅の北側 に位置する...」
	相対方向	自己中心的な方向（左，右，正面，背面，上り坂）	「駅の右側 に位置する...」
S4 ルート	経由地	ルート上の中間地点	“東寺を経由して…”
	沿道	線形のランドマーク（道路・河川）との関連性	“堀川通／鴨川沿い…”
		エンティティ構成要素	
E1 種別	エンティティ種別	エンティティの分類（例：施設カテゴリ）	“寺院／ベーカリーはどこですか？”
E2 人物	名称	場所に関連する人物の名前	“空海と関連…”
	関係種別	関係タイプ（創立者／訪問者／関連人物／…）	“空海によって創立された…”
E3 イベント	名称	歴史的イベントの名称	“応仁の乱に関連する…”
	関係種別	関係タイプ（発生場所／戦場／会談場所／…）	“応仁の乱が発生した場所…”
E4 属性	制度的	制度的または法的ステータス（指定・区域区分）	“重要文化財に指定／風致地区内”
	物理的	物理的・形態的特性（形状・材質・年代）	“木造構造／796 年建立”
	機能的	機能的役割または用途（目的・収容能力）	“礼拝／宿泊に使用”
	運用的	運用条件（営業時間・料金・アクセシビリティ）	“8:00〜17:00 営業／車椅子対応”
	知覚的	知覚的または体験的特性（風光明媚・静寂・歴史的）	“風光明媚で静寂な場所”
	メタ	識別子・別名・データ来歴	“Wikidata QIDを保持／別名：教王護国寺”

表 4　空間要素およびエンティティ構成要素の分類体系．

B　モデルのパラメータ

表 5 に，本実験で使用した両モデルのハイパーパラメータおよび API 設定の詳細を示す．

パラメータ	値
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

表 5　評価対象モデルのハイパーパラメータ設定

C　実験で使用したプロンプト例

実験では，2 モデル共通で次のプロンプトを用いた．指示の部分は全ての質問で共通であり，質問内容は質問ごとに異なる．

> 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．
> 大分県由布市にある由布院駅から 0.5km 圏内にある仏教寺院はどこ？