

Geo-Spatial Concept Probing of Large Language Models: Abstraction, Compositionality, and Grounding

Karim Radouane
karim.radouane@irit.fr
University of Toulouse, IRIT
Toulouse, France

Jose G Moreno
jose.moreno@irit.fr
University of Toulouse, IRIT
Toulouse, France

Lynda Tamine
lynda.tamine@irit.fr
University of Toulouse, IRIT
Toulouse, France

Abstract

Understanding concepts is fundamental to generalization. Despite their impressive performance on a wide range of tasks, Large Language Models (LLMs) still struggle with genuine concept understanding. Prior work has evaluated conceptual understanding in LLMs using natural-language benchmarks or narrowly scoped synthetic tasks, but these settings often conflate multiple skills or lack precise control over the underlying concepts and their properties. To support controlled probing of concepts in LLMs, we design tests on their core properties: abstraction, compositionality, and groundness. We set up a concept-centric benchmark, targeting spatial concepts such as direction, distance, topology, and their compositions, and use question answering tasks serving as a proxy. We conduct extensive experiments across multiple LLM architectures and training regimes to analyze how model scale and design impact conceptual understanding. The results reveal clear limitations in current LLMs and provide insights into the factors shaping their ability to acquire and compose structured concepts. Our findings shed light on how concept-based LLMs can be redesigned for improved information access and knowledge management. The code will be available at <https://github.com/rd20karim/concept-probing>.

CCS Concepts

• Information systems → Retrieval models and ranking.

Keywords

Large Language Models, Probing, Geography, Concepts.

1 Introduction

Large Language Models (LLMs) exhibit a wide range of capabilities, but the question of whether they can be models of human language understanding is still an open philosophical and scientific debate that impacts cognitive science, machine learning (ML), and natural language processing (NLP), and information research disciplines [13, 35, 52], to cite but a few. One major subject of debate is the source of their capabilities: do they behave as *statistical parrots*, or are they able to organize symbolic, structured representations of *concepts* that they manipulate in the generation process? The question is key since concepts are the cornerstone of intelligence, allowing better generalization, and interpretability [36]. Recent work has developed probing and learning techniques for concepts in LLMs. Probing can focus on factual knowledge through downstream task evaluation [9, 17, 28, 54] or on their internal representations to explain their predictions (e.g., *mechanistic interpretability* [33, 49]). Concept learning refers to the learning of concept representations and their explicit utilization in the internal layers of models, such as *concept bottleneck models* [20].

Given the open-ended nature of concepts across tasks and domains, other works evaluated whether *concept properties* acknowledged in the literature [10, 22, 24, 44] (e.g., compositionality [22, 44]), can be represented or even learned in LLMs. Evaluation of *concept property* provides a more fine-grained insight into the *concept understanding* of LLMs, enabling better control over both concept probing and learning.

Research gap. Recently, there has been growing interest in studying LLMs’ understanding of concepts such as truth [2], time and space [15], patient gender [1], and gender bias [53]. Other works addressed the study of concept properties such as complexity [18], compositionality [22, 44], and grounding [23, 32]. However, while previous work provides insights on the abilities of LLMs to succeed in (cross-modal) tasks through task performance involving concepts of interest, they do not systematically assess concepts per se for the following reasons: (i) they mostly consider multi-modal LLMs (text and image) and concept probing is addressed through representation sharing between modalities (e.g., [23]); (ii) the few works that considered text-only LLMs covered only one core property of concepts (e.g., abstraction [15], grounding [35]) limiting the scope of concept probing; (iii) the design of probes is not guided by the concepts per se, but instead by downstream tasks, thereby confounding conclusions about both LLMs’ concept understanding and task skills (e.g., [41, 51]).

Goal and research questions. In this work, we seek to fill the above research gap by investigating whether LLMs can build, from only the **text modality**, **concept** representations of world models using a **controlled concept-centric benchmark**. Inspired by previous work [24], we approach this challenge by framing it as a set of **probing tests** about **each of the core properties** of conceptual representations: **abstraction**, **compositionality**, and **grounding**. Under this perspective, our main goal can be formalized as:

How can we translate abstraction, compositionality, and grounding into testable hypotheses and develop the empirical probing tests allowing us to quantify the extent to which LLMs align with?

We choose three popular spatial concepts—*direction*, *distance*, and *topology*—which are particularly challenging for dealing with spatial commonsense since they are rarely expressed in texts, though our probing methodology is designed to generalize across spatial concepts and extend to other concepts, tasks, and domains. Specifically, we address the following research questions:

(RQ1) Are LLMs’ performances on a geo-spatial question-answering task sufficient to assess their actual ability to understand underlying concepts?

(RQ2) Do LLMs encode abstract geo-spatial concept representations? Are concept representations generalizable across linguistic tokens and geographic regions?

(RQ3) Are LLMs able to compose geo-spatial concepts? Can LLMs’ predictions be explained by composing concept representations?

(RQ4) Can concept representations be grounded in real-world knowledge?

We build a **concept-centric benchmark** and probe popular text-only LLMs (Llama-8B, Mistral-7B/8B, Qwen-0.6B, 1.7B, 4B, 8B) through binary and multi-choice question-answering (MCQ) tasks to analyze their internal representations across layers and concepts.

Main findings.

- For RQ1, we assess LLM performance on our benchmark as a preliminary proxy for concept understanding. We find moderate MCQ accuracy coexists with low consistency under Binary-QA, exposing unstable concept understanding. This limitation is most analytically tractable for the *distance* concept. **These findings clearly motivate moving beyond geo-spatial QA performance toward targeted probing of core conceptual properties.**
- For RQ2, we measure the extent to which concepts are explicitly encoded in LLMs, and their instances represent types as an abstraction across tokens and regions. Our findings indicate that most models encode concepts with a good level of generalization across instances, except the Mistral family models which consistently underperform across generalization splits and layers.
- For RQ3, compositionality is assessed through: the *compositionality gap*, the additivity of latent representations, and the extent to which predictions can be explained by the composition of latent embeddings and/or atomic predictions. We find that most LLM variants exhibit a compositional structure from a cosine similarity perspective, except for the Mistral models. The correlation analyses using different proposed approaches reveal that compositionality is a key factor for successful concept understanding. In particular, models that preserve correlation across layers perform well, whereas the Mistral family models fail to meet these criteria.
- For RQ4, despite enriching concept-based questions with real-world factual information, such as geometrical measurements, it does not improve related QA-task performance in LLMs even with an explicit distance threshold definition. This limitation highlights a persistent gap in the LLMs models’ ability to ground geo-spatial concepts, revealing a misalignment between representations in linguistic space and their corresponding numerical representations.

2 Related Work

The primary objective of this work is to investigate LLMs’ understanding of concepts through abstraction, compositionality, and grounding. As a use case, we consider geography-related concepts and use performance on question-answering and probing tasks as a proxy. Therefore, we review previous work on (i) concepts in LLMs; (ii) probing LLMs on geography knowledge, and (iii) geo-spatial

question-answering. Table 1 compares prior close work on probing LLMs’ concepts with ours.

2.1 Concepts in LLMs

The definition of *concept*, which has recently gained renewed attention in explainable artificial intelligence (XAI), remains elusive and varies across disciplinary boundaries (e.g., cognitive science, ML [10, 36, 43]). In *cognitive science*, a *concept* is a fundamental unit of structured knowledge representing an interpretable abstraction of entities (e.g., shape, color) that functions like a symbol used for problem-solving, reasoning, or organizing information [10, 14, 50]. In *ML*, a concept is an abstract representation that captures a meaningful class, attribute, or feature within a model’s learned space that correlates with high-level semantic [11, 37, 44]. Across these disciplines, a conceptual representation is generally assumed to require three core properties [10, 22, 24, 44]: (i) *abstraction*¹: different instances of a concept evaluate to the same semantic type. This property is strongly related to the ability of concept representations to tackle out-of-distribution generalization [12, 24]; (ii) *compositionality*: a concept can be *atomic* or composed of a set of concepts. A *composite* concept is a function of its constituents. Previous work addressed the compositionality of LLMs from two sides. *Compositional reasoning* is the ability to process reasoning chains [16, 21, 39, 42] while *conjunctive compositionality* is the ability to disentangle and combine the concept constituents (e.g., via logical AND [31]) to recall or understand the composite concept [29, 31]; and (iii) *grounding*: universally refers to the fact that concepts are assumed to apply to situational context reflecting physical things in the real-world [10, 23, 51]. Grounding tests [3] generally rely on multimodal mapping between concept modalities, where a “modality” could be either a sensor (e.g., text vs. image) or a mode (e.g., plain text and a textual description of a grid [32]).

The question of whether LLMs can give rise to complex conceptual representations has led to a large body of work that falls in the mechanistic interpretability [12, 30, 33]. It has been shown that concepts such as truth [2], time and space [15], patient gender [1], and gender bias [53] can be localized in LLM representations. In their recent work [18], Jian et al. categorized the complexity of concepts based on their level of abstraction and showed that tasks requiring complex concepts require deeper layers for an accurate understanding. Other works studied, particularly, whether concept properties such as compositionality [22, 39, 44] and grounding [23, 32, 51] hold in LLMs through multi-modal downstream tasks (i.e., using texts and images).

In this work, we probe LLMs from a concept-centric perspective that explicitly models concepts and their three core properties—*abstraction*, *grounding*, and *compositionality*—and, unlike prior work, introduce a concept-guided probing benchmark to evaluate LLMs’ internal concept representations and the sensitivity of natural language question-answering performance to the presence of these concepts along with their three core properties.

