AGILE: GIScience Series, 7, 6, 2026. https://doi.org/10.5194/agile-giss-7-6-2026
Proceedings of the 29th AGILE Conference on Geographic Information Science, 16–19 June 2026.
Eds.: Evelyn Uuemaa, Alexander Kmoch, Holger Virro, and Simon Scheider
This contribution underwent peer review based on a full paper submission.
© Author(s) 2026. This work is distributed under the Creative Commons Attribution 4.0 License.

Geo-Alignment of Vague Cognitive Regions: Representing Uneven Cognitive Geographies of Large Language Models

Mina Karimi [1], Krzysztof Janowicz [1], Zilong Liu [1], Songlin Wang [1], and Annika Süß[1]

[1]Department of Geography and Regional Research, University of Vienna, Austria

Correspondence: Mina Karimi (mina.karimi@univie.ac.at)

Abstract. Vague cognitive regions (VCRs) such as *Levant* or *Bible Belt* play a central role in how people reason about space and place, despite lacking clear boundaries or formal definitions. Prior studies in behavioral geography and GIScience have used human surveys or crowd-sourced social media data to delineate human perception and cognition of such regions. In this paper, we introduce a new AI-based approach, which uses foundation models, particularly large language models (LLMs), to represent VCRs. Then, we challenge the results by arguing that LLMs exhibit uneven cognitive geographies, representing some VCRs more coherently and stably than others. We introduce geo-alignment as an analytical lens to examine the discrepancies between different representations. Focusing on comparative cases such as *the Alps*, *Northern-Southern California*, *the Sahara*, and *Kashmir*, we show variations that systematically shape LLM-derived VCRs. Rather than treating misalignment as a modeling error, we conceptualize it as a signal of unequal global cognitive visibility embedded in training data. The paper contributes a methodological framework for analyzing VCRs through the lens of geo-alignment and advances a methodological GIScience perspective on the spatial knowledge encoded in LLMs.

Submission Type. Theory, Case Study, Analysis

BoK Concepts. [CF2-1] Perception and cognition of geographic phenomena; [CF3] Cognitive, linguistic and social foundations, [GS5] Ethical aspects

Keywords. Geo-Alignment, Geographic Bias, Vague Cognitive Region, Large Language Models, Knowledge Representation

1 Introduction

Vague cognitive regions (VCRs) such as *downtown* or the *Global South* are a core component of how humans conceptualize and communicate about places and regions. Yet, they lack precise geometric definitions (Montello, 2003) that would make them suitable for a direct transition into (geographic) information systems. These regions are not merely imprecise versions of, for instance, administrative units. Instead, they are cognitive and linguistic constructs shaped by public narratives, shared or individual associations and beliefs, culture, and lived experiences (Montello et al., 2003).

Understanding how such regions are conceptualized, communicated, and operationalized has been a long-standing interest in geography and spatial cognition, particularly within behavioral geography (Montello, 2003, 2013). Research has shown that VCRs are characterized by *core–periphery structures* (Copus, 2001; Vanolo, 2010), *prototypical locations* (Lloyd, 1994; Lakoff, 2007), and *graded (fuzzy) membership* (Zadeh, 1965; Wang and Hall, 1996). Many everyday spatial queries and information-retrieval tasks, such as "central downtown apartment", depend heavily on the spatial extent *and* graded structure of those regions (Purves et al., 2018). Everyday spatial relations, such as "inside" or "close to" as used in such queries, seem to imply straightforward topological containment. However, this assumption becomes problematic when the referenced region, such as the *city center*, has vague or indeterminate boundaries.

With recent advances in foundation models, particularly large language models (LLMs), new and complementary opportunities have emerged for extracting spatial knowledge directly from natural language at an unprecedented scale. LLMs are increasingly used to identify place references, infer spatial relations (Cohn and Blackwell, 2024), answer geographic questions, and generate region-like representations from textual descriptions without being explicitly designed to do so (Majic et al., 2024). This raises important questions about how geographic knowledge is encoded in these models and whether their spatial representations align with established geographic understanding.

However, an implicit assumption underlies much of this emerging line of work: **that LLMs encode a broadly consistent and neutral understanding of geographic space**. Growing evidence shows that this assumption is problematic. Training data are unevenly distributed across regions, languages (Ballatore et al., 2017), and cultures, leading to representation bias (Shankar et al., 2017; Janowicz, 2023) and the emergence of strong prototypes and defaults (Liu et al., 2025, 2026). As a result, geographic knowledge encoded in LLMs may reflect uneven global information landscapes rather than uniform spatial understanding, raising questions about how such representations relate to established cognitive-geographic theories of regions.

In this paper, we study these patterns through the concept of *geo-alignment*. In a broader sense, *geo-alignment* of AI refers to pluralistic approaches (Sorensen et al., 2024) to ensure that AI systems further a multitude of region-dependent goals, values, and norms (Janowicz et al., 2025a). Here, we focus more specifically on the spatial dimension of *geo-alignment* and examine the degree to which spatial representations produced by LLMs align to each other and vary across regions.

Because most VCRs lack authoritative ground-truth boundaries, evaluating such alignment is challenging. Instead of comparing model outputs to a single reference geometry, we examine the internal coherence and spatial structure of LLM-generated representations across different models and regions, and interpret these patterns through established concepts such as core–periphery structures and graded membership. Where possible, we also anchor our analysis with regions for which behavioral survey data exist, such as the well-studied case of *Northern* and *Southern California* (Montello et al., 2014). This perspective raises several research questions:

- **RQ1:** To what degree do state-of-the-art LLMs produce coherent spatial representations of VCRs?
- **RQ2:** To what degree do different LLMs agree in their representations of these regions?
- **RQ3:** How does the spatial coherence and stability of LLM-derived VCRs vary across different types of regions (e.g., physical, environmental, vernacular, or geopolitically contested regions)?
- **RQ4:** What do differences between models reveal about uneven geographic knowledge encoded in LLM training data?

The contributions of this paper are fourfold:

- We propose a systematic, grid-based workflow for extracting and comparing LLM representations of VCRs by querying multiple state-of-the-art models at regular spatial units, capturing both binary classifications and graded similarity scores to address boundary uncertainty using concepts of fuzzy region membership and core–periphery structures from cognitive and behavioral geography.
- Through four case studies, namely, *the Alps*, *Northern–Southern California*, *the Sahara*, and *Kashmir*, we demonstrate that LLMs show uneven and region-specific cognitive representations.
- We analyze where and how representations of regions agree within LLMs, and what these (dis)agreements show about the structure of geographic knowledge encoded in LLMs.
- This paper focuses on regional-scale cognitive representations and shows unevenness not simply as bias, but as a form of geo-misalignment where the quality (e.g., coherence) of representation varies geographically, and can be interpreted through established theories of VCRs.

By integrating research on vague regions, behavioral and cognitive geography, and AI bias, this work contributes to our growing understanding of how AI systems conceptualize geographic space. This matters, as the public's access to geographic knowledge is increasingly funneled through AI pipelines. The remainder of this paper is structured as follows. Section 2 introduces the related work. We propose our methods in Section 3. Section 4 presents our results. Then, we discuss our results and findings in Section 5. Section 6 concludes this paper with potential limitations.

2 Background

2.1 Vague Cognitive Regions in Geography

Understanding how VCRs are formed, communicated, and conceptualized has therefore been a core interest of geographers, spatial data scientists, and cognitive scientists, particularly at the intersection of behavioral geography, spatial cognition, and formal qualitative spatial representation (Cohn et al., 1997; Montello, 2013). Research has long shown that VCRs exhibit a **core–periphery** structure (Montello et al., 2003, 2014), where central areas are widely agreed upon while peripheral areas show increasing disagreement (Copus, 2001). These regions are often organized around **prototypical** locations (Lloyd, 1994; Lakoff, 2007) that serve as cognitive anchors for spatial reasoning, reflecting broader cognitive processes of categorization and salience (Forgas, 1983). Similar insights have emerged in geographic information retrieval research, where VCRs are modeled with graded membership rather than crisp boundaries (Jones et al., 2008; Purves et al., 2018). Together, these perspectives emphasize that spatial concepts are socially constructed and unevenly distributed across cultural and informational contexts.

Empirical studies investigating such regions have primarily relied on human-subject experiments. Through interviews, sketch maps, and grid-based rating tasks, researchers have approximated how people (collectively) delineate regions such as *downtown* (Montello et al., 2003) or *Northern California* versus *Southern California* (Montello et al., 2014). A consistent finding across these studies is the presence of well-agreed-upon regional cores accompanied by increasing disagreement toward the boundaries, reflecting the graded structural characteristic of vague spatial categories.

Many of these regions, with vernacular names such as *SoCal* or *the Balkans*, do not appear in official gazetteers (Wijegunarathna et al., 2025). This absence complicates their representation in geographic information systems and raises challenges for georeferencing and spatial reasoning (Jones et al., 2008; Hill, 2009). As a result, VCRs have been a key topic at the intersection of GIScience and qualitative spatial representation (Cohn et al., 1997).

