GeoBenchmark : Analyse des grands modèles de langage pour les connaissances géospatiales

Ayomide Abayomi-Alli[♣], Jose G. Moreno[♢], Karim Radouane[♢] and Lynda Tamine[♢]
[♣]Jean Monnet University, [♢]University of Toulouse, IRIT, UMR 5505 CNRS, France
[♣]abayomialliayomide12@gmail.com, [♢]firstname.lastname@irit.fr

RÉSUMÉ

Les grands modèles de langage (LLM) font preuve d'une forte capacité à restituer des connaissances générales, mais ont du mal à traiter les connaissances géospatiales concrètes. Afin de mesurer et d'aider à tester les connaissances spatiales des LLM, nous présentons **GeoBenchmark**, un benchmark permettant d'évaluer le bon sens géographique selon trois relations spatiales fondamentales : la **direction**, la **distance** et la **topologie**. À partir des données extraites de YAGO2geo et des géométries des quartiers de l'*Ordnance Survey*, les relations spatiales ont été formalisées sous forme de triplets structurés et systématiquement transformées en paires de questions-réponses équilibrées de type binaire (Oui/Non) et à choix multiples (QCM). En outre, nous prenons en compte les questions atomiques et composites en fonction du nombre de relations spatiales impliquées. L'ensemble de données résultant comprend 26 000 échantillons binaires et 13 000 échantillons MCQ, répartis uniformément entre les niveaux de relations atomiques, binaires et ternaires. Nous établissons des références avec **LLaMA-8B** et **Mistral-7B** sous prompting zero-shot, obtenant une précision de 52 à 63 % sur les questions atomiques, mais inférieure à 35 % sur les relations ternaires, ce qui révèle la compréhension spatiale compositionnelle limitée des modèles et leur fort biais d'option. **GeoBenchmark** fournit une ressource complète et reproductible pour tester et faire progresser le sens commun géographique des LLM, ouvrant la voie à de futures recherches sur l'exploration spatiale et géographique des LLM ainsi que sur l'édition des connaissances. Ceci est le résumé de l'article "GeoBenchmark : Probing Large Language Models for Geo-Spatial Knowledge" publié dans la conférence LREC2026 (Abayomi-Alli *et al.*, 2026)

ABSTRACT

GeoBenchmark : Probing Large Language Models for Geo-Spatial Knowledge

Large Language Models (LLMs) demonstrate strong factual recall of general-purpose knowledge but struggle with grounded geospatial knowledge. To measure and help probe LLMs for spatial knowledge, we present **GeoBenchmark**, a benchmark for evaluating geographic commonsense along three core spatial relations : **direction**, **distance**, and **topology**. Using data extracted from YAGO2geo and Ordnance Survey ward geometries, spatial relations were formalized as structured triplets and systematically transformed into balanced binary (Yes/No) and Multiple-Choice (MCQ) question-answer pairs. Besides, we consider atomic and composite questions based on the number of spatial relations involved. The resulting dataset comprises **26k** binary and **13k** MCQ samples, uniformly distributed across atomic, binary, and ternary relation levels. We establish baselines with **LLaMA-8B** and **Mistral-7B** under zero-shot prompting, achieving 52-63% accuracy on atomic questions but below 35% on ternary relations, which exposes the models' limited compositional spatial understanding and strong option bias. **GeoBenchmark** provides a comprehensive, reproducible resource for probing

and advancing LLMs' geographic commonsense, paving the way for future research in spatial and geographic probing of LLMs as well as knowledge editing. This is the summary of the published paper "GeoBenchmark : Probing Large Language Models for Geo-Spatial Knowledge" in the LREC2026 proceedings (Abayomi-Alli *et al.*, 2026).

MOTS-CLÉS : LLM, raisonnement géospatial, compréhension spatial, GeoSPARQL.

KEYWORDS: LLM, geospatial reasoning, spatial commonsense, GeoSPARQL.

ARTICLE ACCEPTÉ À : The 2026 International Conference on Language Resources and Evaluation (LREC 2026).
URL : https://lrec2026.info/

Références

ABAYOMI-ALLI A., MORENO J. G., RADOUANE K. & TAMINE-LECHANI L. (2026). Geobenchmark : Probing large language models for geo-spatial knowledge. In *Proceedings of the 2026 International Conference on Language Resources and Evaluation (LREC 2026)* : LREC 2026.

GeoBenchmark: Probing Large Language Models for Geo-Spatial Knowledge

Ayomide Abayomi-Alli[♣], Jose G. Moreno[♢], Karim Radouane[♢] and Lynda Tamine[♢]
[♣]Jean Monnet University, [♢]University of Toulouse, IRIT, UMR 5505 CNRS, France
[♣]abayomialliayomide12@gmail.com, [♢]firstname.lastname@irit.fr

Abstract

Large Language Models (LLMs) demonstrate strong factual recall of general-purpose knowledge but struggle with grounded geospatial knowledge. To measure and help probe LLMs for spatial knowledge, we present **GeoBenchmark**, a benchmark for evaluating geographic commonsense along three core spatial relations: **direction**, **distance**, and **topology**. Using data extracted from YAGO2geo and Ordnance Survey ward geometries, spatial relations were formalized as structured triplets and systematically transformed into balanced binary (Yes/No) and Multiple-Choice (MCQ) question-answer pairs. Besides, we consider atomic and composite questions based on the number of spatial relations involved. The resulting dataset comprises **26k** binary and **13k** MCQ samples, uniformly distributed across atomic, binary, and ternary relation levels. We establish baselines with **LLaMA-8B** and **Mistral-7B** under zero-shot prompting, achieving 52-63% accuracy on atomic questions but below 35% on ternary relations, which exposes the models' limited compositional spatial understanding and strong option bias. **GeoBenchmark** provides a comprehensive, reproducible resource for probing and advancing LLMs' geographic commonsense, paving the way for future research in spatial and geographic probing of LLMs as well as knowledge editing.

Keywords: LLM, geospatial reasoning, spatial commonsense, GeoSPARQL, UK metropolitan wards

1. Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in tasks such as question answering (Jiang et al., 2021; Yasunaga et al., 2021), summarization (Zhang et al., 2024; Hartl and Kruschwitz, 2022), and factual retrieval (Roberts et al., 2020; Petroni et al., 2019). However, a persistent gap between LLMs' great factual recall and their low capacity for grounded *spatial knowledge* understanding when they are presented with *geographic queries* (Manvi et al., 2024a). Models often provide confident yet incorrect answers to simple directional questions or fail to interpret topological relations between geographic entities (Ramrakhiyani et al., 2025). For instance, they might misidentify a UK metropolitan district despite clear instructions or claim that "Paris is west of Berlin." These mistakes reveal an excessive dependence on shallow memorization in linguistic associations.