2.2 Probing LLMs for geo-spatial knowledge

Probing LLMs for geo-spatial knowledge offers a lens into two categories of work. The first category aims to evaluate the capabilities

¹Abstraction is also referred to as *systematicity*. [10]

Related Work	Concepts				Properties			Probing Tests	
	Dir	Dis	Top	Other	Abst.	Comp.	Ground	Knowledge	Repr.
Ramrakhiani et al. [41]	-	-	-	Location	-	-	-	✓	-
Patel and Pavlick [34]	✓	-	-	Color	-	-	✓	-	✓
Manvi et al. [28]	-	-	-	Location	-	-	-	✓	-
Gurnee and Tegmark [15]	✓	-	-	Time	✓	-	-	✓	✓
Park et al. [32]	✓	-	-	-	-	-	✓	-	✓
Yamada et al. [51]	-	-	-	-	-	-	-	✓	-
Our work	✓	✓	✓	-	✓	✓	✓	✓	✓

Table 1: Related work on concept probing in LLMs with a focus on text modality. The table summarizes investigated concepts and properties, and categorizes probing tests by whether they target knowledge or/and internal representations.

of LLMs to recall geographic facts using downstream tasks’ performance as a proxy [17, 28, 47, 54]. For instance, Manvi et al. [28] demonstrated that LLMs are highly sensitive to prompt formats but that fine-tuning using map data from OpenStreetMap enhances the accuracy and robustness of these models. However, it has been shown that fine-tuning LLMs on specific datasets introduces bias against geographic areas with lower socio-economic conditions [27].

The second category of work [15, 34] aims to examine the association between internal representations and extrinsic properties of geo-spatial knowledge. Gurnee and Tegmark [15] showed that LLMs internally store concepts like latitude, longitude, and time in the early layers, and they can predict a location map. The work developed by Patel and Pavlick [34] represents a notable conceptual advancement in understanding the capabilities of LLMs to ground the learned conceptual representation –built upon a linguistic world– with real-world representation based on a few in-context examples. They showed that large-size models exhibit high-level capabilities for grounding and generalization, indicating that they can learn a latent structure of the world.

2.3 Geospatial question-answering

Geospatial question answering (GeoQA) aims to enable generation or retrieval models to answer questions involving geographic entities or concepts and that require spatial operations [26]. Mai et al. [26] report many challenges in GeoQA systems, among which are the *vagueness of geographic concepts* invoked in the questions, the *difficulty in identifying the correct spatial relations between the entities* involved in the questions, and the difficulty in encoding spatial entities. With the advancement of LLMs, a large body of work examined their capabilities to address these challenges, by covering a wide range of GeoQA tasks, including place retrieval [25], and spatial reasoning on topological relations [6–8, 17]. The main limitations of LLMs that arise from this literature are: (i) LLMs are not spatially replicable over the globe, nor efficiently robust across tasks; (ii) LLMs are not efficiently transferable to tasks handling different spatial scales; and (iii) LLMs’ do not generalise the “neighbourhood” relationship.

The challenges mentioned above motivate us to: (i) target spatial concepts as a use case for our probing study, but without loss of generality; and (ii) design a question-answering benchmark test built upon these concepts.

3 Background, Terminology and Notations

3.1 Definitions

Concept. Across disciplines, a *concept* C is a high-level and human-interpretable unit of information. Specifically, we adopt a definition rooted in knowledge representation [5] and widely used in Information Retrieval (IR) and NLP, where a concept is an abstract category or class that defines a set of object entities sharing common properties and roles (e.g., concept of *direction*). Concepts are interpreted as unary *predicates*, and roles as binary *relations* between object entities.

Concept instance. A concept instance c is a concrete exemplar of a concept, representing a specific real-world entity that belongs to the extension of that concept. In IR and NLP, a concept is approximated by a finite vocabulary of representative lexical words c that represent its *instances* (e.g., *east*, *west* are instances of concept *direction*). We note $I(C)$ the finite representative set of instances of concept C . For the sake of simplicity, we consider in practice $C = I(C)$.

Concept instance representation. We consider symbolic concept instance representations using binary relational predicates that express *facts*. A fact is represented as a *positive triplet* $\langle x, r, y \rangle$, where x is the subject entity, r is the core relation (predicate) expressing the concept instance c , and y is the object entity (e.g., $\langle \text{Prescot}, \text{west_of}, \text{Todmorden} \rangle$ represents *west*, an instance of concept *direction* viewed as a the binary relation *west_of* between *Prescot* and *Todmorden* object entities). We also consider the negated triplet $\langle x, \neg r, y \rangle$ which embeds a fact with an opposite relation to r (e.g., $\neg \text{west_of}$ is *east_of*).

Concept properties. Following [10, 24], we consider three core properties of concepts:

- **Abstraction:** a concept C is assumed to be the abstraction of its instances $I(C)$. Therefore, $I(C)$ form the *semantic type* of C .
- **Compositionality:** following compositional distributional semantics models [39, 44, 46], we focus on *compositionality* through conjunctive (i.e., logical AND) concept composition [31]. The composition of *atomic concepts* $C_1, \dots, C_n$ yields a new *composite concept* $C_1 \times \dots \times C_n$, whose instances are formed by composing the instances of the constituent atomic concepts (e.g., $\langle \text{Prescot}, \text{west_of} \wedge \text{far_from}, \text{Todmorden} \rangle$).
- **Grounding** is the awareness of concepts, expressed using natural language constructs, with the physical entities in the world they are assumed to apply to [10, 23, 51]. As done in previous work [34], we only use the *text* modality to represent concepts in two different modes. In our work, we specifically use numerical fact-based geometric measurements and associated question-based formulations.

3.2 Concept probing methodology overview

To probe concept C through the properties of *abstraction*, *compositionality*, and *grounding*, we build a concept-centric question-answering benchmark where questions $q \in Q$ are generated from triplets $\langle x, r, y \rangle$ representing instances of either atomic or composite concepts (§ 4). Using this benchmark, we assess each concept property through specific tests (§ 6) by using

Concept C	Instance (c) / Relation (r_c)	Negation ($\neg r_c$)
Direction	north/north_of south/south_of east/east_of west/west_of	south_of north_of west_of east_of
Distance	far/far_from close/close_to	close_to far_from
Topology	within/is_within borders/is_bordering	is_not_within is_not_bordering

Table 2: Concepts with paired instance–relation entries and their corresponding negated relations.

two proxies:

- *Task performance*: to evaluate whether LLMs succeed on a test, we compute the *accuracy* and *consistency* metrics on the question-answering task. Let Q_t denote the question set. We define per-question *accuracy* as $\text{Acc}(q) = \mathbb{I}[\hat{y}(q) = y(q)]$, where $\hat{y}(q)$ is the LLM’ prediction and $y(q)$ is the ground-truth answer. Accuracy is computed as the average over all questions Q_t as $\text{Acc} = \frac{1}{|Q_t|} \sum_{q \in Q_t} \text{Acc}(q)$. We access *consistency*, regarding a concept property test, as a measure of whether a model correctly answers the paired questions $(q, \bar{q})$ that embed concept instances and their *negation* through opposite relational facts (e.g., if $\langle x, \text{east_of}, y \rangle$ is true, then $\langle x, \text{west_of}, y \rangle$ should be false). Consistency is computed as $\text{Consist.} = \frac{1}{|Q_t|} \sum_{q \in Q} \text{Acc}(q) \cdot \text{Acc}(\bar{q})$, where $\text{Acc}(\bar{q})$ is the per-question accuracy of the negated question.

- *Probing performance*: We complement task performance evaluation with probing task evaluation [4]. The probing tasks rely on linear classifiers to test whether the internal concept instance representations $L(q)$, $q \in Q$ are aligned with the requirements of the property being tested.

4 The Concept-Centric Probing Benchmark

4.1 Benchmarking geographic concepts

Without loss of generality, we investigate in this work three key spatial concepts— *Direction*, *Topology*, and *Distance*. Let $\mathcal{W}$ be a set of geographic entities. We note $\mathcal{M}$ the set of wards that belong to UK metropolitan districts,² $\mathcal{M} \in \mathcal{P}(\mathcal{W})$ where $\mathcal{P}$ denotes the power set. The distance, direction, and topology concepts can be expressed using a wide range of instances. In our work, we consider the following concept instances:

$$\begin{aligned} C_{dir} &= \{\text{north, south, east, west}\}, \\ C_{dis} &= \{\text{close, far}\}, \quad C_{top} = \{\text{within, borders}\}. \end{aligned} \quad (1)$$

Each concept C is represented using a set of instances c as relational triplets $R_C = \{(x, r_c, y) / (x, y) \in \mathcal{W} \times \mathcal{W}, c \in I(C)\}$. For the sake of simplicity, we consider $R_C = \cup_{c \in I(C)} R_c$ (e.g., $R_{dis} = R_{close} \cup R_{far}$). To allow the evaluation of consistency, we also consider negated concept instance representations $\{(x, \neg r_c, y) / (x, y) \in \mathcal{W} \times \mathcal{W}, c \in I(C)\}$. Table 2 presents the concepts, their instances, associated relations, and their negations.