2.2 Computational Approaches to Modeling VCRs

While survey-based approaches provide rich insights into how (groups of) people perceive and reason about space, and remain a strong empirical foundation for studying VCRs, they are difficult to scale (Gao et al., 2017). They are labor-intensive and costly, which can impose a high cognitive burden on participants as spatial resolution increases, and are often limited to small study areas or specific populations. Therefore, repeating such studies across large geographic extents, at fine spatial resolutions, or over time is challenging and often impractical. Clearly, we cannot simply ask dozens of participants to delineate all the hundreds of thousands of major vague cognitive regions and then repeat the process to track changes.

To address these limitations, later work turned to data-driven and social-sensing approaches (Resch, 2013; Liu et al., 2015). By analyzing large volumes of user-generated content from social media, travel blogs, and other online sources, researchers were able to infer regional concepts from collective digital traces (Gao et al., 2017). These approaches offered clear advantages in scalability and temporal coverage, and closely reproduced patterns similar to those found in human surveys. However, social sensing also introduced new challenges. Social media data are unevenly distributed, biased toward specific populations and regions (Karimi and Mesgari, 2023), and shaped by platform-specific factors (Hecht and Stephens, 2014; McKenzie et al., 2020). Place references can be noisy or ambiguous, and the need to predefine search terms may exclude alternative or less dominant regions overall. Hence, social sensing only provides a partial and uneven view on vague cognitive regions.

Recent advances in LLMs provide a new opportunity to examine how geographic knowledge embedded in large textual corpora becomes encoded in AI systems (Razavi et al., 2025). If LLMs encode geographic knowledge derived from large corpora of text containing geographic references and descriptions, they may implicitly encode spatial concepts and regional knowledge, and their internal representations may reflect similar core-periphery cognitive structures. Investigating these representations offers a complementary perspective on how geographic knowledge circulates through digital infrastructures and becomes embedded in contemporary AI systems (Janowicz et al., 2025b). However, because training data are unevenly distributed across regions and languages, the geographic knowledge encoded in LLMs may itself be uneven. Hence, investigating how LLMs represent VCRs provides insight not only into geographic knowledge representation but also into the biases and knowledge gaps embedded in large-scale AI systems (Liu et al., 2025; Shypula et al., 2025).

In this study, we build on these research traditions by examining how LLMs represent VCRs and how these representations vary across models and regions. Rather than treating LLM outputs as substitutes for human cognition, we interpret them as reflections of the geographic knowledge embedded in training data. Hence, we argue that uneven representations should not be reduced simply to technical *errors*, but can instead be understood as an *opportunity* to study the geographic distortions present in large-scale AI models, their training data, and their deployment. This perspective allows us not only to reproduce spatial patterns, but to interpret them in light of established theories of VCRs and spatial cognition.

3 Data and Methods

In this section, we outline our approach in detail. We introduce the selection of case studies, spatial resolution, LLMs, output formats, and the prompting settings.

3.1 Case Studies

There are potentially hundreds of thousands of VCRs around the world, such as *Levant*, *Outback*, *Central Europe*, *Koreatown*, *Bible Belt*, *Amazon Basin*, *Balkans*, *Swiss Plateau*, *Cottage Quarter*. In this paper, we select four of them, which are located in different regions across the world. These regions are the *Alps*, *Northern-Southern California*, *Kashmir*, and the *Sahara*.

All data used in this study is composed of AI-generated responses to the prompts we designed. Since these responses are machine-generated and, in some cases, do not strictly follow the specified system instructions, additional cleaning and preprocessing is necessary. To maintain consistency, all responses are standardized to the required output format: either *True* or *False* (*North* or *South* for *California*) for binary outputs, or a similarity score within the range $[0, 1]$ for similarity-based outputs.

3.2 Models

We first introduce the LLMs used in this study. We selected three LLMs for our analysis; see Table 1. Our selection includes both open-source and proprietary models from different providers. The *version* denotes the specific model snapshot used in our experiments. The *supported modalities* address the acceptable types of input data for the model. We extract the *knowledge cutoff* dates from each model's official documentation. In case of *DeepSeek-V3*, we verify the cutoff date based on the publicly released training data and model documentation; for proprietary models, we rely on the developer's published information. We use OpenAI API[1], DeepSeek API[2], and Google Gemini API[3] to get the responses from the GPT-4o, DeepSeek-V3, and Gemini-2.5-Flash models.

The models were selected based on their state-of-the-art performance, API availability for reproducibility and systematic querying, and their relevance and increasing use within the GeoAI and GIScience communities. Using the latest stable versions enables a systematic comparison of geo-alignment across models with differing training data and architectures.

3.3 Spatial Resolution

To apply the proposed approach across all case studies, we follow a unique methodological structure. First, the study areas are defined. For the *Alps*, this includes the countries that contain portions of the Alpine region: Austria, France, Germany, Italy, Liechtenstein, Slovenia, and Switzerland. For *Northern-Southern California*, the study area corresponds to the U.S. state of California. Because *Kashmir* is a politically contested region spanning three countries—Pakistan, India, and China—we do not consider the full national territories in our analysis. Instead, we focus on the core Kashmir region and apply a $100km$ spatial buffer, extending approximately 25% beyond the estimated region extent to capture the full range of potential model responses including peripheral areas. For the *Sahara*, the study area comprises all countries that contain parts of the desert and Sahel, including Algeria, Benin, Burkina Faso, Cameroon, Central African Republic, Chad, Côte d'Ivoire, Djibouti, Egypt, Eritrea, Ethiopia, Gambia, Ghana, Guinea, Guinea-Bissau, Liberia, Libya, Mali, Mauritania, Morocco, Niger, Nigeria, Senegal, Sierra Leone, Somalia, South Sudan, Sudan, Togo, and Tunisia.

Secondly, we divide the study area into $20km \times 20km$ square grids. The $20km \times 20km$ grid represents a trade-off between spatial detail and computational feasibility, preserving important boundary effects while limiting the number of API calls. A consistent resolution is applied across all regions to ensure comparability. Although alternative resolutions may suit large regions such as the *Sahara*, varying grid sizes would introduce methodological inconsistency. Table 2 shows the total number of grids and area covered by them for each VCR. Then, we compute the central coordinates of each grid using `Point_in_Polygon` analysis. These geographical coordinates are calculated in WGS84 format and are later used as the input of our designed prompts to collect the responses of LLMs. Prior work (Karimi et al., 2026 under review) has shown that model outputs remain largely invariant across representational choices such as center points, versus WKT geometries for cells or hexagons as used by Montello et al. (2014).

3.4 Output Format

After obtaining the central coordinates of the overlaid grids on the study area, we apply our proposed LLM-based approach to the centroids of grids. We design our approach using two different prompting strategies, including:

- Binary outputs: In this strategy, We force the models to make definitive decisions and create crisp boundaries. Hence, we ask them to respond with binary values, either `True` if the coordinate is inside the *region* or `False` if the coordinate is outside the *region*. The values are `N` or `S` for *California*. This also ensures answers are always one-token long.
- Similarity outputs: In this strategy, we account for the inherent fuzzy nature of VCRs by asking models to make non-binary decisions and to return a similarity score in the range $[0, 1]$, indicating the degree of belonging of the coordinates to the queried VCR. Scaling aside, this setup is in line with the data-synthesis study by Gao et al. (2017).

These two strategies allow us to capture both the concrete boundary interpretations of regions and the gradual, fuzzy transitions that characterize VCRs.

3.5 Prompting Configurations

The system prompt, user prompt, and model response are assessed within a single interaction using the selected model, applied to the centroid input query, with the LLM-sampling temperature fixed at 0. In brief, the *system prompt* establishes the model's context and specifies roles and constraints, the *user prompt* represents the task or question to be answered, and the model's *temperature* controls how narrowly or broadly probable output tokens are sampled, with higher values producing a flatter distribution. For all models, the maximum number of tokens is set to 10. However, for Gemini-2.5-Flash, this low-token limit results in non-informative outputs, as the model often repeats the prompt or generates

[1] https://platform.openai.com
[2] https://platform.deepseek.com
[3] https://ai.google.dev/

Table 1. Details of three selected large language models used in our experiments.

Model Name	Version	Open Source	Supported Modalities	Knowledge Cutoff Date
GPT-4o	gpt-4o-2024-08-06	No	Text & Image	October 2023
DeepSeek-V3	deepseek-v3-0324	Yes	Text	July 2024
Gemini-2.5-Flash	Gemini-2.5-Flash	No	Text & Image	January 2025

Table 2. The number of grids and total area of each vague cognitive region.

Region	Number of Grids	Area (km^2)
Alps	7,946	3,178,400
California	1,674	657,999.701
Kashmir	1,822	719,568.469
Sahara	52,753	21,030,390.439

partial reasoning before reaching an answer. To address this issue, the `max_tokens` parameter was increased (100) to avoid truncated outputs; this primarily affected intermediate reasoning rather than final responses, which were consistently extracted and verified.

Listings 1 and 2 show the prompt setting designed for the experiment. For each coordinate, the prompt defined in Listing 1 is employed to get the binary values (`True` if the coordinate is inside the *region* or `False` otherwise) as the response of the LLMs. Similarly, the prompt shown in Listing 2 is employed for each coordinate to obtain the similarity value of the region. While for *California*, the models are forced to respond with `N` and `S` in binary-based prompting. For similarity-based prompting, higher similarity scores indicate alignment with *NorCal*, while lower scores indicate alignment with *SoCal*. This is done because for this VCR we actually do have ground truth from Montello et al. (2014).