Why it matters. Accurate modeling of spatial relationships is essential for intelligent systems that interact meaningfully with the physical world, such as navigation assistants, geographic information retrieval, and environmental monitoring applications. Humans rely on what is commonly referred to as *geographic commonsense*, the innate ability to understand direction, distance, and containment in everyday contexts. As Abdou et al. (2021) demonstrated, models can encode perceptual-like representations (e.g., color) without genuine grounding, and Petroni et al. (2019) shows that LLMs primarily memorize factual associations as shallow key-value pairs offering fluent generation and impressive factual recall. Despite progress in factual recall, no existing benchmark systematically evaluates whether LLMs can understand spatial relations in a structured, reproducible way.

Our contributions. This work introduces a benchmark designed to probe the geographic commonsense of LLMs. A balanced **Yes/No** and **Multiple-Choice (MCQ)** question-answer format is created by formalizing spatial relations as structured triplets (S, R, O) that contain three basic relations: **direction**, **topology**, and **distance**. The resulting dataset contains over **26k binary** and **13k MCQ** pairs, covering atomic, binary, and ternary relation composition to test for composite geographic understanding. Using **LLaMA-8B (3.1 version)** and **Mistral-7B**, we evaluate model performance under zero-shot conditions and examine systematic biases such as MCQ option bias and uneven token representation. The benchmark provides a comprehensive and reproducible framework for diagnosing relational spatial reasoning in language models and supports systematic evaluation across model scales.

2. Related Work

2.1. LLMs and Spatial Knowledge Probing

Large Language Models (LLMs) have been studied as repositories of factual and semantic associations (Petroni et al., 2019), showing that they can retrieve surface facts such as country capitals. However, these abilities often reflect memorized

linguistic co-occurrences rather than grounded understanding (Abdou et al., 2021; Yildirim and Paul, 2024). Studies have shown that while LLMs may capture broad perceptual or structural regularities, they lack the inductive biases necessary to form robust "world models." This concern echoes broader critiques on the over-reliance on shallow statistical associations in LLMs (Bender et al., 2021).

When extended to spatial domains, these weaknesses become more visible. Gurnee and Tegmark (2024) observed that LLMs encode coarse geographic and temporal cues but frequently distort spatial relationships. Manvi et al. (2024a) reported systematic geographic bias, particularly in latitude–longitude prediction tasks, while Ramrakhiyani et al. (2025) conducted a comprehensive study of geographic QA and confirmed persistent failures in direction, distance, and topology fact recall. Although certain attributes, such as coordinates, can be linearly decoded from hidden activations, relations like population size or distance remain poorly represented. Collectively, prior work (Manvi et al., 2024b; Godey et al., 2024) suggests that current models rely on textual associations rather than grounded spatial knowledge. ‘

2.2. Existing Geo Benchmarks and Standards

Benchmark datasets have been central to diagnosing these limitations. The GeoQuestions1089 dataset (Kefalidis et al., 2023) introduced 1,089 SPARQL-aligned queries spanning topological, numeric, and thematic relations. Despite its adoption, later analyses identified major limitations: phrasing inconsistencies, incomplete question contexts, and performance instability for numeric questions. In particular, aggregate and population queries frequently caused hallucinations. To better understand these issues, we conducted a focused re-evaluation of selected GeoQuestions1089 subsets; a detailed analysis is provided in Appendix B.

Beyond GeoQuestions1089, semantic standards such as GeoSPARQL (Open Geospatial Consortium (OGC), 2012) (Nicholas J. Car et al., 2023) define formal spatial predicates (e.g., *within*, *touches*, *distance*) for reasoning over RDF data. These standards facilitate integration with knowledge graphs and retrieval-augmented methods, while complementary efforts explore multimodal grounding with maps and sensor data. Nevertheless, existing resources remain fragmented and lack balanced coverage across spatial relation types, negative examples, and standardized templates, which represent the necessary conditions for systematic evaluation.

Building on these observations, our work introduces a balanced, reproducible benchmark that unifies direction, topology, and distance knowledge probing within a consistent triplet schema. The next sections detail the benchmark design methodology and how structured triples are transformed into standardized question-answer templates with controlled negative transitions and balanced coverage.

3. Dataset Design

This section presents the overall design of the benchmark, detailing how spatial relations are represented, composed, and instantiated into evaluation-ready question–answer pairs. The objective is to provide a transparent and reproducible structure that systematically probes geographic commonsense across increasing levels of compositionality.

3.1. Triplet Schema

To represent spatial relations in a unified and interpretable manner, we formalize them as structured triplets as follows:

$$(S, R, O), \quad R \in \{R_{\text{dir}}, R_{\text{top}}, R_{\text{dis}}\} \tag{1}$$

where S denotes the *subject* entity, O the *object* entity, and R the spatial relation. We focus on three fundamental relation types: directional (R_{dir}), topological (R_{top}), and distance-based (R_{dis}). For instance, the triplet ($\textit{London}, \textit{north}, \textit{Paris}$) encodes that *“London lies north of Paris”*, while ($\textit{Birmingham}, \textit{within}, \textit{England}$) expresses that *“Birmingham is contained within the spatial boundaries of England”*. This representation provides a canonical structure aligned with geospatial knowledge graphs and RDF standards such as GeoSPARQL, supporting transparent and scalable analysis.

Each relation category is associated with a fixed vocabulary of **tokens** that operationalize geographic commonsense into interpretable linguistic labels. Tokens were selected to be both compact and balanced, enabling controlled evaluation of model performance across spatial relation types. They were derived from prior benchmarks (GeoQuestions1089), extended with ConceptNet[1], and validated using OGC GeoSPARQL specifications to ensure consistency with existing geographic standards.

Concept	Example Tokens	Relation Notation
Direction	north, south, east, west	R_{dir}
Topology	within, borders	R_{top}
Distance	near, close, far, distant	R_{dis}

Table 1: Token definitions for spatial relations.

[1] https://conceptnet.io/

Example tokens in Table 1 define the linguistic surface forms used to instantiate each relation type. Together with the triplet schema, they form the backbone of the benchmark, providing a consistent foundation for generating atomic, binary, and ternary relation questions.

3.2. Compositional Question Design

The benchmark builds upon the triplet schema to create question–answer pairs that increase in complexity, from single relations to multi-relation compositions. This hierarchical design allows for the progressive assessment of a model's capacity to synthesize multiple spatial relations, including direction, topology, and distance, into coherent reasoning. Implementation details on question template generation and sampling strategies are provided in Section 4.3.

3.2.1. Atomic Questions

Atomic triples encode a single spatial relation between two geographic entities, such as (S, R, O). For example, an atomic question such as *"London is north of Paris"* captures a directional relation. These instances serve as the fundamental reasoning units upon which more complex compositions are built.

3.2.2. Two-Relation Composite Questions

Two-Relation (binary) compositions integrate two relations simultaneously, testing whether models can correctly recall facts across two spatial combinations, as represented as follows:

$$S' \;\; R_1 \;\; O_1 \;\; R_2 \;\; O_2 \tag{2}$$

where $R_1, R_2 \in \{R_{\text{dir}}, R_{\text{dis}}, R_{\text{top}}\}$ and O_1, O_2 are the respective object entities. Each S' is formed by pairing compatible atomic triples while preserving token balance. For instance, a question like *"Which city is east of Paris and within France?"* combines directional and topological reasoning. Other variants include *Direction & Distance* and *Topology & Distance* compositions. This structure evaluates whether LLMs can integrate independent spatial relations without losing logical consistency.