4.2 GeoQA dataset Generation

Concept instance generation. We focus on the UK metropolitan district wards² as location entities, which form three discontinuous regions (upper, middle, and lower). Concept instances in the form of triplets $\langle x, r, y \rangle$ are generated from the middle continuous region $\mathcal{W}$, while the upper region is reserved for out-of-distribution (OOD) generalization. Specifically, using geometric measurements

from GraphDB³ and YAGO2GEO⁴, we compute bearings, pairwise distances, and topological relations between wards associated with *direction*, *distance*, and *topology* concepts, respectively. For any $x, y \in \mathcal{W}$, we denote their distance and bearing by $d(x, y)$ and $\theta(x, y)$, respectively, and map these measurements to relation concept instances $\langle x, r, y \rangle$ via the functions $\phi_{d,th}$ and ϕ_θ , defined as follows:

$$\phi_\theta(x, y) = \begin{cases} E, & \theta \in [45^\circ, 135^\circ], \\ S, & \theta \in [135^\circ, 225^\circ], \\ W, & \theta \in [225^\circ, 315^\circ], \\ N, & \text{otherwise.} \end{cases} \quad \phi_{d,th}(x, y) = \begin{cases} \text{close}, & d(x, y) \leq th, \\ \text{far}, & d(x, y) > th. \end{cases}$$

Figure 1: Direction function ϕ_θ mapping angles to cardinal directions. Distance threshold function $\phi_{d,th}$ mapping distances to ‘close’ or ‘far’.

For example, $\phi_\theta(x, y) = E$ yields the triplet $(x, \text{east_of}, y)$, while $\phi_{d,th}(x, y) = \text{close}$ yields $(x, \text{close_to}, y)$; analogous triplets are generated for all remaining atomic distance and direction relations. We select the distance threshold based on the pairwise distance distribution of geometric wards, setting th as the mean of the distribution, resulting in $th = 47.76$ km. Since our goal is to investigate the internal LLM perception of distance, this threshold is not provided in the question context; its impact is further analyzed in Section 5.2. For topological relations, following YAGO2geo, the relations *is_within* and *is_bordering* between spatial entities are derived by applying GeoSPARQL/OGC topological predicates over their geometries.

The generation of all such atomic spatial triplets is described in Part 1 of Algorithm 1 (lines 4-16). We further construct higher-level concept instances through *conjunctive composition*, by systematically combining atomic relations that share the same subject and object. Given two atomic triplets (x, r_1, y) and (x, r_2, y) , where r_1 and r_2 belong to different relational families (e.g., distance and direction), we generate a composite triplet $(x, r_1 \wedge r_2, y)$ and proceed analogously for three relations. This composition process exhaustively covers all valid pairwise combinations of atomic relations, and is further extended to higher-order compositions (i.e., triplets of relations) whenever applicable. **By explicitly generating atomic, pairwise, and higher-order composed relations, the benchmark enables controlled evaluations of abstraction and compositional generalization.** The compositional generation procedure is detailed in Part 2 of Algorithm 1 (lines 17-22).

Question generation. We generate Binary-QA (Yes/No) and MCQ (3-options) from each concept instance represented as triplets generated by Algorithm 1. Given the resulting set of positive triplets T^+ , we construct negative triplets T^- by negating each relation r (Table 2), yielding the complete set $\mathcal{T} = T^+ \cup T^-$. We then apply a predefined set of templates to each triplet $t \in \mathcal{T}$, illustrated in Table 3. The prompt templates used for binary and MCQ tasks are defined are presented in Figure 2.

Ground-truth answer generation. Ground-truth answers for both, Binary-QA and MCQ tasks, are generated using Algorithm 2.

²Metropolitan boroughs on Wikipedia: en.wikipedia.org/wiki/Metropolitan_borough.

³<https://graphdb.ontotext.com/>

⁴<https://yago2geo.di.uoa.gr/>

Binary-QA Task: Answer the following geography question on UK metropolitan district wards. Respond only with yes or no.
MCQ Task: Answer the following geography question on UK metropolitan district wards. Respond only with the correct option A, B, or C.

Figure 2: Prompt templates for Binary-QA and MCQ tasks.

Algorithm 1 Triplet Generation

```

1: Inputs: wards  $\mathcal{W}$ , metropolitan districts  $\mathcal{M}$ 
2: Output:  $T^+$  (atomic and compositional positive triplets)
3: Initialize  $R_{dis}, R_{dir}, R_{top}, R_{is\_within}, T^+ \leftarrow \emptyset$ 
4: for all  $x, y \in \mathcal{W}$  do ▷ Part 1: Atomic triplet
5:    $d \leftarrow d(x, y), \theta \leftarrow \theta(x, y)$  ▷ compute distance and angle
6:    $r_{dis} \leftarrow \phi_d(x, y), r_{dir} \leftarrow \phi_\theta(x, y)$  ▷ dis. and dir. relation
7:    $R_{dis} \leftarrow R_{dis} \cup \{(x, y)\}; R_{dir} \leftarrow R_{dir} \cup \{(x, y)\}$ 
8:   if border( $x, y$ ) then ▷ Using YAGO2GEO
9:      $r_{top} \leftarrow is\_bordering$ 
10:     $R_{top} \leftarrow R_{top} \cup \{(x, y)\}$ 
11:   end if
12:    $r_{top} \leftarrow is\_within$ 
13:   Find  $z \in \mathcal{M}$ , such that  $(x, r_{top}, z)$  ▷ Using YAGO2GEO
14:    $R_{top} \leftarrow R_{top} \cup \{(x, z)\}$ 
15:    $T^+ \leftarrow T^+ \cup \{(x, r, y) \mid r \in \{r_{dir}, r_{dis}, r_{top}\}\}$ 
16: end for
17: for all  $(x, y) \in T^+$  do ▷ Part 2: Compositional triplet
18:    $S \leftarrow \{r \mid (x, r, y) \in T^+\}$ 
19:   for all subsets  $U \subseteq S$  with  $|U| \geq 2$  do
20:      $r_U \leftarrow \bigwedge_{r \in U} r$  ▷ logical AND (e.g.  $r_1 \wedge r_2$ )
21:      $T^+ \leftarrow T^+ \cup \{(x, r_U, y)\}$ 
22:   end for
23: end for
24: return  $T^+$ 

```

Affirmation	Negation	Composition Example
Is X far from Y?	Is X close to Y?	Is X far from Y and close to Y?
Is X close to Y?	Is X far from Y?	Is X close to Y and far from Y?
Is X west of Y?	Is X east of Y?	Is X west of Y and bordering Y?
Does X border Y?	Is X not bordering Y?	Does X border Y and north of Y?
Is X within Z?	Is X not within Z?	Is X within Z and close to Y?

Table 3: Examples of binary atomic and compositional questions, for MCQ task, we use slightly different template (e.g., *Which ward is close to Y?* followed by options).

MCQ tasks additionally require the generation of incorrect options, which are produced using a distractor generator denoted by D_n :

$$D_n(R, y) = \{d_1, \dots, d_n \mid d_i \sim \{x \in \mathcal{W} \mid (x, y) \notin R\}\}.$$

In our case $n = 2$, the procedure is to choose the target $w^* \in \mathcal{W}$ as the correct answer, sample two distinct distractors $w_1, w_2 \in \mathcal{W} \setminus \{w^*\}$ satisfying $(w_i, y) \notin R$, form the option set $O = \{w^*, w_1, w_2\}$, and place w^* at a uniformly random position $i^* \in \{1, 2, 3\}$.

QA-Dataset statistics. Let $|\cdot|$ denotes the cardinality measure, we have $|\mathcal{W}| = 506$ wards and $|\mathcal{M}| = 25$ metropolitan district. Statistics of our generated Geo-QA dataset, $\mathcal{D}_{QA}$, are presented in Table 4.

Algorithm 2 Ground Truth Generation of Answers (A)

```

1: Inputs:  $k \in \{\text{Binary-QA, MCQ}\}, T^+, \mathcal{W}, D_n$ 
2: Outputs: Answer set A
3: for all  $(x, R, y) \in T^+$  do ▷ Binary task answers
4:   if  $k = \text{Binary-QA}$  then
5:      $A(x, R, y) \leftarrow \{\text{Yes}\}; A(x, \neg R, y) \leftarrow \{\text{No}\}$ 
6:   else if  $k = \text{MCQ}$  then ▷ MCQ task answers
7:      $A(x, R, y) \leftarrow \{x, d_1, d_2\}$  (i.e.  $x \cup D_2(R, y)$ )
8:      $\bar{x} \sim \{z \in \mathcal{W} \mid (z, y) \in \neg R\}$  ▷ Sampling ( $\sim$ )
9:      $A(x, \neg R, y) \leftarrow \{\bar{x}, \bar{d}_1, \bar{d}_2\}$  (i.e.  $x \cup D_2(\neg R, y)$ )
10:     $A(y) \leftarrow \bar{x}$ 
11:   end if
12: end for
13: return A

```

Task(s)	Combinations	Total
Binary-QA	1-Concept: dir, dist, top	769,260
	2-Concept: dir_top, dist_top, dist_dir	771,930
	3-Concept: dist_dir_top	258,200
	Overall Total	1,799,390
MCQ	1-Concept: dir, dist, top	4,096
	2-Concept: dir_top, dist_top, dist_dir	70,100
	3-Concept: dist_dir_top	41,098
	Overall Total	115,294

Table 4: Statistics of the Geo-QA dataset ($\mathcal{D}_{QA}$) per task for Binary-QA and MCQ.

5 Probing LLMs on geo-spatial QA

Our key objective here is to investigate whether geo-QA task performance is a reliable and sufficient proxy to probe LLMs on their understanding of concepts.