Listing 1 Input format, system prompt, user prompt, and model response in a single session using the selected model for binary prompts on centroids applied to the Alps, Kashmir, and the Sahara.

```
Input: Geographic coordinates of the centroids
↪  ({lat}, {lon})

System prompt: Return only True or False.

User: Is this coordinate ({lat}, {lon}) inside the
↪  {region_name} or not?

LLM: { "response": "True" or "False" }
```

3.6 Agreement Metrics

To examine the level of agreement among different spatial delineations, we use three well-known metrics designed to quantify the agreement across alternative representations of the same VCR. Specifically, we employ the *Jaccard Index*, also known as *Intersection over Union*, the *Dice Coefficient*, and the *Pearson Correlation Coefficient*. The first two metrics are applied to binary outputs, whereas the last one is used to assess similarity-based outputs.

Listing 2 The input format, system prompt, user prompt, and model response in a single session using the selected model for similarity prompts on centroids applied to the Alps, Kashmir, and the Sahara.

```
Input: Geographic coordinates of the centroids
↪  ({lat}, {lon})

System prompt: Return only a number between 0 and 1,
↪  as the similarity score, where 0 means no
↪  similarity and 1 means entirely similar.

User Prompt: How similar is the region at coordinate
↪  ({lat}, {lon}) to [region_name]?

LLM: { "response": "Similarity score in [0, 1]" }
```

3.6.1 Jaccard Index

To quantify how closely two spatial region delineations align, we employ the Jaccard Index. Rather than relying on absolute spatial coverage, this metric evaluates relative agreement by comparing the shared area of two regions (R_1 and R_2) with their total combined extent. In other words, it measures the proportion of spatial units (grids) that are common to both delineations, as formalized in Eq. (1).

$$J(R_1, R_2) = \frac{|R_1 \cap R_2|}{|R_1 \cup R_2|} \tag{1}$$

The index ranges from 0 to 1, where values approaching 1 indicate strong spatial agreement between the two delineations. A value of 0 reflects no overlap, whereas a value of 1 corresponds to complete coincidence. In this study, the Jaccard Index serves as a measure of spatial consensus, which enables comparisons of how different models delineate a VCR.

3.6.2 Dice Coefficient

The Dice Coefficient provides a complementary perspective by placing greater emphasis on the overlapping portion of two regions. This property makes it particularly responsive in situations where the

total sizes of the regions differ. The measure is defined in Eq. (2).

$$D(R_1, R_2) = \frac{2|R_1 \cap R_2|}{|R_1| + |R_2|} = \frac{2J}{1+J} \tag{2}$$

Similar to the Jaccard Index, the Dice values range between 0 and 1, with higher values indicating greater similarity. In this analysis, it is used to quantify regional overlap, offering a different sensitivity compared to the Jaccard Index alone.

3.6.3 Pearson Correlation Coefficient

While overlap-based metrics are effective for binary comparisons, they do not capture similarities in how regions are graded or weighted. To address this limitation, we employ the Pearson Correlation Coefficient (ρ), which measures the strength of a linear relationship between two continuous variables (S_1 and S_2). As shown in Eq. (3), ρ relates the covariance between the variables to their respective standard deviations.

$$\rho_{S_1,S_2} = \frac{\mathrm{cov}(S_1, S_2)}{\sigma_{S_1}\sigma_{S_2}} \tag{3}$$

The coefficient takes values between -1 and 1, where positive values indicate consistent variation, and negative values indicate inverse relationships. Unlike the Jaccard and Dice measures, which are applied to binary classifications, we use Pearson Correlation Coefficient to assess how similarly models assign graded similarity values across space. In other words, we evaluate the extent to which different models produce comparable fuzzy-like membership scores for the same spatial units.

3.6.4 Moran's I Spatial Autocorrelation

We employ Moran's I as a measure of spatial autocorrelation to quantify the degree to which LLM-generated representations of VCRs exhibit spatial clustering or dispersion. This allows us to assess the extent to which model outputs form coherent spatial patterns, which is directly relevant to evaluating the structure and consistency of VCRs. This approach aligns closely with the notion of spatial coherence addressed in our research questions.

3.7 Data and Software Availability

All data and code are available in a public GitHub repository[4]. The data include the spatial extent of the case studies, the revised prompts, as well as the generated responses from different LLMs. The codes include our data collection process, data analysis, and visualization.

To support exploration and interpretation of the proposed LLM-based approach, we developed a web-based map application[5] that visualizes the resulting geographic representations. Fig. 1 presents the interface of the web application we have developed to represent the spatial patterns produced by LLMs in an interactive manner, illustrating LLM-derived VCRs. The system is implemented using standard Web technologies, including HTML, CSS, and JavaScript, together with the OpenLayers library for interactive mapping. Spatial data are managed and served through the open-source GeoServer, while external Web-based access to the application is enabled via ngrok.

4 Results

Here, we analyze the results obtained in two output formats for each region across various models. Generally, binary (True/False) outputs tend to emphasize categorical inclusion. They produce sharp, sometimes fragmented boundaries, especially near region edges. In addition, they mask internal uncertainty as cells are either "in" or "out", even when the region overall is vague. However, similarity scores capture a gradual cognitive transition from the core to the periphery and are more aligned with the idea of VCRs. They indicate how each LLM conceptualizes centrality, extent, and ambiguity.

To examine the overlap between different LLMs and quantify the agreements between models in representing VCRs, we present the results of applying metrics introduced in Section 3.6 including the *Jaccard Index*, the *Dice Coefficient*, and the *Pearson Correlation Coefficient*. The Jaccard Index measures exact overlap relative to union. The Dice Coefficient is an overlap metric that emphasizes shared elements. The Pearson Correlation Coefficient measures linear consistency of model outputs. To assess output stability, we repeated a subset of prompts 10 times and observed highly consistent results, with negligible variation in spatial patterns. The uncertainty observed at region boundaries corresponds to the concept of fuzzy or graded membership (Zadeh, 1965; Wang and Hall, 1996).

4.1 Alps

While the *Alps* may appear as a clearly (physically) delineated region, they are not. There are multiple (international) conventions defining the area belonging to the *Alps* and many less formal notions, e.g., including the Alpine foothills. Fig. 2 shows the results obtained by applying our LLM-based approach to estimate the

[4] https://github.com/MinaKarimi/Geo-alignment-of-AI-in-VCRs

[5] The application is available at https://sean-stanchable-scrofulously.ngrok-free.dev/sdss/.

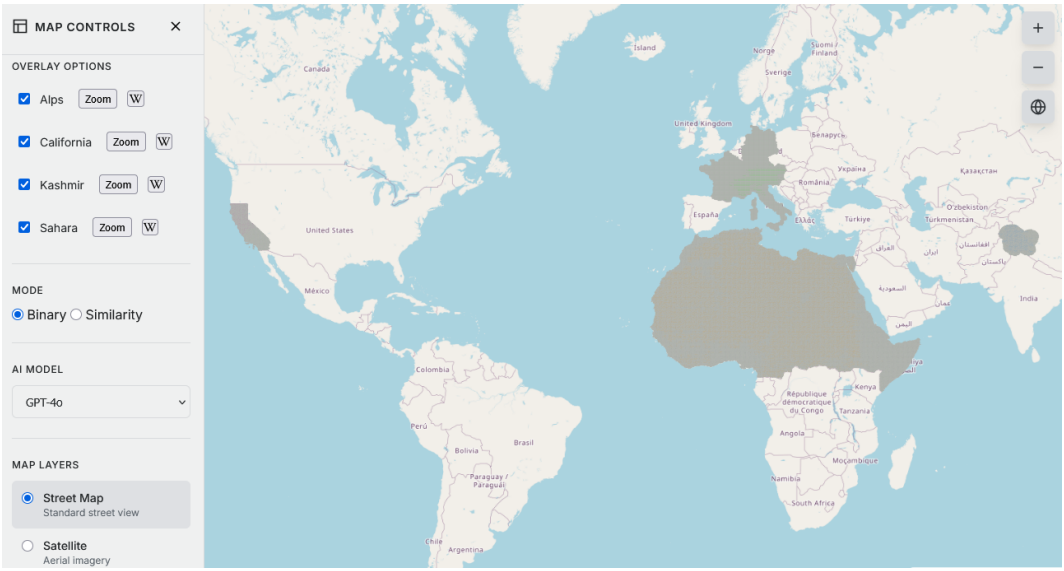


Figure 1. The interface of LVCRS: LLM-derived VCR System.

Alps region using two visualization methods: binary classification (True/False) and similarity scores.

Across all three LLMs, the representations of the *Alps* show a clearly identifiable strong core, which aligns with the east–west Alpine arc and indicates the prototype theory of cognitive models in representing VCRs (Lloyd, 1994; Lakoff, 2007). The high agreements between different LLMs is visible. This confirms the *Alps* as a semi-vague cognitive region, in which the core area is physically (and culturally) grounded but the foothills are mediated. This means that all models correctly identify the Alpine arc with their central mountain core, while they differ in peripheral areas. This finding is consistent with the core-periphery structure of VCRs (Vanolo, 2010; Copus, 2001).