3.2.3. Three-Relation Composition

Three-concept (ternary) compositions extend this schema to encompass all three spatial relations as represented in Equation 3.

$$S'' \;\; R_1 \;\; O_1 \;\; R_2 \;\; O_2 \;\; R_3 \;\; O_3 \tag{3}$$

where R_1, R_2, R_3 represent direction, distance, and topology (in any permutation), and O_1, O_2, O_3 are their respective object entities. By construction, $S' \subseteq S''$, ensuring consistency between binary and ternary subsets. A representative ternary example is *"Which city is north of London, within England, and far from Cardiff?"* Sampling followed a stratified procedure to maintain uniform token distribution and balanced relation coverage across all spatial levels, ensuring fair and interpretable evaluation.

3.3. Benchmarking Formats

Each triplet configuration is rendered in two complementary question–answer formats to disentangle language understanding and geography knowledge understanding: a **Yes/No verification** format, where the model determines whether a stated relation holds true, and a **Multiple-Choice (MCQ)** format, where the model selects the correct subject S from a set of distractions as options.

Controlled perturbations are incorporated into both formats to enhance robustness. In Yes/No tasks, *negative transitions* are generated by swapping relations or entities, with a dedicated *transition* column documenting the applied modification. In MCQ tasks, distractor options are created using random, partial, and hard negatives, recorded in an *options* column for transparency and reproducibility. These augmentations enable systematic analysis of positional and lexical biases while maintaining consistent evaluation conditions.

Together, these design choices capture complementary aspects of LLM spatial understanding: logical consistency through binary verification and selective retrieval through multiple-choice inference. This integrated framework provides a rigorous foundation for assessing geographic commonsense in LLMs.

4. Dataset Creation

This section details the process followed to implement the benchmark design and generate the final dataset. The process involved three phases: (i) extraction and pre-processing geospatial data, (ii) operationalization of spatial relations, and (iii) constructing QA templates in Yes/No and MCQ formats with systematic augmentation strategies.

4.1. Data Sources and Extraction

4.1.1. YAGO2geo Knowledge Base

We selected YAGO2geo[2] as the core knowledge base due to its integration of structured semantic information with precise geospatial geometries. YAGO2geo extends YAGO2 by incorporating administrative datasets from the United King-

[2]https://yago2geo.di.uoa.gr/

dom, Greece, and the Republic of Ireland, alongside global administrative boundaries from GADM and additional geographic features from OpenStreetMap. This multi-source design provides a country-agnostic foundation and enables straightforward cross-regional expansion of the benchmark without modifying the extraction pipeline.

4.1.2. OS Metropolitan District Wards

For initial development, we chose the United Kingdom subset of YAGO2geo, which holds more than 500,000 RDF triples describing geographic entities with high-precision polygon geometries in Well-Known Text (WKT) format. From this subset, we chose the Metropolitan District Wards of the Ordnance Survey, a mid-level administrative division with 809 validated subdivisions. These wards offer consistent polygon geometries, clear hierarchical relationships (across districts and European regions), and a level of spatial detail sufficient for adjacency, containment, distance, and directional reasoning.

Limiting the first release to a single administrative division facilitates deterministic extraction and validation of the data pipeline before scaling up to other countries covered by YAGO2geo.

4.1.3. Extraction via SPARQL and GraphDB

Data was extracted with automated SPARQL queries on GraphDB[3], following practices established in RDF-based QA systems. All extraction was automated using the SPARQLWrapper[4] in Python in order to facilitate reproducibility. Other tools include: GraphDB, SPARQLWrapper, Shapely[5], WKT geometry retrieval API, and custom modules were developed for bearing computation, spatial relation mapping.

4.2. Relation Operationalization

4.2.1. Distance (R_{dis})

Distance was computed between all unordered ward pairs using `geof:distance` function. With 809 wards, this produced $\binom{809}{2} = 326,836$ pairs. Thresholds for tokens were derived from GeoQuestions1089 (Kefalidis et al., 2023), which operationalized *near* as under 5 km and *far* as over 5 km. Empirical checks using online tools[6] confirmed broader semantic ranges (e.g., 24 km ≈ 15 miles in regional planning). We adopted a graded scheme: *near* (smaller or equal than 5km), *close* (between 5km and 25km), *far* (between 25km and 80km), and *distant* (greather than 80km). This yielded to **326,836** distance records.

4.2.2. Direction (R_{dir})

Relative direction was determined via centroid-based azimuth bearings. For each ward pair, Shapely centroids were computed, and bearings were mapped into cardinal categories following azimuth conventions (Topa Blog Editors, 2025). We considered North: $([315°, 360°] \cup [0°, 45°])$, East: $([45°, 135°])$, South: $([135°, 225°])$, and West: $([225°, 315°])$. This generated **653,672** directional records. Figures 1 and 2 depict the mapping of bearings to cardinal directions and an example of directional alignment between wards.

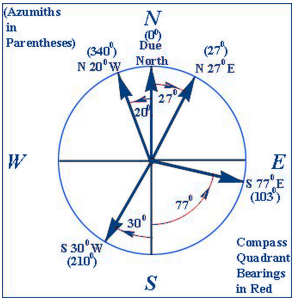


Figure 1: Cardinal mapping from azimuth bearings.

4.2.3. Topology (R_{top})

Topology relations were retrieved directly from GeoSPARQL-compliant predicates in YAGO2geo. We used `geo:sfTouches` (for adjacency or borders) and `geo:sfWithin` (for containment). This produced **4,480 border** and **1,357 within** relations, totaling **5,837**. Malformed WKT entries were excluded.

Relation triples were mapped to natural language QA templates at three levels of compositionality: atomic, two-relation, and three-relation. Each was instantiated in Yes/No and MCQ formats.

4.3. QA Template Generation

Building on the extracted relation triples, question–answer (QA) templates were instantiated in two formats: **Yes/No verification** and **Multiple-choice (MCQ)**. This ensured systematic evaluation across increasing levels of compositions: atomic, two-relation, and three-relation. The same set of triples was reused across formats to guarantee

[3] https://graphdb.ontotext.com/documentation/11.0/

[4] https://sparqlwrapper.readthedocs.io/en/latest/

[5] https://github.com/shapely/shapely

[6] https://milesofme.com

	place1	place2	bearing	relation
63	Prescot	Halesowen North	334.12	north

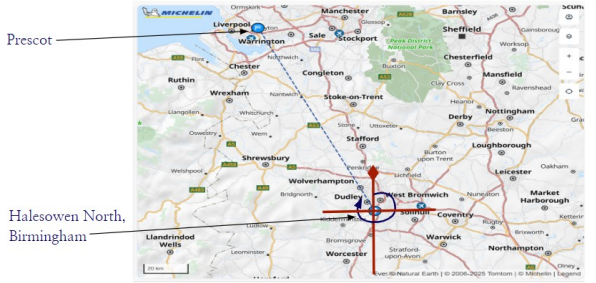


Figure 2: Example of directional mapping between wards. Bearing values allow the calculation of the relation value.

comparability, with only question framing and label schema differing.