5.1 Preliminary evaluation

Before testing LLM’s understanding of concepts on core properties, we first aim to evaluate their performance on the QA dataset of our benchmark, $\mathcal{D}_{QA}$ (§ Table 4). We evaluate popular text-only LLMs (Llama-8B, Mistral-7B/8B, Qwen-0.6B, 1.7B, 4B, 8B) and report the main results in Table 5. We can see that across all settings, MCQ accuracy consistently exceeds Binary-QA accuracy, confirming that structured answer choices substantially ease the task. Larger instruction-tuned models achieve the highest accuracies (e.g., Llama-3.1-8B at 71.7% MCQ and Qwen3-4B at 56.14% Binary-QA). However, consistency remains low overall, especially for Binary-QA (often $< 26\%$), indicating unstable factual recall despite moderate accuracy. While explicit thinking is costly, we can observe that it is largely ineffective: it yields marginal gains at best and often degrades performance (e.g., LLaMA-3.1-8B MCQ accuracy drops from 71.7% to 50.6%, with consistency decreasing from 48.7% to 16.6%). Smaller models sometimes reach comparable or slightly improved accuracy under explicit thinking but still suffer noticeable consistency declines, suggesting correct answers without robust consistency. Consequently, we focus henceforth on models without explicit thinking, relying instead on their inner latent reasoning for our different tests on LLMs’ concept understanding.

Focusing on concept-level analysis, Figure 3 highlights performance across concept dimensions. In particular, we can see that

Model/Task	Binary_QA				MCQ			
	Acc (%)		Consist. (%)		Acc (%)		Consist. (%)	
	✓	×	✓	×	✓	×	✓	×
Llama-3.1-8B-Inst.	61.58	54.95	40.04	25.68	50.60	71.70	16.61	48.72
Mistral-2410-8B-Inst.	61.26	53.39	38.22	22.94	49.10	56.16	23.73	30.96
Mistral-7B-v0.3-Inst.	51.64	51.65	22.55	18.09	51.87	55.31	21.63	31.64
Mistral-7B-v0.2-Inst.	55.42	51.28	27.24	17.10	58.54	52.73	22.40	32.32
Qwen3-8B	62.30	47.99	41.02	10.83	47.22	53.52	15.96	24.25
Qwen3-4B-Inst.	62.73	56.22	39.98	20.81	50.00	56.14	22.67	25.05
Qwen3-1.7B	56.05	49.09	30.74	0.30	46.17	54.29	19.47	28.54
Qwen3-0.6B	53.52	50.25	19.86	0.62	49.06	49.90	21.09	22.02

Table 5: MCQ and Binary-QA performance with shared metrics. ✓ indicates explicit thinking (reasoning-enabled for Qwen models and CoT for the others); ✗ indicates standard prompting.

the LLaMA3-8B model exhibits relatively higher accuracy across most settings compared to other models, reaching above 70% on the *dist_dir_top* composition, with broader coverage across the seven dimensions representing the full concept space. However, consistency remains comparatively lower, reaching around 60% at best on *top* concept (MCQ task). Overall, models tend to perform better under standard prompting than with explicit reasoning, with compositional settings not consistently outperforming atomic ones in the MCQ task.

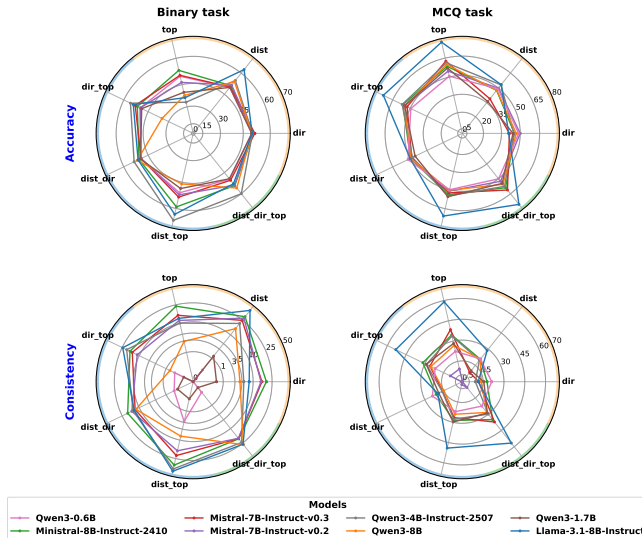


Figure 3: Performance trends per concept type and composition from one to three concepts across Binary-QA and MCQ tasks.

5.2 Distance threshold analysis

In this part, we investigate the internal definition of the distance concept per LLMs, regions and tasks. Binary and MCQ distance labels are derived from a fixed threshold T_0 set to the pairwise distance median. A key question is whether T_0 reflects the models’ *implicit* notion of closeness or imposes an external boundary misaligned with LLM perception. Moreover, to evaluate the effect of **regional scale**, we regenerate the data similarly as UK for the US region. For each region, and task, we estimate each model’s **revealed threshold** T^* as the intersection of the kernel density

estimation (KDE) curves fitted to the close- and far-labelled distance distributions extracted from existing predictions. For binary questions, question polarity combined with Yes/No prediction, determines question label, for MCQ, the distance to the *chosen* ward is assigned with question relation (close/far). Results are summarized in Table 6. We highlight three main findings: **(i) Closeness perception is systematically stricter than the dataset label.** In the binary task, most models yield $T_{UK}^* \ll T_0$, treating a large fraction of *close*-labeled pairs as *far*. This compression is consistent across families, suggesting a corpus-level prior on proximity rather than a model-specific artifact. **(ii) Task format re-calibrates the implicit threshold.** MCQ reverses the binary pattern: all models cluster near $T^*/T_0 \approx 0.92\text{--}0.94$ (UK), closely tracking T_0 . The three named candidate wards act as implicit distance anchors, pulling the model’s boundary toward the dataset scale, regardless of its underlying spatial prior. This re-calibration is format-induced rather than evidence of stronger spatial understanding, and cautions against interpreting MCQ accuracy as a reliable proxy for absolute closeness concept strength. **(iii) Distance concept understanding degrades at continental scale.** Smaller Qwen models (0.6B, 1.7B) show near-chance US binary accuracy ($\approx 50\%$) with negligible Δ , indicating no close/far discrimination at the 1,637 km scale, whereas larger models maintain meaningful gains (Qwen3-4B: Acc +6.1, Cons +5.3; Qwen3-8B: Acc +5.7)—suggesting continental-scale distance concepts are capacity-dependent. The same small models yield non-trivial improvements in UK binary (Qwen3-4B: Acc +3.2, Cons +4.8), confirming the collapse is scale-induced rather than a general model limitation. This degradation manifests differently in MCQ: nearly all models yield *negative* accuracy deltas at T^* across both regions (e.g. Qwen3-1.7B US: -2.1 , Qwen3-4B UK: -0.5), meaning the fixed label T_0 already outperforms the model’s own revealed boundary in the anchored-option setting—further evidence that MCQ performance reflects format calibration rather than intrinsic spatial understanding.

To sum up. QA-based evaluation has inherent limitations as a proxy for concept understanding: surface-level accuracy does not reliably reflect robust understanding across tasks and regions, particularly for challenging concepts such as distance. Consequently, a more targeted probing approach is required to analyze the core properties of concepts in depth, which is the central focus of the following Section 6.

6 Concept Probing

To evaluate concept *abstraction*, *compositionality*, and *grounding*, we construct a dedicated **probing dataset** $\mathcal{D}_{PB}$ by subsampling 1,000 binary and 1,000 ternary representative compositional questions from our Geo-QA corpus $\mathcal{D}_{QA}$ and including all corresponding atomic decompositions and their negations, resulting in $2 \times 7,000 = 14,000$ QA instances. Particularly for the MCQ task, we ensure that each compositional question and its atomics share the same options, this is mandatory for consistent decomposition. Statistics of $\mathcal{D}_{PB}$ are presented in Table 7.

6.1 Testing concept abstraction

To test LLMs on the property of *abstraction*, we answer two core questions tightly related to semantic type representativeness and generalizability: (i) Are concept instances explicitly encoded in the

Model	UK ($T_0 = 47.76$ km)						US ($T_0 = 1636.66$ km)					
	Binary			MCQ			Binary			MCQ		
	T^*	Acc (%)	Cons (%)	T^*	Acc (%)	Cons (%)	T^*	Acc (%)	Cons (%)	T^*	Acc (%)	Cons (%)
Llama-3.1-8B	16.80	70.0 (+7.6)	48.3 (+7.6)	43.22	41.0 (+2.7)	13.4 (+3.0)	1824.39	48.8 (-2.9)	25.4 (-2.9)	1365.79	43.4 (+1.7)	0.0 (+0.0)
Minstral-8B	20.88	60.3 (+9.0)	38.8 (+9.0)	43.99	39.8 (+2.4)	16.0 (+1.6)	1225.86	58.5 (+8.4)	38.0 (+8.4)	1570.75	31.6 (+0.5)	0.0 (+0.0)
Mistral-7B-v0.2	20.78	65.3 (+14.5)	41.5 (+14.5)	56.36	67.6 (+17.0)	22.2 (+11.1)	1034.62	64.2 (+13.4)	47.2 (+13.5)	1495.82	35.9 (-2.8)	14.3 (+0.0)
Mistral-7B-v0.3	19.70	64.4 (+14.0)	38.3 (+14.0)	28.05	44.2 (-1.0)	16.8 (-2.0)	344.80	80.0 (+30.8)	63.6 (+30.8)	1486.11	32.5 (+0.0)	14.3 (+4.8)
Qwen3-0.6B	47.37	50.0 (+0.0)	0.1 (+0.0)	45.09	35.9 (+2.4)	11.9 (+0.2)	1293.36	50.1 (+0.1)	2.0 (+0.1)	1508.52	32.9 (+0.7)	0.0 (+0.0)
Qwen3-1.7B	7.54	50.1 (+0.0)	1.3 (+0.0)	44.43	35.5 (+1.8)	11.9 (+2.0)	1724.44	50.3 (-0.4)	10.7 (-0.4)	1604.81	24.7 (-0.3)	0.0 (+0.0)
Qwen3-4B	30.86	55.9 (+0.9)	21.6 (+0.9)	43.65	35.6 (+2.7)	14.3 (+1.2)	1055.08	60.0 (+9.5)	32.1 (+9.5)	1503.55	36.9 (+2.8)	15.6 (+0.0)
Qwen3-8B	31.74	57.1 (+2.9)	18.1 (+2.9)	44.89	36.8 (+1.9)	11.6 (+0.6)	1582.63	52.8 (+0.9)	28.0 (+0.9)	1453.31	37.5 (+1.1)	0.0 (+0.0)

Table 6: Revealed threshold T^* (km) and performance relative to the GT label T_0 . Acc (%) and Cons (%) are evaluated at T^* ; deltas in parentheses show gain over evaluation at T_0 (+ = improvement, - = degradation).