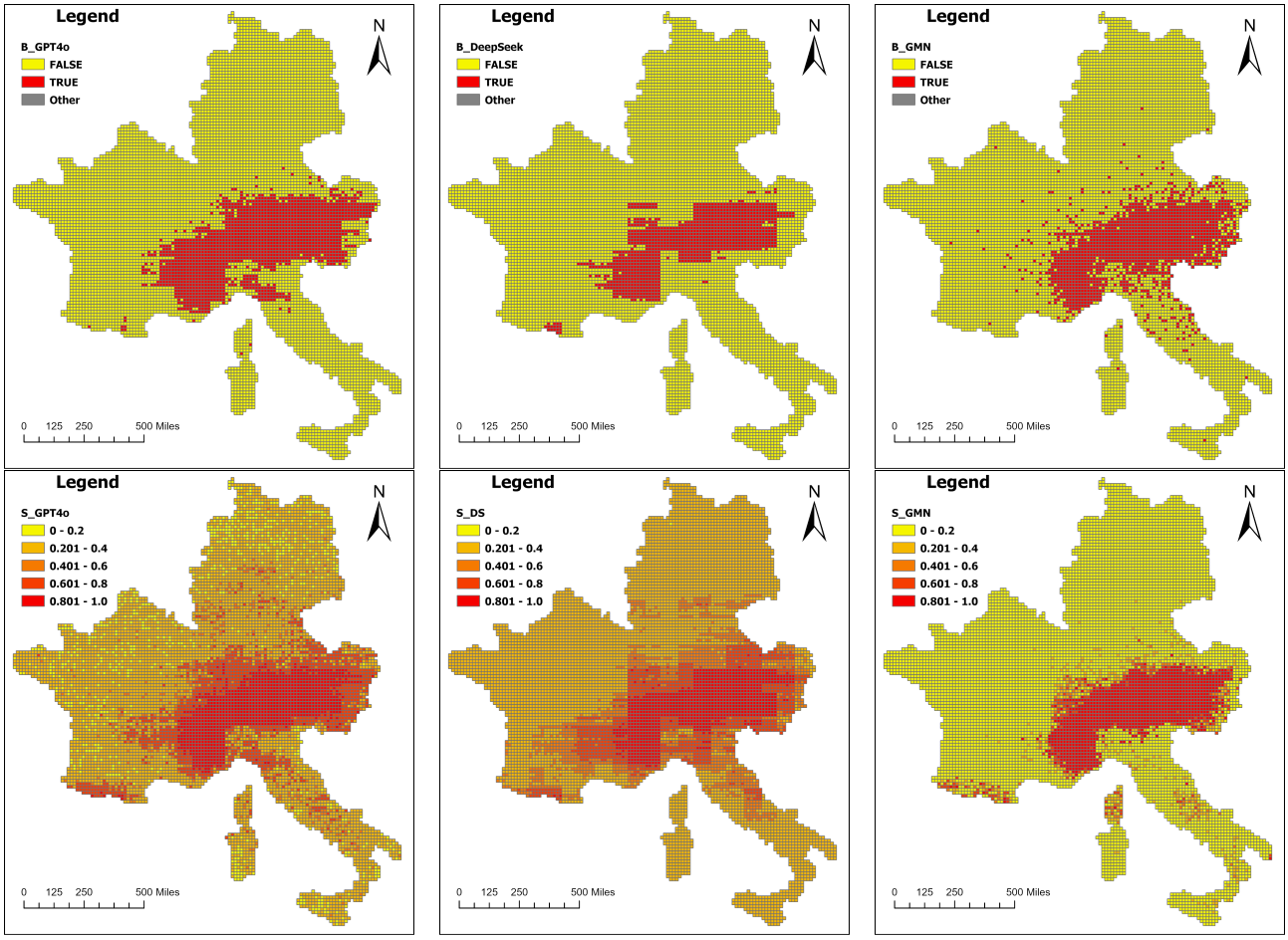
Binary delineations of GPT-4o model are highly compact and continuous. `True` cells follow the Alpine chain with a minimal leakage into surrounding lowlands. Non-mountainous areas are clearly excluded from the main region. Similarity delineations also show a strong central ridge of high similarity scores with a gradual decay towards the edges, specifically north and south of the study area where there are more flat regions.

DeepSeek-V3 binary representation still follows the Alpine arc but with slight fragmentation and more irregular boundaries. The block-like inclusion shows its limitation in representing the *Alps*. Although it performs better in similarity-based representation, but it delineates broader mid-range similarity values with less contrast between core and periphery. In addition, the *Alps* emerge as a regional corridor, not just a mountain chain.

The widest inclusion among the three models is represented in Gemini-2.5-Flash binary results. Noticeable spillover into adjacent regions and some scattered "True" regions across the study area are clearly visible. High similarity scores in the core zone are located in Central Europe. There is a sharper distinction between the core and periphery compared to GPT-4o and DeepSeek-V3.

In distributional semantic models, conceptual relationships emerge from patterns of linguistic co-occurrence, not strict geography. Words that frequently appear in similar contexts become semantically associated, forming clusters in vector space (Lenci and Sahlgren, 2023). As a result, terms such as *Alps* become linked with terms such as "Switzerland", "Austria", "Alpine lifestyle", "Ski resorts", "Central Europe", etc. Over time, this builds a higher-level thematic groupings like **"Alpine Europe"** that reflect discourse-based associations rather than strict spatial proximity or just "high-elevation mountains" (Mehler et al., 2020). That may explain why binary outputs include foothills and peripheral zones, or some high similarity values appear outside the highest elevations in culturally similar regions.

Table 3 shows the pairwise agreements between three LLMs using three metrics in the *Alps*. The overlap between GPT4o and Gemini reveals the best overlap across the *Alps* with a Jaccard Index of 0.60, and Dice Coefficient of 0.75 while slightly lower Pearson Correlation Coefficient (0.82). This indicates that while they agree on items, the exact values or probabilities may vary more. The overlap between GPT4o and DeepSeek is moderate (Jaccard Index is 0.56 and Dice Coefficient is 0.72), but interestingly, it returns the highest linear correlation with Pearson Correlation Coefficient of 0.86. This finding shows that representations of these two models are more consistent in magnitude even if exact matches are fewer. The overlap between DeepSeek and Gemini is lowest across two metrics (Jaccard 0.53, Dice 0.69), which considers binary representations, while the similarity-based representations correlate with a Pearson value of 0.83. Hence, this pair has the least agreement and consistency.



(a) GPT-4o (b) DeepSeek-V3 (c) Gemini-2.5-Flash

Figure 2. The LLM-based representations of the Alps using the proposed approach in two output formats.

The LLMs show moderate to high (not near perfect) consensus on the *Alps* data. GPT4o seems central since it has relatively strong agreement with both DeepSeek and Gemini with GPT4o-Gemini agreeing most on shared items, while GPT4o-DeepSeek shows the strongest linear consistency. DeepSeek-Gemini is the weakest and the least aligned, implying their outputs differ more for the *Alps* dataset. This reveals some systematic differences in how these models interpret the *Alps* dataset.

In addition, the choice of output format (binary-based with concrete crisp boundaries or similarity-based with gradual transitional zones), which leads to the choice of agreement metric matters. Binary- and similarity-based representations give slightly different insights since the results from overlap metrics (Jaccard/Dice) are different from correlation metric (Pearson).

Table 3. Pairwise agreements between LLMs: the Alps.