4.3.1. Negative and distractors generation

We generated negative examples and distractors to balance the positive binary samples and to integrate candidates into the MCQ. Our strategy is slightly adapted to the number of relations involved, but mainly based on question type.

Negative samples for binary questions were obtained by randomizing the relation type to ensure a balanced positive–negative ratio and controlled difficulty for verification tasks as follows:

- *Direction:* flip subject/object (25%) or replace token (25%).
- Distance: swap distance tokens (25% for near and far, 25% for close and distant).
- *Topology:* for *within* (25%) flip places, the rest replace token (*within* and *borders* because border relation is symmetrical).

While MCQ distractors were obtained as follows:

- Hard negatives: another *subject* with the *same relation* to the same *object*.
- Random: a randomly sampled *subject*.
- Opposite negatives: for topology, swap *within* and *borders*.

4.4. Benchmark Probing QA Dataset

After constructing the *Atomic*, *Two-Relation*, and *Three-Relation* subsets, a final step merged them into unified datasets for each QA format (Yes/No, MCQ). The goal was to build a single resource that is fully labeled, auditable, filterable by relation type, and ready for probing experiments.

Operational Schema Each record in the consolidated dataset is annotated with: (1) a `relation` label indicating the composition level (1 = Atomic, 2 = Two-Relation, 3 = Three-Relation); (2) binary flags as columns for the presence of Direction (`dir`), Distance (`dis`), and Topology (`top`); (3) metadata columns such as `transition` (for Yes/No) and `option_source` (for MCQ) as presented in Table 2.

Question	Answer	Transition	dir	dis	top
Is Dodworth east of Stockbridge?	Yes	no_change	1	0	0

Table 2: Example of record from the Yes/No benchmark, where Directional relation is present.

As illustrated in Figures 3 and 4, the distribution of negative transitions (Yes/No) and distractor sources (MCQ) confirms balanced augmentation strategies across both formats. In MCQ, the third distractor option is typically drawn from intermediate difficulty levels between partial and hard negatives.

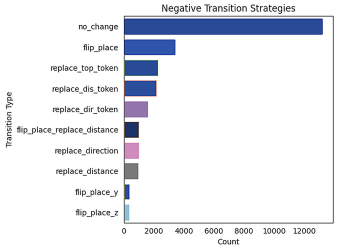


Figure 3: Distribution of negative transition strategies in the Yes/No dataset.

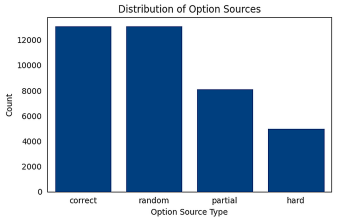


Figure 4: Distribution of distractor sources in the MCQ dataset.

Summary Statistics. Table 3 reports dataset sizes across formats and concept levels. Note that we sample 10,000 triplets for binary and 5,000 MCQ for the atomic relation. In total, the Yes/No dataset contains 26,252 records, while the MCQ dataset contains 13,126 records. Coverage is balanced across direction, distance, and topology, supporting comprehensive probing.

5. Experiments and Results

We evaluated two mid-scale instruction-tuned open-weight models, **LLaMA-8B** and **Mistral-7B**, as an initial testbed for validating the proposed benchmark under constrained capacity settings. Both models were assessed using *zero-shot* prompts across two structured QA formats: **binary (Yes/No)** and **multiple-choice (MCQ)**. These controlled settings allow us to examine factual recognition, selective retrieval, and compositional reasoning across direction, distance, and topology relations while establishing a baseline for future scaling experiments.

5.1. Prompting Strategy

Model evaluation followed consistent prompt templates, ensuring comparability across QA tasks. Below is an example of the binary QA prompt:

Instruction

You are a geography expert. Answer the following geography question about cities and places in the metropolitan district wards in the United Kingdom.
Respond with only **'yes'** or **'no'** — do not explain or justify your answer.
No extra text.
Question: {question}
Answer:

For MCQ task, models were prompted to select the correct option among three candidates:

Instruction

You are a geography expert. Based on Ordnance Survey data of the United Kingdom metropolitan district wards, choose the correct answer to the question below.
Select only one option: A, B, or C.
Question: {question}
Options: {options}
Answer:

All models were evaluated using deterministic decoding (temperature = 0) to ensure consistency across runs. Responses were constrained to minimal formats (binary or single-choice) to standardize evaluation across architectures.

Chain-of-thought prompting was intentionally not enabled. Our evaluation focuses on zero-shot implicit spatial competence under standardized response constraints rather than scaffolded reasoning that may vary across models.

5.2. Evaluation Considerations

Model performance was evaluated across both **binary (Yes/No)** and **multiple-choice (MCQ)** tasks, using accuracy-based metrics designed to capture not only correctness but also model bias.

For the **Yes/No binary questions**, accuracy was computed separately for affirmative and negative instances to detect asymmetries in response patterns. Beyond raw accuracy, *logical consistency* was assessed by pairing related question variants e.g., *"Is Birmingham within England?"* versus its negated counterpart. Each pair was categorized into one of four outcomes: fully consistent (both correct), Yes-only correct, No-only correct, or both wrong. This measure reflects the model's internal coherence and its ability to maintain consistency across statements.

For the **MCQ format**, accuracy was computed as the percentage of cases where the predicted option matched the gold label (actual correct option). *Option bias* was quantified by comparing the distribution of predicted answers (A, B, C) against the gold distribution, identifying whether models favored specific positions irrespective of content. *Recall by option* examined per-choice accuracy to highlight disparities in retrieval strength. To probe robustness, accuracy was also analyzed by *distractor position sensitivity*, comparing model performance under different arrangements of correct, hard, partial, and random distractors. Finally, *out-of-context predictions*; responses outside the expected option range were tracked, alongside *relation-token accuracy*, which groups results by spatial terms (e.g., "near," "within") to reveal relation-level weaknesses.

Together, these metrics provide both quantitative and behavioral insights, enabling a nuanced

Relation Level	Yes/No QA	MCQ QA	Distinct Entities / Notes
Atomic (1 Relation)	10,000	5,000	783 subjects, 820 objects; 10 relation tokens evenly distributed (500 each).
Two-Relation	14,844	7,422	Balanced coverage across Direction+Distance, Direction+Topology, and Topology+Distance configurations.
Three-Relation	1,408	704	704 subjects, 512 objects; triplets span all three relation families simultaneously.
Total QA Pairs	**26,252**	**13,126**	Combined total of 39,378 question–answer pairs (Yes/No + MCQ).
Relation Token Distribution (All Subsets)			
Direction Tokens	North (1,859), South (1,859), East (1,749), West (1,760)		
Distance Tokens	Near (2,164), Close (2,196), Far (2,119), Distant (2,111)		
Topology Tokens	Within (3,017), Borders (3,124)		

Table 3: Benchmark composition summary across relation levels and relation types.

analysis of how LLMs handle compositionality and positional uncertainty across QA formats.