Probing Data ($\mathcal{D}_{PB}$): Concept types						
dir	dist	dist_dir_top	top	dir_top	dist_top	dist_dir
1541	1523	1000	844	374	354	272

Table 7: Positive counts for each combination in the probing data with uniform sampling per number of concepts.

hidden layer and reliably classified as concept types? (ii) Do learned concept representations generalize to unseen instances or novel combined instances?

Experimental design. To investigate whether frozen LLMs encode geographical concepts at different depths, we employ a layer-wise linear probing approach. For each layer l of a pre-trained model, we extract the average token embedding for a question q , producing a layer-specific representation $h = f_l(q)$. A probe classifier trained on h predicts one of seven conceptual classes C (e.g., directional, topological, distance, or their compositions). Specifically, for each q , the probe outputs $\hat{y} = \text{softmax}(W_l h + b_l)$ where $c_q \in C$ is the predicted label and W_l, b_l are layer-specific parameters.

To evaluate the generalizability of LLM representations, we construct several data splits: a standard random train/test split to probe concept encoding (question (i)), and two OOD settings, to probe generalization and *systematicity* [10] (question (ii)): the region-based OOD uses QA pairs from the *upper-region*, while the token-based OOD withholds specific concept combinations. All splits are summarized in Table 8.

Results. Table 9 and Figure 4 show that most evaluated LLMs achieve near-perfect performance on the standard random split (99.95–99.98% accuracy) and generalize well to OOD settings, reaching 75–83% on the Geo-Level split, 80–83% on the single Token-level split, and over 99% on the compositional Token-level split. In contrast, the Mistral-family models exhibit substantially lower accuracy across all splits – random split 52.8–58.0%, Geo-Level split 37.8–42.2%, single Token-level split 22.5–29.4%, and compositional Token-level split 48.5–55.4% – indicating significantly weaker concept encoding and generalization under comparable experimental settings.

To sum up: Most models clearly recognize concepts via their latent representations, which is a consistent finding with previous work that has shown the ability of LLMs to represent the space through the geographic entity location task [15]. However, the Mistral family consistently underperforms across all splits and layers, suggesting limited concept abstraction capacity in its architecture compared to the other tested LLMs.

Split Type	Train	Test
<i>Token-Level Split (Atomic-Based)</i>		
Direction	{N, S}	{E, W}
Distance	{C}	{F}
Topology	{Wi}	{B}
<i>Combination-Based</i>		
Atomics	50% per token	50% per token
Dis×Dir	{C, F}×{N, S}	{C, F}×{E, W}
Dis×Top	{C×Wi, F×B}	{C×B, F×Wi}
Dir×Top	{N×Wi, S×B}	{E×Wi, W×B}
Dis×Dir×Top	{C, F}×{N, S}×{Wi, B}	{C, F}×{E, W}×{Wi, B}
<i>Geo-Level Split²</i>		
Geo-OOD	{mid-region}	{upper-region}

Table 8: Atomic-based token-level and combination-based compositional OOD splits, followed by geographic OOD evaluation. Token abbreviations: N=North, S=South, E=East, W=West, C=Close, F=Far, Wi=Within, B=Borders.

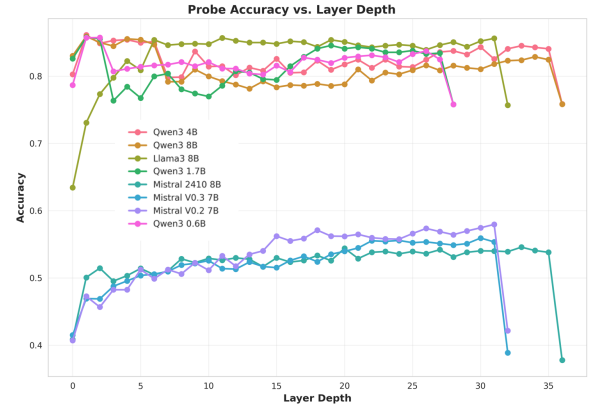
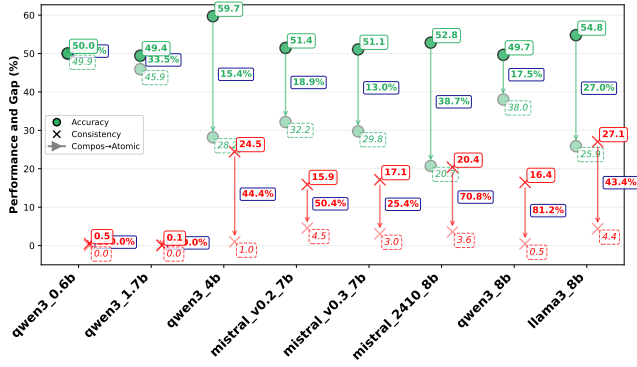


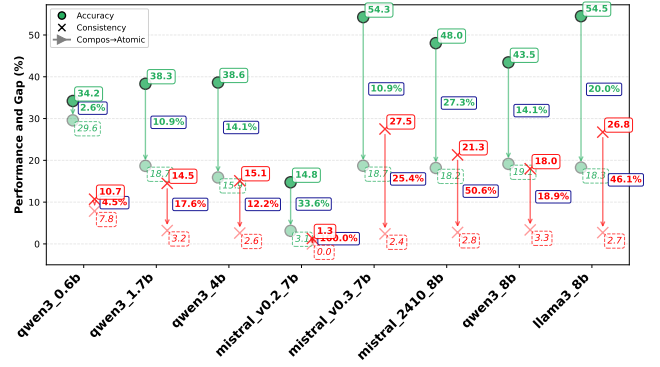
Figure 4: Binary Task: Concepts classification accuracy per LLM layer (depth) for Geo-Level split. The other splits shows similar trends for both tasks.

6.2 Testing concept compositionality

To test the compositionality of LLMs, we proceed in two stages: first, we measure the compositionality gap (§6.2.1), then we test the compositionality of the representations and the predictions (§6.2.2).



(a) Binary-QA Task.



(b) MCQ Task.

Figure 5: Compositional gap analysis (in blue rectangle), highlighting differences between composite and atomic question performance w.r.t accuracy of compositional questions and their atomics. The arrow points from compositional to atomics, illustrating the gap.

Model	L	Token-Level Test			Geo-Level Test
		Random	Single-Based	Combination-Based	Geo
qwen3_4b	36	99.98	81.65	99.45	75.86
qwen3_8b	36	99.98	80.76	99.61	75.83
llama3_8b	32	99.95	82.85	99.90	75.69
qwen3_1.7b	28	99.98	80.43	99.10	75.81
mistral_2410_8b	36	53.05	22.47	48.47	37.76
mistral_v0.3_7b	32	52.83	24.17	49.15	38.88
mistral_v0.2_7b	32	58.02	29.40	55.37	42.17
qwen3_0.6b	28	99.95	80.87	99.33	75.81

Table 9: Final-layer probe results across evaluation settings. Single-Based and Combination-Based correspond to atomic and compositional token-level OOD tests, respectively. Geo-Level Test evaluates geographic generalization.

6.2.1 Measuring the compositionality gap. Instead of following [39] that measures the cases where a model correctly answers all constituent subquestions but fails on the corresponding compositional question to test multi-hop or relational chaining as *compositionality gap*, we probe LLMs through QA tasks with conjunctive concept compositions. For each compositional question $q \in Q$, let $S(q)$ denote its set of sub-questions. Let $Q_t^c = \{q \in Q_t \mid \text{SubAcc}(q) = 1\}$ be the set of compositional questions for a given task t that has its subquestions correctly answered, where $\text{SubAcc}(q) = \mathbb{I}[\forall s \in S(q) : \hat{y}(s) = y(s)]$.

Evaluation metrics. To measure the compositionality gap, we define the following metrics:

- *Per-question CGA.* We define the per-question compositional accuracy as:

$$\text{CGA}(q) = 1 - \text{Acc}(q) = 1 - \mathbb{I}[\hat{y}(q) = y(q)] \quad \forall q \in Q_t^c$$

where $\bar{q}$ denote the negative version of question q , $S(q)$ denotes the set of subquestions of q , $\hat{y}(q)$ is the model prediction, and $y(q)$ is the ground truth answer.

- *Compositional Gap Accuracy (CGA).* The proportion of compositional questions answered incorrectly for which all corresponding subquestions were answered correctly:

$$\text{CGA} = \frac{1}{|Q_t^c|} \sum_{q \in Q_t^c} \text{CGA}(q)$$

- *Compositional Gap Consistency (CGC).* We define a consistency-based compositional gap using paired positive (q) and negative versions ($\bar{q}$) of each question, which decouples true performance from random chance. CGC measures the proportion of compositional questions answered incorrectly for which all corresponding sub-questions were answered consistently correct:

$$\text{CGC} = \frac{1}{|Q_t^c|} \sum_{q \in Q_t^c} \text{CGA}(q) \cdot \text{CGA}(\bar{q})$$

Results. The main results are reported in Figure 5. Unlike [39], in conjunctive-based compositionality, LLMs perform better on compositional questions than on atomic ones. Notably, for the Binary-QA task, *Mistral-v0.3-7B* shows the smallest compositional gap in accuracy (13%), suggesting relatively better compositional generalization. However, this does not translate into strong consistency: its rate is 17.1%, slightly above *Mistral-v0.2-7B* (15.9%), while both remain far from ideal. This advantage in gap size should be interpreted alongside absolute performance, which is higher for *Qwen3-4B* (59.7%). Unsurprisingly, smaller models show the largest compositional and consistency gaps. For the MCQ task, the *llama3-8B* model has the highest compositional accuracy (54.5%) but with a higher compositional gap compared to small-size models. Interestingly, smaller models on this task show lower gaps than in the binary task, yet overall compositional accuracy remains very low ($< 39\%$).