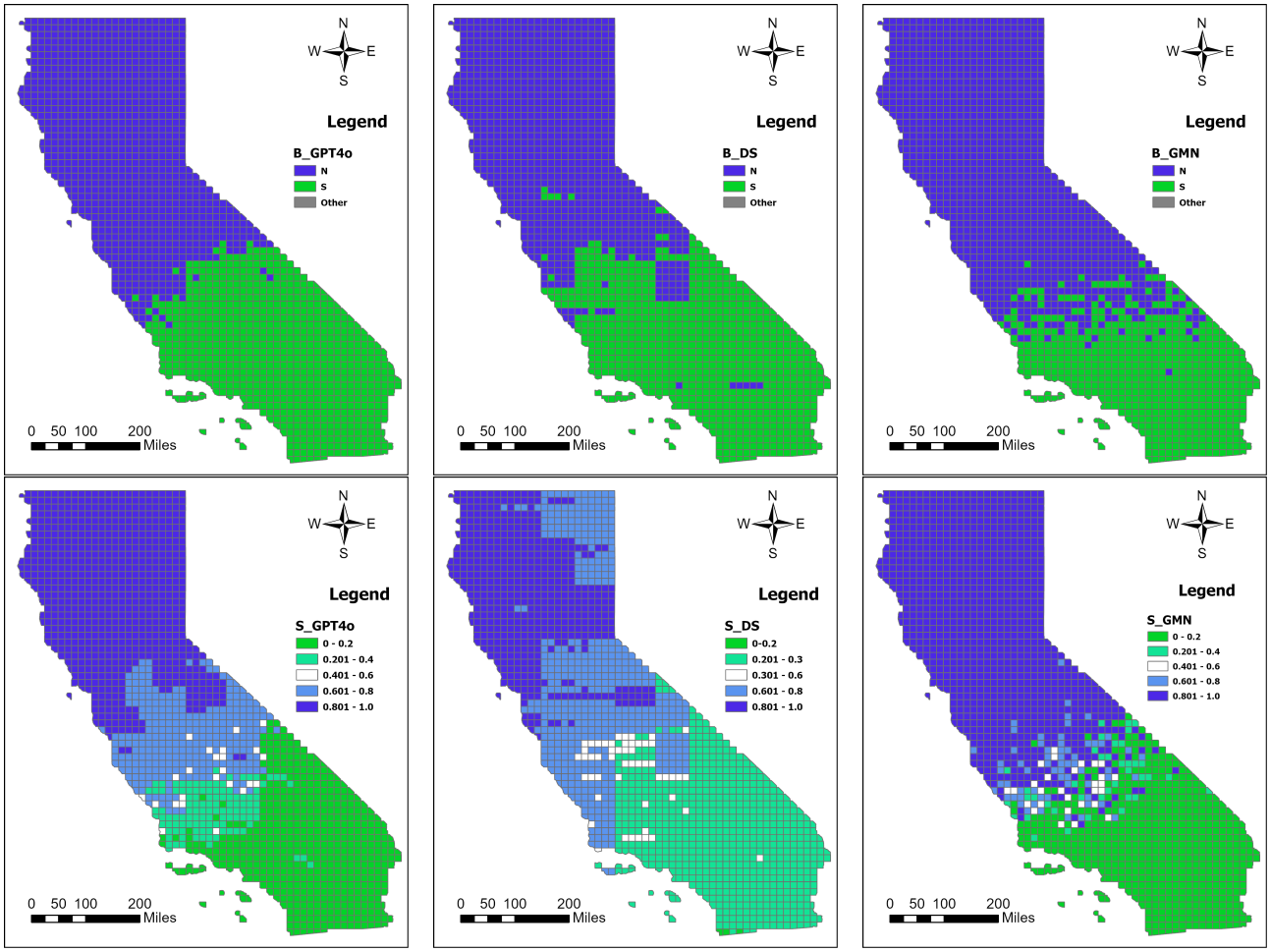
Model Pair	Jaccard	Dice	Pearson
GPT4o–DeepSeek	0.56	0.72	0.86
GPT4o–Gemini	0.60	0.75	0.82
DeepSeek–Gemini	0.53	0.69	0.83

4.2 California

Unlike the *Alps* or the *Sahara*, there is no geomorphological discontinuity that separates the two regions. Instead, the distinction is entirely vernacular, rooted in cultural identity, everyday discourse, and media representations. This makes *California* an especially useful case to examine how LLMs internalize and reproduce cognitively constructed regions rather than physically grounded ones. Fig. 3 presents the delineated *NorCal* and *SoCal* regions, visualized using two output formats: binary classification and similarity scores.

GPT-4o produces a clear north–south division in its binary output, with a relatively sharp boundary aligned with a commonly described cited cultural demarcations (roughly Central California). The similarity output confirms this pattern by showing a strong contrast between northern and southern parts of the state and a transitional zone between them.

DeepSeek-V3, by contrast, presents a more ambiguous output. In the binary delineation, classifications not only near the north–south boundary but even in more northern or southern are scattered, showing high uncertainty. The



(a) GPT-4o (b) DeepSeek-V3 (c) Gemini-2.5-Flash

Figure 3. The LLM-based representations of Northern and Southern California using the proposed approach in two output formats.

similarity delineation also emphasizes this ambiguity, indicating a broad transition zone characterized by mid-range similarity values across not only the Central California but also in the north-east and south-west areas.

Gemini-2.5-Flash specifies more area as *NorCal* compared to *SoCal*. Its binary delineation shows the least coherent structure among the three models in the border. Its binary representation is noisy, especially in the central part and a bit of south part of the state. The similarity output is uneven and fragmented, with higher values in the transition zone. Generally, Gemini assigns more area to *NorCal* compared to *SoCal*.

The pairwise agreements between three LLMs using three metrics in *California* are listed in Table 4. The Jaccard Index of 0.87 for GPT4o-DeepSeek presents the strongest exact overlap, while the value of 0.79 for DeepSeek-Gemini shows the weakest overlap. The Dice Coefficient between GPT4o and DeepSeek is 0.93, which shows an extremely strong agreement. The lowest value is between DeepSeek and Gemini (0.88), which is still high. This confirms that models largely agree on which elements belong to *NorCal* versus *SoCal*, even if there are slight differences.

The highest Pearson Correlation is visible GPT4o-Gemini (0.93), which demonstrates the strongest linear correlation, slightly higher than the correlation of GPT4o-DeepSeek(0.92). The lowest correlation value (0.87) is obtained from the comparison of DeepSeek and Gemini, but it is still reasonably strong. This indicates that not only do the models agree on inclusion, but their weighting of elements is very consistent. This indicates that GPT4o and Gemini agree slightly less on exact items but are very consistent in scoring or probabilities, showing some differences in boundaries or priorities. The DeepSeek-Gemini shows the lowest agreement across all metrics (Jaccard 0.79, Dice 0.88, Pearson 0.87).

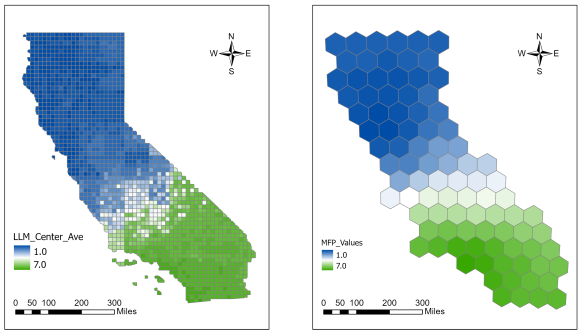
The strong agreement observed for *California* shows high cultural salience, which aligns with the concept of cultural salience in behavioral geography (Montello, 2013)—the prominence, noticeability, and perceived relevance of particular topics, beliefs, or places within the shared culture of a group (Forgas, 1983). Regions that appear frequently in cultural discourse tend to develop stronger

shared mental representations and more clearly defined cores (Montello et al., 2014; Purves et al., 2018). It is the most "stable" region across LLMs, likely because its geographic definition is clearer and straightforward (state boundaries, major features) and it appears frequently in discourse.

Table 4. Pairwise agreements between LLMs: California.

Model Pair	Jaccard	Dice	Pearson
GPT4o–DeepSeek	0.87	0.93	0.92
GPT4o–Gemini	0.83	0.91	0.93
DeepSeek–Gemini	0.79	0.88	0.87

In addition, we compare our results for *California* with the available human survey (Montello et al., 2014). To do that, we calculate the average similarity values of three models and scale them in the range of $[1, 7]$, which corresponds to the former study, which 1 indicates more similarity to *NorCal* and 7 to *SoCal*. Fig. 4 visualizes this comparison.



(a) LLM Average (b) MFP Study

Figure 4. The comparison of the average LLM scores with human survey results (MFP Study).

4.3 Kashmir

Fig. 5 shows different delineations of *Kashmir* using the proposed approach in two output formats.

The results represent that GPT-4o produces a compact and contiguous core, with limited spillover. However, peripheral areas are selectively included. The similarity scores are strong central peak with clear radial decay. The similarity-based representation indicates a "core Kashmir" area, surrounded by zones of declining confidence. Hence, GPT-4o model appears to rely on a geopolitically grounded understanding, which includes core valley plus surroundings.