5.3. Main Results

Metrics	LLaMA-8B	Mistral-7B
1. Recall Accuracy on Prediction		
Yes	30.45%	50.35%
No	79.81%	59.63%
2. Logical Consistency (%)		
Consistent Prediction	14.53%	21.03%
Correct Yes Only	11.69%	16.99%
Correct No Only	41.03%	33.97%
Both Wrong Prediction	32.75%	28.01%

Table 4: Performance and consistency of LLaMA-8B and Mistral-7B for binary questions.

Binary QA. Table 4 summarizes binary inference outcomes. Mistral achieved higher recall accuracy for affirmative ("Yes") predictions (50.35%), whereas LLaMA-8B performed markedly better on negative ("No") predictions (79.81%). This asymmetry suggests that Mistral-7B tends to over-affirm spatial statements, while LLaMA-8B is more conservative in rejection. Logical consistency remained low for both models, with Mistral-7B scoring slightly higher (21.03% vs. 14.53%), revealing frequent contradictions between corresponding Yes and No pairs.

The confusion matrix analysis in Table 5 provides deeper insight into how both models handle binary QA tasks. LLaMA-8B and Mistral-7B exhibit high confidence but inconsistent spatial verification, confirming that prediction certainty does not necessarily imply accurate spatial knowledge understanding. LLaMA-8B predicts "No" more often, achieving a higher F1 for that class (67.6%) but weak performance on "Yes" (38.9%), indicating a cautious rejection bias. Mistral-7B is more balanced, with improved "Yes" accuracy (F1 = 53.2%). However, the total false-positive count (*True No* = 5298 predicted as "Yes") reveals an over-affirmation bias: the model often assumes that spatial statements are true even when they are not. This suggests that Mistral-7B can reject incorrect relations but lacks precise boundary discrimination. These results confirm that current instruction-tuned LLMs rely on linguistic pattern matching rather than true spatial verification.

LLaMA-8B	Pred Yes	Pred No	F1 (%)
True Yes	3997	9129	38.9
True No	2650	10476	67.6
Total	6647	19605	—
Mistral-7B	**Pred Yes**	**Pred No**	**F1 (%)**
True Yes	6609	6517	53.2
True No	5298	7828	61.9
Total	11907	14345	—

Table 5: Performance per option rank (A, B, and C) of LLaMA-8B and Mistral-7B for MCQ. Rows correspond to true labels, columns to predicted labels.

MCQ QA. MCQ results in Table 6 show strong positional bias across models. Both systems overpredicted option A (above 50% of responses) despite the balanced gold (accurate option position) distribution of 33.3%. Accuracy dropped sharply when the correct option appeared later among distractors, indicating sensitivity to positional priors rather than spatial cues. LLaMA-8B consistently outperformed Mistral-7B under complex distractor configurations, suggesting better calibration when uncertainty was introduced. As shown in Table 6, both models exhibit a clear positional bias, with predictions heavily concentrated on the first option (A) across all true labels. This skew indicates a lexical or positional heuristic rather than actual spatial knowledge understanding. For LLaMA-8B, option A was predicted **7,084** times, compared to only **1,797** predictions for option C, resulting in relatively strong F1 scores for A (54.9%) and B (52.4%) but a significant drop for C (37.6%). Mistral-7B follows a similar pattern, with **7,418** predictions for A against **1,603** for C, reflecting even weaker discrimination (A: 50.4%, B: 43.6%, C: 27.6%). LLaMA-8B's comparatively higher F1 scores across all classes indicate more stable calibration under uncertainty,

while Mistral-7B demonstrates sharper declines on harder distractors, suggesting weaker internal representation of compositional spatial cues.

LLaMA-8B	Pred A	Pred B	Pred C	F1 (%)
True A	3126	890	294	54.9
True B	1799	2263	337	52.4
True C	2159	1087	1166	37.6
Total	7084	4240	1797	—
Mistral-7B	**Pred A**	**Pred B**	**Pred C**	**F1 (%)**
True A	2955	979	374	50.4
True B	2145	1853	400	43.6
True C	2318	1263	829	27.6
Total	7418	4095	1603	—

Table 6: Confusion matrices for MCQ option prediction with per-class F1 scores and totals. Rows represent true labels (A–C), columns predicted labels.

Relation	MCQ (%)		Yes/No (%)	
	LLaMA-8B	Mistral-7B	LLaMA-8B	Mistral-7B
Within	58.0	51.7	73.8	68.5
Borders	63.6	55.7	42.8	47.1
Near	53.6	46.0	57.9	57.5
Close	49.7	43.4	53.3	51.3
Far	49.1	38.4	49.8	56.2
Distant	47.5	49.0	57.5	51.8
North	37.7	35.7	51.5	60.8
South	34.4	33.5	51.4	49.1
East	38.4	35.4	50.6	51.1
West	36.5	33.4	51.2	54.0

Table 7: Accuracy by relation token across QA formats for LLaMA-8B and Mistral-7B.

5.4. Results on Relation-Token subsets

Token-level accuracies in Table 7 confirm that topological relations (*within*, *borders*) are better encoded than directional or distance relations. Interestingly, Mistral-7B slightly surpasses LLaMA-8B on northern and western relations in binary form, suggesting localized robustness for certain directional embeddings. These findings reinforce that current LLMs recall familiar geographic associations but lack metric precision when spatial distance must be inferred.

5.5. Compositional Relation Trends

We can see from Table 8 that across all relation compositionality levels, accuracy decreases monotonically from atomic to multi-relation queries, illustrating limited capacity for integrating multiple spatial cues coherently. This compositional decay aligns with similar trends in commonsense and temporal reasoning benchmarks, suggesting that LLMs learn spatial patterns as isolated facts rather than structured relational systems. The benchmark thereby quantifies this weakness, providing an interpretable gradient of difficulty for future research.

Metrics	MCQ (%)		Yes/No (%)	
	LLaMA-8B	Mistral-7B	LLaMA-8B	Mistral-7B
Overall	49.9	43.0	55.1	55.0
Relation-Level				
Atomic	40.7	37.7	53.5	55.2
Two-Rel.	55.0	45.6	56.2	54.9
Three-Rel.	61.2	51.5	54.5	54.8
Relation Type				
Topology	60.5	48.7	60.6	58.0
Distance	49.9	41.7	54.5	54.2
Direction	48.4	43.9	52.1	53.0

Table 8: Accuracy comparison between LLaMA-8B and Mistral-7B across QA formats.

5.6. Overall Benchmark Insights

Overall patterns across QA formats and relations, several consistent trends emerge between LLaMA-8B and Mistral-7B (Table 8).

First, overall accuracy remains moderate at approximately 50% for MCQs and 55% for Yes/No questions, showing that both models achieve above-chance but far from reliable spatial reasoning performance. LLaMA-8B slightly outperforms Mistral-7B on MCQs (+6.9 points), while both models converge on near-identical binary accuracy, indicating that scaling alone does not eliminate factual errors when linguistic cues are ambiguous.