6.2.2 Measuring compositionality. Here, we attempt to answer two core questions: (i) Are concept representations compositional? (ii) Can LLMs’ predictions be explained through compositional representations?

Experimental design. To investigate whether internal representations learned by selected LLMs exhibit compositional structure (i.e., question (i)), we follow previous work [44], by testing compositionality as approximate additivity. We analyze cosine similarity scores between $L(q)$ and $L_{\text{add}} = L(q_1) + L(q_2)$, enabling systematic comparison of compositional behavior across models. We then

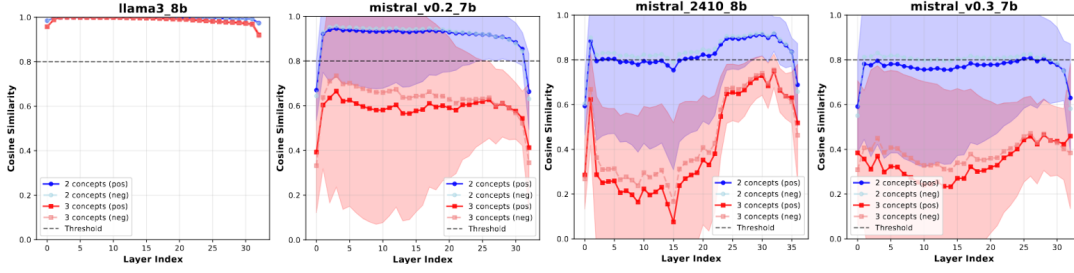
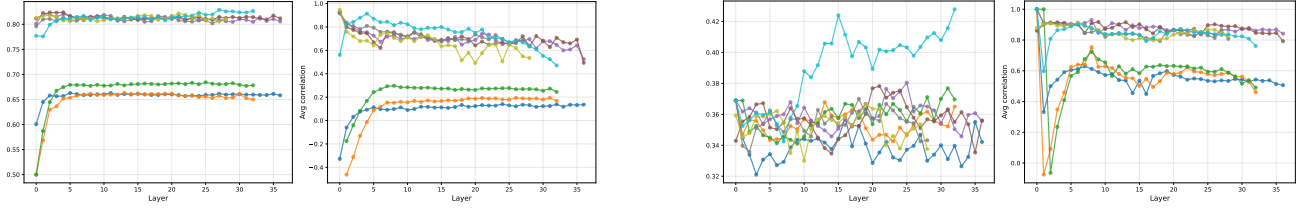


Figure 6: Binary-QA Task: Cosine similarity per layer between each compositional question embedding q and its decomposition, considering 2-Concepts ($q_1 + q_2$) and 3-Concepts ($q_1 + q_2 + q_3$). For MCQ task, we observe similar curves trend. Qwen models have very similar trend as llama3_8b for both tasks.



(a) Binary Task. Mistral models lose logits correlation across layers, directly explaining performance gap compared to the other LLMs.

(b) MCQ Task. Mistral models again exhibit weaker compositionality between atomics and composite questions compared to other models..

Figure 7: Logits-based Pearson correlation coefficient across LLM layers for both tasks.

compute a cosine similarity score $s \in [-1, 1]$ as:

$$s = \text{sim}(\mathbf{L}(q), \mathbf{L}_{\text{add}}) = \frac{\mathbf{L}(q)^\top \mathbf{L}_{\text{add}}}{\|\mathbf{L}(q)\| \|\mathbf{L}_{\text{add}}\|}.$$

The score s measures how well the additive approximation by atomics matches the model’s composite representation.

To test prediction compositionality (i.e., question (ii)), we train multinomial logistic regression probes on frozen LLM layer representations rather than relying on the model’s native logits of the final layer only. This layer-wise probing framework enables analysis across all layers, revealing where compositional structure is encoded. By decoupling evaluation from LLMs’ output-formatting artifacts of auto-regressive generation and focusing on fixed embeddings, the approach isolates representational content and yields a more robust assessment of how internal model features support compositional predictions. Probes are trained on a subset of factual items and evaluated on held-out facts, assessing not fact inference but the consistency of factual and compositional encoding within the model’s representational space and whether it emerges across LLMs from its frozen representation, and explain compositional prediction by its atomic predictions.

To test the level of prediction compositionality, we measure correlations between predictions from the full composite representation $\mathbf{L}(q)$ and the combination of atomic representations $\mathbf{L}_{\text{add}}$. This is performed using logits and embeddings, and probability-based methods. Let $\mathbf{v}_{\text{comp}}$ denote the embedding of a composite question, $\mathbf{v}_i$ the embeddings of its n atomic sub-questions, f the trained classifier, and $\mathcal{L}(f(\mathbf{v}))$ the logit for the true answer class:

- **Logit Additivity:** Correlation between $\mathcal{L}(f(\mathbf{v}_{\text{comp}}))$ and $\sum_{i=1}^n \mathcal{L}(f(\mathbf{v}_i))$, testing whether logit values compose linearly.
- **Embedding Summation:** Correlation between $P(f(\mathbf{v}_{\text{comp}}))$ and $P(f(\sum_{i=1}^n \mathbf{v}_i))$, evaluating whether summed embeddings predict the same class probabilities as composite embeddings.
- **Probability Averaging:** Correlation between $P(f(\mathbf{v}_{\text{comp}}))$ and $\frac{1}{n} \sum_{i=1}^n P(f(\mathbf{v}_i))$, testing whether individual atomic predictions aggregate to composite predictions.

Each metric probes a distinct compositionality hypothesis: decision-boundary linearity, embedding-space compositionality, and compositional confidence.

Results. From an **embedding-based perspective**, Figure 6 reports cosine similarity between compositional questions and their sub-questions across LLM layers. In both Binary-QA and MCQ tasks, Mistral models exhibit the lowest and most variable compositionality correlation. In both settings, 2-Concept compositions consistently show higher compositionality than 3-Concept ones. From a **prediction-based perspective**, Mistral achieves moderate Binary-QA accuracy ($\sim 65\%$) with weak compositional correlations (logit correlations ~ 0.4), and lower MCQ accuracy ($\sim 35\%$). In contrast, Qwen and LLaMA attain higher Binary-QA accuracy ($\sim 80\%$) and comparable MCQ performance ($\sim 35\text{--}39\%$), while maintaining consistently strong correlations across settings (logits ≥ 0.85 , probabilities ≥ 0.7), indicating a substantially stronger alignment between compositional structure and predictive behavior (cf. Table 10).

To sum up. Across experiments, LLMs exhibit a compositional consistency gap in QA tasks. In contrast, probing latent representations across layers reveals that predictive compositionality has a clear impact on performance: models with compositional embeddings

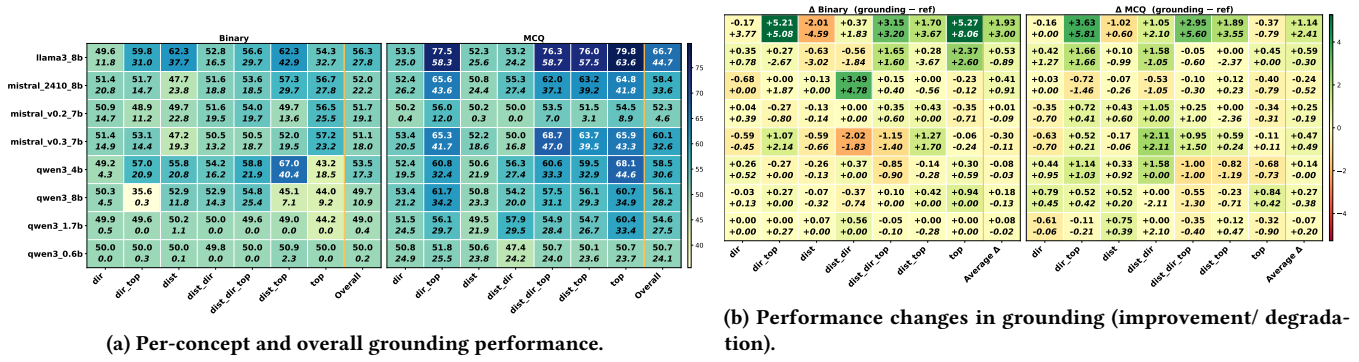


Figure 8: Effect of grounding information on model performance compared to settings without grounding information per concept type. Consistency is reported below accuracy in each cell of the heatmaps, the final column shows the overall and average results.

	Mistral						Qwen						Llama	
Method	7b_v0.2	7b_v0.3	8b_2410	0.6b	1.7b	4b	8b	37	33	35.66	81.36	35.66	81.32	39.01
# layers	33	33	37	29	29	37	37	37	33	33	33	33	33	33
acc (%)	67.03	55.84	64.88	35.39	65.75	33.89	80.89	35.48	81.27	35.22	81.18	36.02	81.36	35.66
logits	0.53	0.60	0.37	0.55	0.33	0.57	0.98	0.86	0.99	0.86	0.99	0.88	0.99	0.85
embed.	0.23	0.51	0.05	0.44	0.08	0.45	0.80	0.83	0.73	0.81	0.76	0.86	0.78	0.84
proba.	0.24	0.59	0.12	0.55	0.10	0.56	0.73	0.85	0.66	0.85	0.70	0.88	0.74	0.84

Table 10: Average correlations and accuracies for Binary/MCQ.

preserve prediction consistency and achieve more robust results, unlike Mistral-like architectures.