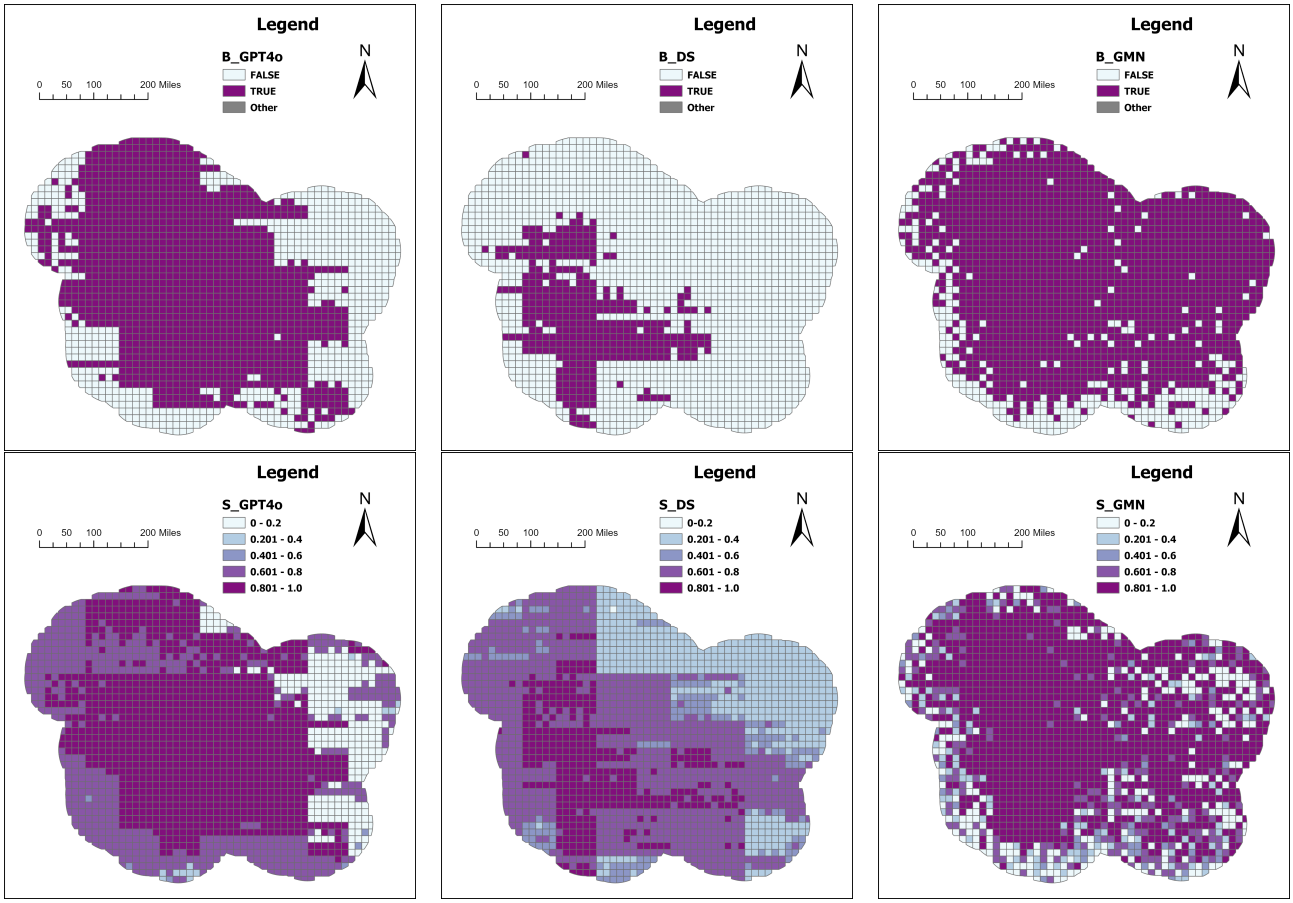
The DeepSeek-V3 results are much more fragmented. The linear or block-like inclusions do not capture the geographical extent of the core region. The similarity results represent a flatter distribution with fewer very-high similarity cells. The core area is less pronounced, with broader mid-range scores. These findings indicate that DeepSeek-V3 seems to encode *Kashmir* more as a contested region than a spatially compact one. This confirms that DeepSeek-V3 is unable to effectively delineate a VCR such as Kashmir and demonstrates a limited understanding of its spatial extent.

Binary-based representation of Gemini-2.5-Flash model indicates a broad inclusion with many scattered `True` cells. The delineation shows less spatial discipline, while noise is clearly visible in the study area. The similarity-based representation shows that high values are widely distributed, even near the edges. In addition, a weak gradient from center to periphery is notable. The results indicate that Gemini-2.5-Flash appears to adopt a looser, culturally expansive notion of *Kashmir*.

Table 5 summarizes the pairwise agreements between three LLMs using three metrics in *Kashmir*. Only GPT4o and Gemini agree on a substantial subset of elements, with a moderate Jaccard Index of 0.64. DeepSeek diverges strongly from both, reaching Jaccard indices of 0.27 and 0.23 for agreement with GPT4o and Gemini, respectively. The Dice Coefficient pattern reflects the same Jaccard pattern. The values in GPT4o-Gemini is strong (0.78) but low in GPT4o-DeepSeek (0.42) and DeepSeek-Gemini (0.37). Even when overlap is weighted more generously in Dice Coefficient, DeepSeek remains an outlier, while GPT4o and Gemini-2.5-Flash converge more on "what belongs to Kashmir".

Pearson Correlations which focus on output consistency, is where *Kashmir* becomes especially revealing. The value of 0.47 for GPT4o–DeepSeek shows a moderate to low linear relationship. However, there is a very weak correlation in GPT4o-Gemini (0.25) and almost near absence between DeepSeek and Gemini (0.14). These values indicate that even when GPT4o-Gemini select similar elements (high Jaccard/Dice), they assign very different weights, intensities, or scores to them. High overlap and very low Pearson in GPT4o-Gemini reveals agreement on what constitutes *Kashmir* but strong disagreement on how important those elements are. Low overlap and moderate correlation in GPT4o–DeepSeek indicates different selections of elements by each model but some shared structure in how outputs are distributed. The metrics for DeepSeek–Gemini are lowest across all metrics, which demonstrates these two models are fundamentally misaligned on *Kashmir*. They disagree on inclusion, emphasis, and relative importance.

These results are interesting, because *Kashmir* shows a very different agreement structure compared to the *Alps* and *California*. In a nutshell, *Kashmir* exhibits low and highly uneven agreement, both in overlap and correlation and emerges as a highly unstable and contested cognitive region, showing fragmented and model-dependent interpretations. The fragmented representation of *Kashmir* aligns with theories that



(a) GPT-4o (b) DeepSeek-V3 (c) Gemini-2.5-Flash

Figure 5. The LLM-based representations of Kashmir using the proposed approach in two output formats.

emphasize the social and political construction of regions, particularly in contested territories (Omrow, 2020).

Table 5. Pairwise agreements between LLMs: Kashmir.

Model Pair	Jaccard	Dice	Pearson
GPT4o–DeepSeek	0.27	0.42	0.47
GPT4o–Gemini	0.64	0.78	0.25
DeepSeek–Gemini	0.23	0.37	0.14

4.4 Sahara

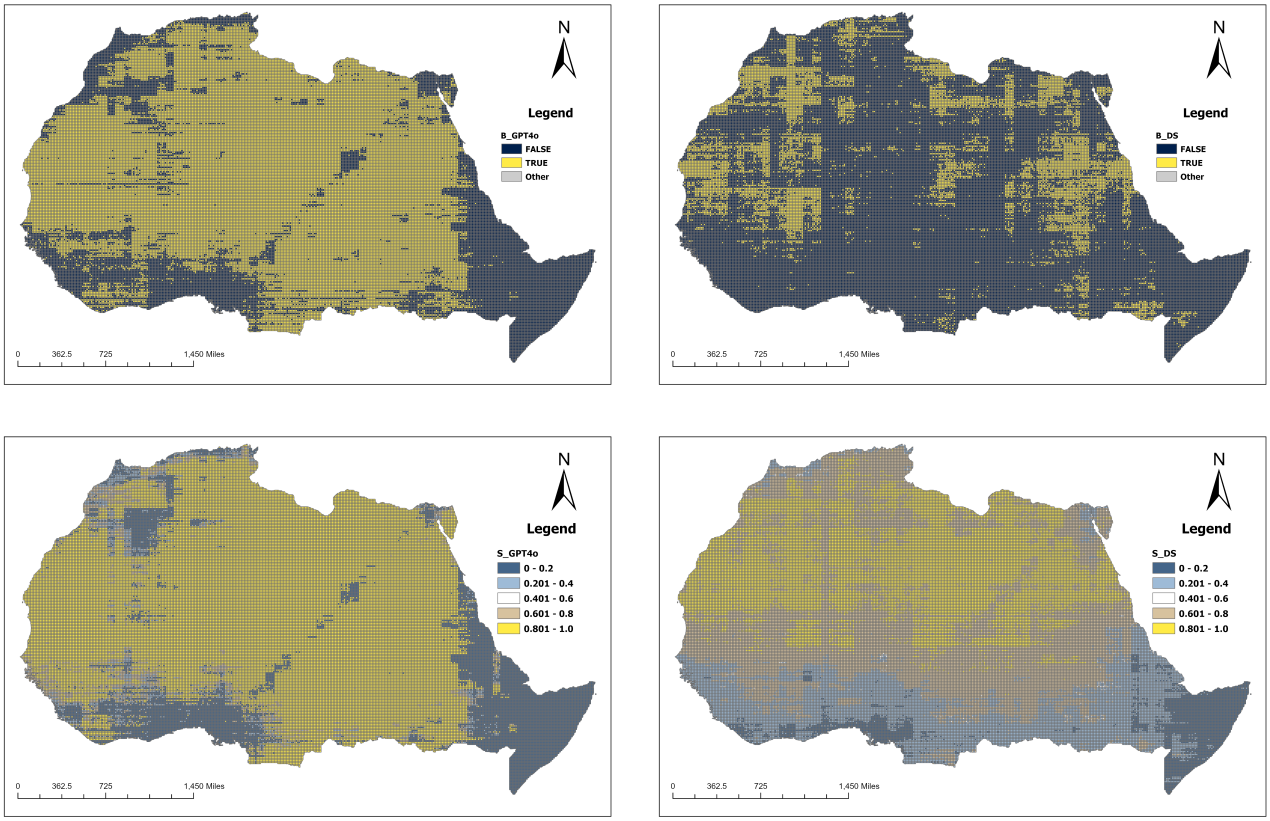
Different delineations of the *Sahara* using the approach are represented in Fig. 6 in two output formats. For the *Sahara*, we only examine two models (GPT-4o and DeepSeek-V3) since we did not collect data from Gemini due to the high number of grid cells (more than $52K$ cells) and the API costs.