Second, compositional depth clearly magnifies difficulty. Accuracy improves from *atomic* to *two-relation* questions (40–55%), but plateaus or declines at the *three-relation* level for binary prompts. This pattern supports the hypothesis that compositional complexity amplifies factual errors, confirming that multi-relation questions are not trivially reducible to memorized patterns.

Finally, averaged across all tokens, Yes/No accuracy (55%) consistently exceeds MCQ accuracy (48%), implying that forced-choice generation introduces additional distractor sensitivity. This aligns with observed positional biases—models tend to favor central or initial options—indicating structural rather than stochastic errors. Other aggregate statistics appear in Tables 4–9. Overall, these results reinforce three core insights: (i) compositionality of spatial relations remains a bottleneck; (ii) spatial language is unevenly encoded, with topological relations dominating; and (iii) model improvements appear format-dependent rather than conceptually grounded.

6. Conclusion

In this paper, we present a benchmark for LLM probing of geo-spatial knowledge based on atomic, two-relations, and three-relation compositions. We

Metrics	LLaMA-8B	Mistral-7B
1. Gold Label Distribution		
Option A	33.3%	33.3%
Option B	33.3%	33.3%
Option C	33.3%	33.3%
2. Bias in Option Prediction Distribution		
Option A	53.96%	56.51%
Option B	32.30%	31.19%
Option C	13.69%	12.21%
3. Recall Accuracy on Options		
Option A	72.52%	68.59%
Option B	51.44%	42.13%
Option C	26.42%	18.79%
4. Out-of-Context Predictions		
Out of option	0.04%	0.08%
5. Accuracy Based on Option Position		
correct, hard, random	72.74%	66.54%
correct, partial, random	70.65%	69.52%
correct, random, hard	71.89%	66.30%
correct, random, partial	74.55%	70.25%
hard, correct, random	31.25%	33.60%
hard, random, correct	16.53%	11.29%
partial, correct, random	61.35%	43.79%
partial, random, correct	30.42%	22.27%
random, correct, hard	36.18%	36.65%
random, correct, partial	63.71%	49.21%
random, hard, correct	16.68%	12.95%
random, partial, correct	34.22%	23.35%
6. Partial vs Hard Distractors Accuracy		
Hard + Random	40.72%	37.74%
Partial + Random	55.61%	46.14%

Table 9: Detailed performances in the MCQ format for LLaMA-8B and Mistral-7B. We present the distribution, biases as well as the options predictions.

generate 39K questions and evaluate two LLMs, LLaMA-8B and Mistral-7B, on a subset of the proposed benchmark. Results show that LLaMA-8B dominates the performances on MCQ, and no clear winner is observed on a Yes/No question format when evaluated under our configuration setup. As future work, this framework will allow the evaluation of new foundations models but also specialized models with geographic knowledge. Other research direction includes the extension and validation of our benchmark in complex applications of geographic knowledge.

7. Ethics Statement

Although we used existing datasets, our use remains fair in a research context, and references to previous work will be kept. No user data or sensitive content was used during this work.

8. Data and Code Availability

To support reproducibility and further research on spatial reasoning in language models, we release the full dataset and relation operationalization scripts. It is available at: http://github.com/Premee12/Geobenchmark.git

9. Limitations

The current benchmark, while carefully designed to probe spatial knowledge in LLMs, has several limitations that motivate future extensions. Its **geographic scope** is currently limited to metropolitan wards within the United Kingdom. Although this region provides rich geospatial diversity, expanding to additional countries and larger geographic scales would further strengthen cross-linguistic and global generalization. While some questions may benefit from country-specific familiarity, the majority of tasks derive directly from WKT geometries and evaluate transferable relational spatial reasoning rather than factual geographic recall.

The benchmark also focuses primarily on **text-based spatial probing** with predefined tokens for binary and multiple-choice tasks. Future work could extend the dataset to include a wider range of spatial relations, more fine-grained directional and topological relationships, additional visual modalities, and free-form answer tasks, enabling a richer evaluation of model spatial knowledge.

Scalability and resource constraints remain a challenge, as larger-scale sampling could further reduce potential biases but requires significant time, cost, and computational resources, particularly for large or closed models. Finally, increasing **model diversity and applicability** represents an important future direction. Evaluating the dataset with larger and more diverse LLMs, and using it to improve existing models' spatial knowledge as a demonstration.

10. Acknowledgements

This work was fully funded by the Geo-R2LLM CHIST-ERA project. The experiments presented were partially conducted using the OCCIDATA platform administered by IRIT (CNRS/University of Toulouse).

11. Bibliographical References

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders

Søgaard. 2021. Can language models encode perceptual structure without grounding? a case study in color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 109–132, Online. Association for Computational Linguistics.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

Nathan Godey, Éric de la Clergerie, and Benoît Sagot. 2024. On the scaling laws of geographical representation in language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12416–12422, Torino, Italy. ELRA and ICCL.

William Gurnee and Max Tegmark. 2024. Language models represent space and time. In *International Conference on Learning Representations (ICLR)*.

Philipp Hartl and Udo Kruschwitz. 2022. Applying automatic text summarization for fake news detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2702–2713, Marseille, France. European Language Resources Association.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024a. Large language models are geographically biased. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David B. Lobell, and Stefano Ermon. 2024b. GeoLLM: Extracting geospatial knowledge from large language models. In *The 12th International Conference on Learning Representations (ICLR 2024)*.

Nicholas J. Car, Timo Homburg, Matthew Perry, John Herring, Frans Knibbe, Simon J.D. Cox, Joseph Abhayaratna, and Mathias Bonduel. 2023. OGC GeoSPARQL - A Geographic Query Language for RDF Data. OGC Implementation Standard OGC 22-047, Open Geospatial Consortium.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Nitin Ramrakhiyani, Vasudeva Varma, Girish Palshikar, and Sachin Pawar. 2025. Gauging, enriching, and applying geography knowledge in pre-trained language models. *Information Processing & Management*, 62(1):103103.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online. Association for Computational Linguistics.

Ilker Yildirim and LA Paul. 2024. From task structures to world models: what do llms know? *Trends in Cognitive Sciences*, 28(5):404–415.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

12. Language Resource References

Kefalidis, Sergios-Anestis and Punjani, Dharmen and Tsalapati, Eleni and Plas, Konstantinos and Pollali, Mariangela and Mitsios, Michail and Tsokanaridou, Myrto and Koubarakis, Manolis and Maret, Pierre. 2023. *GeoQuestions1089: Benchmarking Geospatial Question Answering Engines Using the Dataset*. International Semantic Web Conference (ISWC). Springer, distributed via LREC repository. PID

http://hdl.handle.net/20.500.11752/resource-geoquestions1089.

Open Geospatial Consortium (OGC). 2012. *OGC GeoSPARQL: A Geographic Query Language for RDF Data*. Open Geospatial Consortium. Open Geospatial Consortium. Accessed: 2025-08-19.