6.3 Testing concept grounding

We test *contextual grounding* via QA performance on tasks defined in a numerically explicit spatial context using coordinates, distances, and bearing angles, where all relevant real-world information is provided in context to LLMs as follows:

Context_Binary: X at (a_X, b_X) , Y at (a_Y, b_Y) , distance $d(X, Y) = N$ km, bearing $\beta_{Y \rightarrow X} = \theta^\circ$, X and Y are considered close if $d(X, Y) \leq T_0$, and far otherwise.

Context_MCQ: Y at (a_Y, b_Y) , $\{X_i \text{ at } (a_i, b_i), d(Y, X_i) = N_i \text{ km}, \beta_{Y \rightarrow X_i} = \theta_i^\circ\}_{i=1}^3$ (X_i ward option i), X and Y are considered close if $d(X, Y) \leq T_0$, and far otherwise.

Grounding. In our context is defined by the ability to map linguistic concepts to their numerical meaning. If an LLM fully achieves grounding, it is expected to successfully answer the corresponding questions. In practice, answer quality provides a continuous measure of grounding. Direction- and distance-based concepts can be precisely tested in coordinate-based contexts, where all quantitative information is explicit. In contrast, topology-based concepts require a larger number of points, often leading to noisy contextual expansions and making practical evaluation challenging.

Experimental design. In this test, the LLM context is augmented with relevant geographical measurements and then asked to answer a corresponding related question. This protocol aims to evaluate LLMs' ability to ground concepts appearing in a given question to equivalent numerical representations.

Results. As shown in Figure 8, topology-independent spatial grounding performance in LLMs varies substantially across both task formulations and model architectures. In the MCQ setting, several models surpass the random baseline (33%), with Qwen3-8B achieving the strongest results (66.7% accuracy and 44.7% consistency on the distance concept). However, overall consistency

remains limited, indicating only weak MCQ grounding ability. In the Binary-QA setting, accuracy largely concentrates around chance level ($\sim 50\%$) and is consistently paired with low consistency, revealing significant instability in concept grounding (e.g., Mistral-8B at $\approx 52\%$). Across geographical concept-mapping tasks, LLMs demonstrate weak and inconsistent grounding, with no systematic advantage for direction- or distance-based concepts over topological ones. In several cases, performance degrades below that of simple factual recall, with only marginal improvement in some instances, achieving at best an average improvement of 1.93% accuracy and 3% consistency (Fig. 8b).

To sum up. Concepts are not naturally grounded in the evaluated LLMs; instead, the results indicate a reliance on in-context concept memorization rather than true numerical meaning grounding. This aligns with prior findings showing that text-only small LMs struggle to map linguistic concepts to non-linguistic referents [34].

7 Conclusion

We investigated geo-spatial concept understanding in LLMs through the lens of abstraction, grounding, and compositionality. Our results show that while LLMs encode and partially compose concepts, factual consistency and real-world grounding remain critical bottlenecks. The core of our probing concept-centric QA benchmark (i.e., questions and ground truth) is based on a task-agnostic generation and filtering of relational facts in the form of triplets (§ Algorithm 1, 2). Such a widely adopted knowledge representation form inherently eases the extendability and reuse of our probing methodology to the study of abstraction, compositionality, and grounding of other concepts (e.g., truth [2]) even in other domains (e.g., gender bias in healthcare [1]).

However, while providing valuable insights, our study has some limitations. First, our experiments rely only on two real-world geographic regions. Second, the studied concepts may not fully capture the complexity of other real-world concepts. Third, our probing experiments are limited to linear classifiers.

Our work has potential implications for information and knowledge management, including the following: (i) **Information access and retrieval:** our findings reveal that LLMs generally have a good level of performance in recognizing OOD concept instances, indicating a significant ability of abstraction. Based

on this finding, new paths moving from the current designs of LLMs for relevance ranking [19, 38, 55] toward new approaches of axiomatic IR [48] or mechanistic interpretability for IR [33] are worth investigation by testing the concept of *relevance* through its core properties that constrain words, documents, and queries; (ii) **Mining multimodal content:** the limited grounding of concepts in current LLMs provides concrete evidence of why multi-modal (e.g. language and vision) models must go beyond traditional two-tower models [40]. Our findings argue for externally grounded models that complement LLMs with parametric mechanisms for explicit grounding in the same line as parametric retrieval augmented generation [45]; (iii) **Evaluation:** our findings reveal a critical limited ability of downstream task evaluation to actually probe LLMs on their understanding of conceptual knowledge. This result calls for the design of new benchmarks suited to concept probing, annotated with core properties, standardized metrics, and reference concepts, configurations with published baselines.

References

- [1] Hiba Ahsan, Arnab Sen Sharma, Silvio Amir, David Bau, and Byron C Wallace. 2025. Elucidating Mechanisms of Demographic Bias in LLMs for Healthcare. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 14614–14631. <https://doi.org/10.18653/v1/2025.findings-emnlp.789>
- [2] Amos Azaria and Tom Mitchell. 2023. The Internal State of an LLM Knows When It’s Lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 967–976. <https://doi.org/10.18653/v1/2023.findings-emnlp.68>
- [3] Lisa Beinborn, Teresa Botschen, and Iryna Gurevych. 2018. Multimodal Grounding for Language Processing. In *Proceedings of the 27th International Conference on Computational Linguistics*, Emily M. Bender, Leon Derczynski, and Pierre Isabelle (Eds.). Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2325–2339. <https://aclanthology.org/C18-1197/>
- [4] Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics* 48, 1 (March 2022), 207–219. https://doi.org/10.1162/coli_a_00422
- [5] Ronald J. Brachman. 1979. Taxonomy, Descriptions, and Individuals in Natural Language Understanding. In *17th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, La Jolla, California, USA, 33–37. <https://doi.org/10.3115/982163.982174>
- [6] Anthony G Cohn and Robert E Blackwell. 2025. Evaluating the Ability of Large Language Models to Reason about Cardinal Directions, Revisited. *arXiv preprint arXiv:2507.12059* (2025).
- [7] Cheng Deng, Tianhang Zhang, Zhongmou He, Yi Xu, Qiyuan Chen, Yuanyuan Shi, Luoyi Fu, Weinan Zhang, Xinbing Wang, Chenghu Zhou, Zhouhan Lin, and Junxian He. 2024. K2: A Foundation Language Model for Geoscience Knowledge Understanding and Utilization. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM ’24)*, 161–170. <https://doi.org/10.1145/3616855.3635772>
- [8] Mahir Labib Dihan, MD Tanvir Hassan, MD TANVIR PARVEZ, Md Hasebul Hasan, Md Almash Alam, Muhammad Aamir Cheema, Mohammed Eunus Ali, and Md Rizwan Parvez. 2025. MapEval: A Map-Based Evaluation of Geo-Spatial Reasoning in Foundation Models. In *Forty-second International Conference on Machine Learning*. <https://openreview.net/forum?id=hS2Ed5XYRq>
- [9] Alexandru Dumitru, Venkatesh V. Adam Jatowt, and Avishek Anand. 2025. Evaluating List Construction and Temporal Understanding capabilities of Large Language Models. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval, ICTIR 2025, Padua, Italy, 18 July 2025*, Hamed Zamani, Laura Dietz, Benjamin Piwowarski, and Sebastian Bruch (Eds.). ACM, 369–379. <https://doi.org/10.1145/3731120.3744606>
- [10] Jerry A. Fodor and Zenon W. Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. , 3–71 pages. [https://doi.org/10.1016/0010-0277\(88\)90031-5](https://doi.org/10.1016/0010-0277(88)90031-5)
- [11] Ruth Fong and Andrea Vedaldi. 2018. Net2Vec: Quantifying and Explaining How Concepts are Encoded by Filters in Deep Neural Networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 8730–8738. <https://api.semanticscholar.org/CorpusID:2738204>
- [12] Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer Feed-Forward Layers Build Predictions by Promoting Concepts in the Vocabulary Space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 30–45. <https://doi.org/10.18653/v1/2022.emnlp-main.3>
- [13] Mariel K. Goddu, Alva Noë, and Evan Thompson. 2024. LLMs don’t know anything: reply to Yildirim and Paul. *Trends in Cognitive Sciences* 28, 11 (2024), 963–964.
- [14] Joseph Goguen. 2005. What Is a Concept?. In *Conceptual Structures: Common Semantics for Sharing Knowledge*, Frithjof Dau, Marie-Laure Mugnier, and Gerd Stumme (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 52–77.
- [15] Wes Gurnee and Max Tegmark. 2024. Language Models Represent Space and Time. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=jE8xbmvFin>
- [16] Jianchao Ji, Zelong Li, Shuyuan Xu, Max Xiong, Juntao Tan, Yingqiang Ge, Hao Wang, and Yongfeng Zhang. 2023. Counterfactual Collaborative Reasoning. In *Proceedings of the 16th ACM International Conference on Web Search and Data Mining (WSDM ’23)*, 249–257. <https://doi.org/10.1145/3539597.3570464>
- [17] Yuhai Ji, Song Gao, Ying Nie, Ivan Majić, and Krzysztof Janowicz. 2025. Foundation models for geospatial reasoning: assessing the capabilities of large language models in understanding geometries and topological spatial relations. *International Journal of Geographical Information Science* (2025), 1–38.
- [18] Mingyu Jin, Qinkai Yu, Jingyuan Huang, Qingcheng Zeng, Zhenting Wang, Wenyue Hua, Haiyan Zhao, Kai Mei, Yanda Meng, Kaize Ding, Fan Yang, Mengnan Du, and Yongfeng Zhang. 2025. Exploring Concept Depth: How Large Language Models Acquire Knowledge and Concept at Different Layers?. In *Proceedings of the 31st International Conference on Computational Linguistics*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 558–573. <https://aclanthology.org/2025.coling-main.37/>
- [19] Ekaterina Khramtsova, Shengyao Zhuang, Mahsa Baktashmotlagh, and Guido Zuccon. 2024. Leveraging llms for unsupervised dense retriever ranking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1307–1317.
- [20] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *International conference on machine learning*. PMLR, 5338–5348.
- [21] 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)*, 2873–2882. Proc. of ML Research Vol. 80.
- [22] Martha Lewis, Nihal Nayak, Peilin Yu, Jack Merullo, Qinan Yu, Stephen Bach, and Ellie Pavlick. 2024. Does CLIP Bind Concepts? Probing Compositionality in Large Image Models. In *Findings of the Association for Computational Linguistics: EACL 2024*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian’s, Malta, 1487–1500. <https://aclanthology.org/2024.findings-eacl.101/>
- [23] Lei Li, Jingjing Xu, Qingxiu Dong, Ce Zheng, Xu Sun, Lingpeng Kong, and Qi Liu. 2023. Can Language Models Understand Physical Concepts?. In *The 2023 Conference on Empirical Methods in Natural Language Processing*. <https://openreview.net/forum?id=HaSS8a3Oe7>
- [24] Charles Lovering and Ellie Pavlick. 2022. Unit Testing for Concepts in Neural Networks. *Transactions of the Association for Computational Linguistics* 10 (2022), 1193–1208. https://doi.org/10.1162/tacl_a_00514
- [25] Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, Chris Cundy, Ziyuan Li, Rui Zhu, and Ni Lao. 2024. On the Opportunities and Challenges of Foundation Models for GeoAI (Vision Paper). *ACM Trans. Spatial Algorithms Syst.* 10, 2, Article 11 (July 2024), 46 pages. <https://doi.org/10.1145/3653070>
- [26] Gengchen Mai, Krzysztof Janowicz, Rui Zhu, Ling Cai, and Ni Lao. 2021. Geographic Question Answering: Challenges, Uniqueness, Classification, and Future Directions. In *AGILE Conf. 8*. <https://doi.org/10.5194/agile-giss-2-8-2021>
- [27] Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024. Large language models are geographically biased. In *Proceedings of the 41st International Conference on Machine Learning (Vienna, Austria) (ICML ’24)*. JMLR.org, Article 1409, 16 pages.
- [28] Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David B. Lobell, and Stefano Ermon. 2024. GeoLLM: Extracting Geospatial Knowledge from Large Language Models. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=TqLx2BwXP3>
- [29] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*. 3111–3119. NeurIPS 2013.
- [30] Jatin Nainani. 2024. Evaluating Brain-Inspired Modular Training in Automated Circuit Discovery for Mechanistic Interpretability. *ArXiv abs/2401.03646* (2024). <https://api.semanticscholar.org/CorpusID:266843932>