Binary representations in GPT-4o are large and continuous inclusion. However, some parts of the region located in *Null Island* are clearly excluded. Although some internal fragmentation is visible in the results, the internal ambiguity is relatively low. Similarity-based representation follows quite similar pattern with high-similarity scores across central *Sahara* and Sahel, without clear core–periphery structure.

DeepSeek-V3 binary representation shows a noticeable internally heterogeneity compared to GPT-4o. There are many grid-level noises inside the desert area. Generally, DeepSeek-V3 is unable to encode the *Sahara* in binary representation. However, similarity representation shows a flatter similarity surface compared to GPT-4o, where peripheral zones receive moderate similarity rather than a sharp decrease.

The pairwise agreements between two LLMs using three metrics in the *Sahara* are shown in Table 6. Less than one quarter of elements overlap between GPT4o and DeepSeek (a Jaccard Index of 0.23), which indicates strong disagreement on "what belongs to the *Sahara*". The low weighted overlap (Dice 0.37) confirms the Jaccard results. Even when overlap is emphasized, agreement remains low. The results shows the presence of different boundary assumptions (core desert vs. extended Sahelian fringe) in each model.

Despite the low overlap, the correlation is moderate (Pearson 0.60). This is a key point that shows the two models disagree on selection of elements, but partially agree on structure by assigning similar relative



(a) GPT-4o (b) DeepSeek-V3

Figure 6. The LLM-based representations of the Sahara using the proposed approach in two output formats.

weights. Compared to *Kashmir*, the *Sahara* appears less discursively fragmented and more characterized by spatial diffuseness and gradual transitions but still not well represented compared to the *Alps* and *California*.

The weaker representation of the *Sahara* may reflect the role of information availability and cultural exposure in shaping cognitive regions (Ballatore et al., 2017; Omrow, 2020), as well as the nature of the region as a desert with uneven visibility. We also observe isolated cells near (0,0), commonly referred to as "Null Island". Similar patterns have been documented in Linked Data and geospatial systems, where erroneous records accumulate at this location (Janowicz et al., 2016), indicating that LLM-derived representations may inherit or reproduce such underlying data biases.

Table 6. Pairwise agreements between LLMs: the Sahara.

Model Pair	Jaccard	Dice	Pearson
GPT4o–DeepSeek	0.23	0.37	0.60

5 Discussion

Comparing model agreement across four case studies reveals clear and systematic differences in how regions are represented. *Northern-Southern California* stands apart from the other regions in the degree to which models converge, both on what belongs to the region and on how strongly different elements are emphasized. This is likely connected to the fact that *California* is a well-known US state and appears repeatedly in English-language texts, media coverage, and many datasets. Studies about *Northern-Southern California* are also likely in the models' training data. Nonetheless, the ability of the models to correctly encode and recall geographic coordinates and to form largely consistent regions is impressive. The result is a relatively uniform representation with less divergence across models.

The *Alps* show a slightly different, though still relatively stable, pattern. Agreement is lower than in *Northern-Southern California*, particularly when overlap is considered, but models follow quite similar internal structures. Disagreement appears mainly at the edges of the region rather than at its core. Similar to *Northern-Southern California*, the *Alps* are heavily documented, visualized, and embedded in a dense western data environment, shaped by extensive cartographic traditions,

tourism discourse, scientific research, and popular representations. This may help to explain how they are represented despite the lack of sharp boundaries.

For the *Sahara*, the picture changes significantly. Models often disagree on inclusion, indicating uncertainty not only about where the region ends but also about what it is. The *Sahara* does not have clear borders, and its transition into surrounding zones is different between the models. Much of the detailed or locally grounded material is likely not in English or does not circulate widely online, which may explain why models trained on different data mixtures produce different delineations. The key take-home message here is not only that individual models show unexpected patterns that **resemble problematic data artifacts**, e.g., GPT-4o in Fig. 6, but that model outputs do not converge among the studied systems. Hence, there is more error and low(er) agreement.

Kashmir produces the most uneven results. Disagreement is not limited to boundaries. Even when models refer to similar elements, they do not weight them in comparable ways. This shows that the issue is not only spatial but also interpretive. *Kashmir* is described through competing political frames across three countries, and the language used to refer to it shifts across sources. These inconsistencies appear to be reflected in the models' outputs, leading to representations that are less stable and more dependent on training data composition. We believe this finding complements the arguments made by Ballatore et al. (2017) before.

Summing up, vague cognitive regions situated in areas where documentation is thinner, more fragmented, or distributed across multiple linguistic and political contexts are represented less consistently, and agreement between models declines. While this may not be surprising, it is not less worrisome as interaction with AI agents more and more shapes the everyday experience of the general public. What is surprising, however, is that the theoretical concept of VCRs in LLM-based GenAI systems is directly coupled to concrete examples. Regions such as *Kashmir* are not just represented with more uncertain at their respective borders, but with less coherence within and across models. For instance, Gemini's representation of *Kashmir* has a substantially lower spatial autocorrelation (Global Morans'I 0.33) due to the patchy, variation in similarity values. This is a significant finding, as it suggests that the continuity of these regions is not preserved in the encoding. Table 7 shows the Global Moran's I values of different models in each VCR.

From a spatial cognition perspective, regions such as the *Sahara*, and *Kashmir* may lack strong prototypical cores and exhibit more diffuse core–periphery structures, leading to lower agreement and more fragmented representations. This would mean that representation quality in LLMs varies not only with data availability, but also with the intrinsic nature of region and its geographic feature type including its environmental, cultural, or geopolitical characteristics. For example,

Table 7. The Global Moran's I values of different LLMs for each VCR.

Region	GPT-4o	DeepSeek-V3	Gemini-2.5-Flash
Alps	0.87	0.95	0.89
California	0.96	0.94	0.91
Kashmir	0.79	0.90	0.33
Sahara	0.86	0.94	-

weaker and less coherent representations in the *Sahara* are not only related to the geographic location, but also to structural and representational characteristics of the region itself. This limitation cannot be explained solely by the overrepresentation of Western countries in existing datasets or by the limited number of English-language narratives. Instead, it is also likely linked to the region's graphical characteristics as a desert. Desert regions generally imply low population density, which leads to fewer recorded activities and reduced relevance in dominant data sources. This reduces the availability and consistency of spatial signals in the data from which LLMs learn. Maybe desert regions, even within the Western world, would exhibit a similarly weak representation in AI systems and lower consistency in LLM outputs. We leave a systematic comparison across different types of vague cognitive regions for future work. For now, the question of how (local) citizens can be involved in co-shaping the latent representation of these regions remains largely unanswered (Sieber et al., 2025).

Prompt language can influence representations. The use of English prompts may introduce a bias toward English-language conceptualizations, which may partly explain the higher agreement observed for California, where the dominant language aligns with the prompt language, compared to other VCRs. More broadly, LLM outputs may reflect characteristics of their training data, including potential regional or cultural biases, though such effects are difficult to isolate within this study. Notably, the fragmented and blocky patterns observed for DeepSeek occur across all regions, indicating broader model-specific behavior rather than a single geopolitical effect. These patterns are likely related to differences in training data distribution and geographic coverage, although model architecture may also contribute. Distinguishing between these effects, however, is beyond the scope of this study. Furthermore, spatial resolution influences the resulting representations. For large regions such as the Sahara, coarser grids may be more time- and cost-efficient, although they reduce spatial detail. In this study, we applied a uniform grid size across all case studies to ensure consistency. In addition, LLM outputs may be sensitive to prompt formulation. In this study, we use a single standardized prompt to ensure comparability across models and regions and our findings of uneven spatial coherence are based on consistent prompting. While alternative phrasings may yield different representations,

exploring prompt sensitivity systematically remains an important question. These limitations highlight the importance of multilingual prompting, spatial resolution, and prompt formulation as interesting directions for future work.

The variation observed here points not only to geographic complexity, but also to uneven global patterns of knowledge production that shape what LLMs learn and how reliably they reproduce it; compare to (Ballatore et al., 2017). These uneven representations have direct implications for geo-alignment, as models align more closely with regions that are already cognitively stabilized within dominant data ecosystems. As a result, LLMs risk reinforcing uneven cognitive geographies, systematically privileging well-documented regions while producing less stable and more fragmented representations of others.

6 Conclusions and Future Directions

This paper examined how large language models (LLMs) represent vague cognitive regions (VCRs) and whether these representations differ across models and region types. This is important as we increasingly learn about geographic space and regions via AI agents and because the latent representation of these models influences how they behave in downstream tasks such as tourism. Using a grid-based querying workflow, we queried three state-of-the-art LLMs across four contrasting case studies: *the Alps*, *Northern–Southern California*, *the Sahara*, and *Kashmir*. By analyzing both binary classifications and graded similarity scores using established evaluation metrics, we were able to examine patterns of agreement across models for each VCR. In the case of *Northern–Southern California*, we were also able to compare the results to the ground truths, although our work here is largely concerned with the novel perspective of inter-model agreement.