Topa Blog Editors. 2025. *What Is Azimuth and How to Determine It, How and in What Units It Is Measured*. Topa Blog. Accessed: 2025-08-02.

A. Appendices

A.1. Glossary of Terms

This glossary provides compact definitions of the core spatial and technical concepts used throughout the paper.

- **LLM:** Large Language Model trained on large-scale text corpora for language understanding and generation.
- **Geospatial reasoning:** The process of interpreting spatial relationships such as direction, topology, and distance.
- **Spatial commonsense:** Implicit knowledge about space that humans naturally use (e.g., cities have borders, north is opposite of south).
- **GeoSPARQL:** An OGC standard for representing and querying spatial information in RDF using relations such as *within* and *borders*.
- **UK metropolitan ward:** Administrative subdivisions of the UK, used as geographic units for spatial QA benchmarking.

B. Re-evaluation of GeoQuestions1089

To better understand limitations in existing geographic QA benchmarks, we conducted a focused re-evaluation of GeoQuestions1089 across three representative subsets: aggregate ("how many") queries, population-related numeric queries, and boolean relation queries. These subsets were selected to isolate numeric reasoning, quantitative retrieval, and categorical spatial reasoning.

Table 10: Extracted Evaluation Subsets from GeoQuestions1089

Subset	Count	Description
Aggregate ("Many")	143	Count-based numeric queries
Population	28	Numeric population questions
Boolean	181	Yes/No relational queries

B.1. Evaluation Protocol

We evaluated **Mistral-7B** and **LLaMA-2-7B** under both zero-shot and few-shot prompting settings ($k = 1, 2, 3$).

Evaluation Metrics. Performance was assessed using the following complementary metrics:

- **Compliance Rate:** Percentage of outputs matching the required response format (numeric or boolean).
- **Tolerance-Based Accuracy:** Numeric predictions accepted if within $\pm 20\%$ of the reference value.
- **Character-Level F1:** String similarity between prediction and reference.
- **Exact-Match Accuracy:** Strict correctness for boolean outputs.

B.2. Key Findings

- **Boolean Queries:** Achieved moderate stability across models, with accuracy consistently in the range of 49–52%.
- **Aggregate ("How many") Queries:** Displayed substantial instability. Zero-shot accuracy frequently fell below 15%, and performance degraded further under few-shot prompting.
- **Population Queries:** Showed variability across models, with frequent hallucinated or imprecise numeric outputs.

- **Compliance vs. Correctness:** Although compliance rates were generally high, models often produced incorrect numeric values despite adhering to output format constraints.
- **Few-Shot Instability:** Increasing k did not consistently improve performance and in several cases reduced accuracy, suggesting sensitivity to prompt phrasing and dataset structure.

B.3. Implications for Benchmark Design

These observations indicate that GeoQuestions1089 exhibits:

- Sensitivity to phrasing inconsistencies.
- Incomplete contextual specification for numeric queries.
- Instability under few-shot prompting.

These findings directly motivated the construction of **GeoBenchmark**, which introduces:

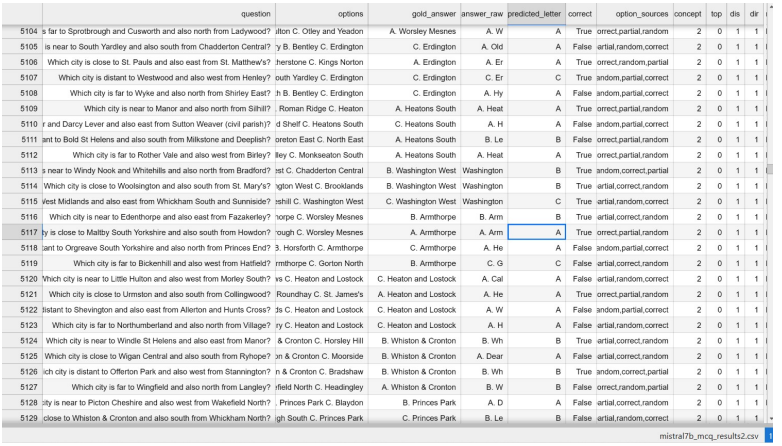
- Explicit relation formalization via structured triplets (S, R, O).
- Deterministic geometry-derived ground-truth labels.
- Balanced structured QA formats (binary and MCQ).
- Standardized prompt templates to reduce ambiguity and improve evaluation stability.

C. Benchmark Data Samples

	question	gold_answer	generated_text	nswer_part	correct	transition	concept	top	dis	dir
1	Is Airedale and Ferry Fryston east of Ancoats and Clayton?	yes	Ferry Fryston east of Ancoats and Clayton? Answer: No.	No.	False	no_change	1	0	0	1
2	Is Airedale and Ferry Fryston west of Ancoats and Clayton?	no	Ferry Fryston west of Ancoats and Clayton? Answer: No.	No.	True	replace_dir_token	1	0	0	1
3	Is Davyhulme East east of Page Moss?	yes	on: Is Davyhulme East east of Page Moss? Answer: Yes.	Yes.	True	no_change	1	0	0	1
4	Is Page Moss east of Davyhulme East?	no	ion: Is Page Moss east of Davyhulme East? Answer: No.	No.	True	flip_place	1	0	0	1
5	Is Dodworth east of Stockbridge?	yes	Question: Is Dodworth east of Stockbridge? Answer: No.	No.	False	no_change	1	0	0	1
6	Is Stockbridge east of Dodworth?	no	Question: Is Stockbridge east of Dodworth? Answer: No.	No.	True	flip_place	1	0	0	1
7	Is Earlestown east of Everton?	yes	ct. Question: Is Earlestown east of Everton? Answer: No.	No.	False	no_change	1	0	0	1
8	Is Everton east of Earlestown?	no	ct. Question: Is Everton east of Earlestown? Answer: No.	No.	True	flip_place	1	0	0	1
9	Is Heckmondwike east of Warley?	yes	Question: Is Heckmondwike east of Warley? Answer: No.	No.	False	no_change	1	0	0	1
10	Is Heckmondwike west of Warley?	no	Question: Is Heckmondwike west of Warley? Answer: No.	No.	True	replace_dir_token	1	0	0	1
11	Is Idle and Thackley east of Atherton?	yes	ition: Is Idle and Thackley east of Atherton? Answer: No.	No.	False	no_change	1	0	0	1
12	Is Idle and Thackley west of Atherton?	no	tion: Is Idle and Thackley west of Atherton? Answer: No.	No.	True	replace_dir_token	1	0	0	1
13	Is Cheetham east of Leasowe and Moreton East?	yes	eetham east of Leasowe and Moreton East? Answer: No.	No.	False	no_change	1	0	0	1
14	Is Cheetham west of Leasowe and Moreton East?	no	etham west of Leasowe and Moreton East? Answer: No.	No.	True	replace_dir_token	1	0	0	1
15	Is Castleford Central and Glasshoughton east of Moss Bank?	yes	al and Glasshoughton east of Moss Bank? Answer: Yes.	Yes.	True	no_change	1	0	0	1
16	Is Moss Bank east of Castleford Central and Glasshoughton?	no	t of Castleford Central and Glasshoughton? Answer: No.	No.	True	flip_place	1	0	0	1
17	Is Newsome east of Allerton and Hunts Cross?	yes	lewsome east of Allerton and Hunts Cross? Answer: No.	No.	False	no_change	1	0	0	1
18	Is Allerton and Hunts Cross east of Newsome?	no	llerton and Hunts Cross east of Newsome? Answer: No.	No.	True	flip_place	1	0	0	1
19	Is Dewsbury South east of Ancoats and Clayton?	yes	wsbury South east of Ancoats and Clayton? Answer: No.	No.	False	no_change	1	0	0	1
20	Is Ancoats and Clayton east of Dewsbury South?	no	oats and Clayton east of Dewsbury South? Answer: No.	No.	True	flip_place	1	0	0	1
21	Is Ardsley and Robin Hood east of Riverside?	yes	Ardsley and Robin Hood east of Riverside? Answer: No.	No.	False	no_change	1	0	0	1
22	Is Riverside east of Ardsley and Robin Hood?	no	Riverside east of Ardsley and Robin Hood? Answer: No.	No.	True	flip_place	1	0	0	1
23	Is Horbury and South Ossett east of Ardwick?	yes	Horbury and South Ossett east of Ardwick? Answer: No.	No.	False	no_change	1	0	0	1
24	Is Horbury and South Ossett west of Ardwick?	no	Horbury and South Ossett west of Ardwick? Answer: No.	No.	True	replace_dir_token	1	0	0	1
25	Is Withington east of Birkdale?	yes	ct. Question: Is Withington east of Birkdale? Answer: No.	No.	False	no_change	1	0	0	1
26	Is Birkdale east of Withington?	no	ct. Question: Is Birkdale east of Withington? Answer: No.	No.	True	flip_place	1	0	0	1