- [31] Masahiro Naito, Sho Yokoi, Geewook Kim, and Hidetoshi Shimodaira. 2021. Revisiting Additive Compositionality: AND, OR and NOT Operations with Word Embeddings. *Proceedings of the ACL-IJCNLP 2021 Student Research Workshop* (2021).
- [32] Core Francisco Park, Andrew Lee, Ekdeep Singh Lubana, Yongyi Yang, Maya Okawa, Kento Nishi, Martin Wattenberg, and Hidenori Tanaka. 2025. ICLR: In-Context Learning of Representations. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=pXlmOmlHJZ>
- [33] Andrew Parry, Catherine Chen, Carsten Eickhoff, and Sean MacAvaney. 2025. MechIR: A Mechanistic Interpretability Framework for Information Retrieval. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part V* (Lucca, Italy). Springer-Verlag, Berlin, Heidelberg, 89–95. https://doi.org/10.1007/978-3-031-88720-8_16
- [34] Roma Patel and Ellie Pavlick. 2022. Mapping Language Models to Grounded Conceptual Spaces. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=gJcEM8sxHK>
- [35] Ellie Pavlick. 2023. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 381, 2251 (2023), 20220041.
- [36] Eleonora Poeta, Gabriele Ciravegna, Eliana Pastor, Tania Cerquitelli, and Elena Baralis. 2025. Concept-based Explainable Artificial Intelligence: A Survey. *ACM Comput. Surv.* (Nov. 2025). <https://doi.org/10.1145/3774643> Just Accepted.
- [37] Sayantan Polley, Subhajit Mondal, Venkata Srinath Mannam, Kushagra Kumar, Subhankar Patra, and Andreas Nürnberger. 2022. X-Vision: Explainable Image Retrieval by Re-Ranking in Semantic Space. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM '22)*. 4955–4959. <https://doi.org/10.1145/3511808.3557187>
- [38] Ronak Pradeep, Sahel Sharifmoghadam, and Jimmy Lin. 2023. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *arXiv preprint arXiv:2309.15088* (2023).
- [39] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and Narrowing the Compositionality Gap in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6–10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 5687–5711. <https://doi.org/10.18653/V1/2023.FINDINGS-EMNLP.378>
- [40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18–24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. <http://proceedings.mlr.press/v139/radford21a.html>
- [41] Nitin Ramrakhiani, Vasudeva Varma, Girish Keshav Palshikar, and Sachin Pawar. 2025. Gauging, enriching and applying geography knowledge in Pre-trained Language Models. *Inf. Process. Manage.* 62, 1 (Jan. 2025), 23 pages. <https://doi.org/10.1016/j.ipm.2024.103892>
- [42] Kenneth H. Rosen. 2011. *Discrete Mathematics and Its Applications* (7 ed.). McGraw–Hill Education.
- [43] Gesina Schwalbe. 2022. Concept embedding analysis: A review. *arXiv preprint arXiv:2203.13909* (2022).
- [44] Adam Stein, Aaditya Naik, Yinjun Wu, Mayur Naik, and Eric Wong. 2024. Towards compositionality in concept learning. In *Proceedings of the 41st International Conference on Machine Learning (Vienna, Austria) (ICML '24)*. JMLR.org, Article 1893, 26 pages.
- [45] Weihang Su, Yichen Tang, Qingyao Ai, Junxi Yan, Changyue Wang, Hongning Wang, Ziyi Ye, Yujia Zhou, and Yiqun Liu. 2025. Parametric Retrieval Augmented Generation. Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3726302.3729957>
- [46] Matthew Trager, Pramuditha Perera, Luca Zancato, Alessandro Achille, Parinder Bhatia, and Stefan O Soatto. 2023. Linear Spaces of Meanings: Compositional Structures in Vision-Language Models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), 15349–15358. <https://api.semanticscholar.org/CorpusID:257766294>
- [47] Nico Van de Weghe, Lars De Sloover, Anthony Cohn, Haosheng Huang, Simon Scheider, Renée Sieber, Sabine Timpf, and Christophe Claramunt. 2025. Opportunities and challenges of integrating geographic information science and large language models. *Journal of Spatial Information Science* 30 (2025), 93–116.
- [48] Michael Völske, Alexander Bondarenko, Maik Fröbe, Benno Stein, Jaspreet Singh, Matthias Hagen, and Avishek Anand. 2021. Towards Axiomatic Explanations for Neural Ranking Models (*ICTIR '21*). Association for Computing Machinery, New York, NY, USA, 13–22. <https://doi.org/10.1145/3471158.3472256>
- [49] Mengru Wang, Yunzhi Yao, Ziwen Xu, Shuofei Qiao, Shumin Deng, Peng Wang, Xiang Chen, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. 2024. Knowledge Mechanisms in Large Language Models: A Survey and Perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 7097–7135. <https://doi.org/10.18653/v1/2024.findings-emnlp.416>
- [50] Yuzhang Xie, Jiaying Lu, Joyce Ho, Fadi Nahab, Xiao Hu, and Carl Yang. 2024. PromptLink: Leveraging Large Language Models for Cross-Source Biomedical Concept Linking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) (*SIGIR '24*). Association for Computing Machinery, New York, NY, USA, 2589–2593. <https://doi.org/10.1145/3626772.3657904>
- [51] Yutaro Yamada, Yihan Bao, Andrew Kyle Lampinen, Junjo Kasai, and Ilker Yildirim. 2024. Evaluating Spatial Understanding of Large Language Models. *Transactions on Machine Learning Research* (2024). <https://openreview.net/forum?id=xkiflKcw3>
- [52] Ilker Yildirim and L.A. Paul. 2024. From task structures to world models: what do LLMs know? *Trends in Cognitive Sciences* 28, 5 (2024), 404–415. <https://doi.org/10.1016/j.tics.2024.02.008>
- [53] Zeping Yu and Sophia Ananiadou. 2025. Understanding and mitigating gender bias in llms via interpretable neuron editing. *arXiv preprint arXiv:2501.14457* (2025).
- [54] Qianheng Zhang, Song Gao, Chen Wei, Yibo Zhao, Ying Nie, Ziru Chen, Shijie Chen, Yu Su, and Huan Sun. 2025. GeoAnalystBench: A GeoAI benchmark for assessing large language models for spatial analysis workflow and code generation. *Transactions in GIS* 29, 7 (2025), e70135.
- [55] Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zucco. 2024. A setwise approach for effective and highly efficient zero-shot ranking with large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 38–47.