The results demonstrate that LLMs can produce spatially structured representations of VCRs, but that these representations are uneven and strongly dependent on the type of region and the model used. Across the case studies, *Northern–Southern California* and the *Alps* exhibit stronger internal coherence, while the *Sahara* and *Kashmir* show greater variability and model-dependent interpretation. Interestingly, as pointed out above, Fig. 6 shows a diagonal data quality artifact (e.g., for the GPT-4o binary case) that may be caused by such a representational distortions within the model and could even be independent of the specific task, an issue that could be explored during follow-up research. It is also worth noting that the models are largely able to encode the graded structure of membership. While they show weaknesses (especially DeepSeek but also Gemini), the results indicate that LLMs can be used in the future to complement (but not replace) human-subject and social-signal studies.

The significance of these findings goes beyond the individual case studies discussed in this paper. As LLMs are increasingly used in spatial analysis, planning, and geographic information systems, it becomes crucial to understand where their regional knowledge is strong or aligned and where it is limited or biased against (Janowicz, 2023). Treating LLM outputs as definitive geographic representations without questioning their uneven coverage risks amplifying dominant narratives while overlooking or marginalizing (historically) less visible regions and perspectives.

Future research could extend this approach by incorporating multilingual prompts to examine how VCRs vary across linguistic contexts. Applying this framework to other types of VCRs, such as cultural landscapes or historical regions, would further clarify the role of language-centric AI in shaping geographic knowledge.

Declaration of Generative AI in writing

The authors declare that they have not used Generative AI tools in the preparation of this manuscript. Specifically, the AI tools were utilized solely for research data collection, as the aim of the study is to examine the potential and limitations of AI systems—particularly LLMs—in delineating VCRs, as well as to assess the unevenness of their representations across different regions of the world and the influence of embedded biases and misalignment. AI tools were not used for generating scientific content, research data, or substantive conclusions. All intellectual and creative work, including the analysis and interpretation of data, is original and has been conducted by the authors without AI assistance.

References

Ballatore, A., Graham, M., and Sen, S.: Digital hegemonies: the localness of search engine results, Annals of the American Association of Geographers, 107, 1194–1215, 2017.

Cohn, A. G. and Blackwell, R. E.: Can Large Language Models Reason about the Region Connection Calculus?, arXiv preprint arXiv:2411.19589, 2024.

Cohn, A. G., Bennett, B., Gooday, J., and Gotts, N. M.: Qualitative spatial representation and reasoning with the region connection calculus, GeoInformatica, 1, 275–316, 1997.

Copus, A. K.: From core-periphery to polycentric development: Concepts of spatial and aspatial peripherality, European planning studies, 9, 539–552, 2001.

Forgas, J. P.: The effects of prototypicality and cultural salience on perceptions of people, Journal of Research in Personality, 17, 153–173, 1983.

Gao, S., Janowicz, K., Montello, D. R., Hu, Y., Yang, J.-A., McKenzie, G., Ju, Y., Gong, L., Adams, B., and Yan, B.: A data-synthesis-driven method for detecting and extracting

vague cognitive regions, International Journal of Geographical Information Science, 31, 1245–1271, 2017.

Hecht, B. and Stephens, M.: A tale of cities: Urban biases in volunteered geographic information, in: proceedings of the international AAAI conference on web and social media, vol. 8, pp. 197–205, 2014.

Hill, L. L.: Georeferencing: The geographic associations of information, Mit Press, 2009.

Janowicz, K.: Philosophical foundations of geoai: Exploring sustainability, diversity, and bias in geoai and spatial data science, in: Handbook of geospatial artificial intelligence, pp. 26–42, CRC Press, 2023.

Janowicz, K., Hu, Y., McKenzie, G., Gao, S., Regalia, B., Mai, G., Zhu, R., Adams, B., and Taylor, K.: Moon landing or safari? a study of systematic errors and their causes in geographic linked data, in: International Conference on Geographic Information Science, pp. 275–290, Springer, 2016.

Janowicz, K., Liu, Z., Mai, G., Wang, Z., Majic, I., Fortacz, A., McKenzie, G., and Gao, S.: Whose Truth? Pluralistic Geo-Alignment for (Agentic) AI, in: Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems, pp. 799–803, 2025a.

Janowicz, K., Mai, G., Huang, W., Zhu, R., Lao, N., and Cai, L.: GeoFM: how will geo-foundation models reshape spatial data science and GeoAI?, International Journal of Geographical Information Science, 39, 1849–1865, 2025b.

Jones, C. B., Purves, R. S., Clough, P. D., and Joho, H.: Modelling vague places with knowledge from the Web, International Journal of Geographical Information Science, 22, 1045–1065, 2008.

Karimi, M. and Mesgari, M. S.: Extracting Place Functionality from Crowdsourced Textual Data Using Semantic Space Modeling, IEEE Access, 11, 129 217–129 229, 2023.

Karimi, M., Janowicz, K., Liu, Z., McKenzie, G., Abbasi, O. R., Gao, S., Adams, B., Hu, Y., and Wang, S.: Through the Eyes of the Machine: How Foundation Models Represent Vague Cognitive Regions, 2026 under review.

Lakoff, G.: Cognitive models and prototype theory, The cognitive linguistics reader, pp. 130–167, 2007.

Lenci, A. and Sahlgren, M.: Distributional semantics, Cambridge University Press, 2023.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., Chi, G., and Shi, L.: Social sensing: A new approach to understanding our socioeconomic environments, Annals of the Association of American Geographers, 105, 512–530, 2015.

Liu, Z., Janowicz, K., Majic, I., Shi, M., Fortacz, A., Karimi, M., Mai, G., and Currier, K.: Operationalizing Geographic Diversity for the Evaluation of AI-Generated Content, Transactions in GIS, 29, e70 057, 2025.

Liu, Z., Janowicz, K., Karimi, M., Shi, M., Majic, I., and Fortacz-Lazan, A.: Golden Gate Bridge, as Always? Eliciting Prototypical Places From Autoregressive Large Language Models via Category Production, Transactions in GIS, 30, e70 242, 2026.

Lloyd, R.: Learning spatial prototypes, Annals of the Association of American Geographers, 84, 418–440, 1994.

Majic, I., Wang, Z., Janowicz, K., and Karimi, M.: Spatial task-explicity matters in prompting large multimodal models for spatial planning, in: Proceedings of the 7th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, pp. 99–105, 2024.

McKenzie, G., Janowicz, K., and Keßler, C.: Uncovering spatiotemporal biases in place-based social sensing, AGILE: GIScience Series, 1, 14, 2020.

Mehler, A., Gleim, R., Gaitsch, R., Hemati, W., and Uslu, T.: From topic networks to distributed cognitive maps: Zipfian topic universes in the area of volunteered geographic information, Complexity, 2020, 4607 025, 2020.

Montello, D. R.: Regions in geography: Process and content, Foundations of geographic information science, pp. 173–189, 2003.

Montello, D. R.: Behavioral and cognitive geography, Geography, 2013.

Montello, D. R., Goodchild, M. F., Gottsegen, J., and Fohl, P.: Where's Downtown? Behavioral Methods for Determining Referents of Vague Spatial Queries, Spatial Cognition and Computation, 3, 185–204, https://doi.org/10.1080/13875868.2003.9683761, 2003.

Montello, D. R., Friedman, A., and Phillips, D. W.: Vague cognitive regions in geography and geographic information science, International Journal of Geographical Information Science, 28, 1802–1820, 2014.

Omrow, D. A.: A map of the World: Cognitive injustice and the Other, Journal of Philosophy and Culture, 8, 22–32, 2020.

Purves, R. S., Clough, P., Jones, C. B., Hall, M. H., and Murdock, V.: Geographic information retrieval: Progress and challenges in spatial search of text, Foundations and Trends® in Information Retrieval, 12, 164–318, 2018.

Razavi, A., Soltangheis, M., Arabzadeh, N., Salamat, S., Zihayat, M., and Bagheri, E.: Benchmarking prompt sensitivity in large language models, in: European Conference on Information Retrieval, pp. 303–313, Springer, 2025.

Resch, B.: People as sensors and collective sensing-contextual observations complementing geo-sensor network measurements, in: Progress in location-based services, pp. 391–406, Springer, 2013.

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., and Sculley, D.: No classification without representation: Assessing geodiversity issues in open data sets for the developing world, arXiv preprint arXiv:1711.08536, 2017.

Shypula, A., Li, S., Zhang, B., Padmakumar, V., Yin, K., and Bastani, O.: Evaluating the diversity and quality of LLM generated content, arXiv preprint arXiv:2504.12522, 2025.

Sieber, R., Brandusescu, A., Sangiambut, S., and Adu-Daako, A.: What is civic participation in artificial intelligence?, Environment and Planning B: Urban Analytics and City Science, 52, 1388–1406, 2025.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., et al.: A roadmap to pluralistic alignment, arXiv preprint arXiv:2402.05070, 2024.

Vanolo, A.: The border between core and periphery: Geographical representations of the world system, Tijdschrift voor economische en sociale geografie, 101, 26–36, 2010.

Wang, F. and Hall, G. B.: Fuzzy representation of geographical boundaries in GIS, International journal of geographical information systems, 10, 573–590, 1996.

Wijegunarathna, K. I., Stock, K., and Jones, C. B.: Digital gazetteers: review and prospects for place name knowledge bases, ACM Computing Surveys, 58, 1–39, 2025.

Zadeh, L. A.: Fuzzy Sets, Information and Control, 8, 338–353, https://doi.org/10.1016/S0019-9958(65)90241-X, 1965.