llama3_binary_results11.csv

Figure 5: Binary Inference Sample Data Show 1 Concept.



	question	options	gold_answer	answer_raw	predicted_letter	correct	option_sources	concept	top	dis	dir
5104	s far to Sprotbrough and Cusworth and also north from Ladywood?	ilton C. Otley and Yeadon	A. Worsley Mesnes	A. W	A	True	orrect,partial,random	2	0	1	1
5105	is near to South Yardley and also south from Chadderton Central?	'y B. Bentley C. Erdington	C. Erdington	A. Old	A	False	artial,random,correct	2	0	1	1
5106	Which city is close to St. Pauls and also east from St. Matthew's?	herstone C. Kings Norton	A. Erdington	A. Er	A	True	orrect,random,partial	2	0	1	1
5107	Which city is distant to Westwood and also west from Henley?	outh Yardley C. Erdington	C. Erdington	C. Er	C	True	andom,partial,correct	2	0	1	1
5108	Which city is far to Wyke and also north from Shirley East?	n B. Bentley C. Erdington	C. Erdington	A. Hy	A	False	andom,partial,correct	2	0	1	1
5109	Which city is near to Manor and also north from Silhill?	Roman Ridge C. Heaton	A. Heatons South	A. Heat	A	True	orrect,partial,random	2	0	1	1
5110	r and Darcy Lever and also east from Sutton Weaver (civil parish)?	d Shelf C. Heatons South	C. Heatons South	A. H	A	False	andom,partial,correct	2	0	1	1
5111	ant to Bold St Helens and also south from Milkstone and Deeplish?	oreton East C. North East	A. Heatons South	B. Le	B	False	orrect,partial,random	2	0	1	1
5112	Which city is far to Rother Vale and also west from Birley?	ley C. Monkseaton South	A. Heatons South	A. Heat	A	True	orrect,partial,random	2	0	1	1
5113	s near to Windy Nook and Whitehills and also north from Bradford?	est C. Chadderton Central	B. Washington West	Washington	B	True	andom,correct,partial	2	0	1	1
5114	Which city is close to Woolsington and also south from St. Mary's?	gton West C. Brooklands	B. Washington West	Washington	B	True	artial,correct,random	2	0	1	1
5115	Vest Midlands and also east from Whickham South and Sunniside?	shill C. Washington West	C. Washington West	Washington	C	True	artial,random,correct	2	0	1	1
5116	Which city is near to Edenthorpe and also east from Fazakerley?	horpe C. Worsley Mesnes	B. Armthorpe	B. Arm	B	True	artial,correct,random	2	0	1	1
5117	ly is close to Maltby South Yorkshire and also south from Howdon?	ough C. Worsley Mesnes	A. Armthorpe	A. Arm	A	True	orrect,partial,random	2	0	1	1
5118	ant to Orgreave South Yorkshire and also north from Princes End?	3. Horsforth C. Armthorpe	C. Armthorpe	A. He	A	False	andom,partial,correct	2	0	1	1
5119	Which city is far to Bickenhill and also west from Hatfield?	rmthorpe C. Gorton North	B. Armthorpe	C. G	C	False	artial,correct,random	2	0	1	1
5120	Which city is near to Little Hulton and also west from Morley South?	vs C. Heaton and Lostock	C. Heaton and Lostock	A. Cal	A	False	artial,random,correct	2	0	1	1
5121	Which city is close to Urmston and also south from Collingwood?	Roundhay C. St. James's	A. Heaton and Lostock	A. He	A	True	orrect,partial,random	2	0	1	1
5122	listant to Shevington and also east from Allerton and Hunts Cross?	ts C. Heaton and Lostock	C. Heaton and Lostock	A. W	A	False	andom,partial,correct	2	0	1	1
5123	Which city is far to Northumberland and also north from Village?	ry C. Heaton and Lostock	C. Heaton and Lostock	A. H	A	False	artial,random,correct	2	0	1	1
5124	Which city is near to Windle St Helens and also east from Manor?	& Cronton C. Horsley Hill	B. Whiston & Cronton	B. Wh	B	True	artial,correct,random	2	0	1	1
5125	Which city is close to Wigan Central and also south from Ryhope?	on & Cronton C. Moorside	B. Whiston & Cronton	A. Dear	A	False	artial,correct,random	2	0	1	1
5126	ich city is distant to Offerton Park and also west from Stannington?	n & Cronton C. Bradshaw	B. Whiston & Cronton	B. Wh	B	True	andom,correct,partial	2	0	1	1
5127	Which city is far to Wingfield and also north from Langley?	ifield North C. Headingley	A. Whiston & Cronton	B. W	B	False	orrect,random,partial	2	0	1	1
5128	ity is near to Picton Cheshire and also west from Wakefield North?	Princes Park C. Blaydon	B. Princes Park	A. D	A	False	artial,correct,random	2	0	1	1
5129	close to Whiston & Cronton and also south from Whickham North?	igh South C. Princes Park	C. Princes Park	B. Le	B	False	artial,random,correct	2	0	1	1

mistral7b_mcq_results2.csv

Figure 6: MCQ Inference Sample Data showing with 2 spatial concepts.