Foundation models for geospatial reasoning: assessing the capabilities of large language models in understanding geometries and topological spatial relations

Yuhan Ji, Song Gao, Ying Nie, Ivan Majić & Krzysztof Janowicz

Published online: 02 Jun 2025.

Submit your article to this journal ↗

View related articles ↗

View Crossmark data ↗

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Check for updates

Foundation models for geospatial reasoning: assessing the capabilities of large language models in understanding geometries and topological spatial relations

Yuhan Ji[a], Song Gao[a] ![ORCID], Ying Nie[a], Ivan Majić[b] and Krzysztof Janowicz[b,c]

[a]GeoDS Lab, Department of Geography, University of Wisconsin, Madison, WI, USA; [b]Department of Geography and Regional Research, University of Vienna, Austria; [c]Department of Geography, University of California, Santa Barbara, CA, USA

ABSTRACT

AI foundation models have demonstrated some capabilities for the understanding of geospatial semantics. However, applying such pre-trained models directly to geospatial datasets remains challenging due to their limited ability to represent and reason with geographical entities, specifically vector-based geometries and natural language descriptions of complex spatial relations. To address these issues, we investigate the extent to which a well-known-text (WKT) representation of geometries and their spatial relations (e.g., topological predicates) are preserved during spatial reasoning when the geospatial vector data are passed to large language models (LLMs) including GPT-3.5-turbo, GPT-4, and DeepSeek-R1-14B. Our workflow employs three distinct approaches to complete the spatial reasoning tasks for comparison, i.e., geometry embedding-based, prompt engineering-based, and everyday language-based evaluation. Our experiment results demonstrate that both the embedding-based and prompt engineering-based approaches to geospatial question-answering tasks with GPT models can achieve an accuracy of over 0.6 on average for the identification of topological spatial relations between two geometries. Among the evaluated models, GPT-4 with few-shot prompting achieved the highest performance with over 0.66 accuracy on topological spatial relation inference. Additionally, GPT-based reasoner is capable of properly comprehending inverse topological spatial relations and including an LLM-generated geometry can enhance the effectiveness for geographic entity retrieval. GPT-4 also exhibits the ability to translate certain vernacular descriptions about places into formal topological relations, and adding the geometry-type or place-type context in prompts may improve inference accuracy, but it varies by instance. The performance of these spatial reasoning tasks unveils the strengths and limitations of the current LLMs in the processing and comprehension of geospatial vector data and offers valuable insights for the refinement of LLMs with geographical knowledge towards the development of geo-foundation models capable of geospatial reasoning.

CONTACT Song Gao ✉ song.gao@wisc.edu

1. Introduction

Our interaction with Artificial Intelligence (AI) based systems is changing radically due to progress in generative Foundation Models (FM) and the conversational, natural-language-driven style of interaction with many of these models. While most prior AI models were developed with a limited range of downstream tasks in mind, foundation models aim to be general-purpose building blocks supporting a broad range of applications. Essentially, they are trained on a substantially broader set of data and, while giving up accuracy for any specific task during development, are easily fine-tuned before or during deployment. Large language models (LLMs) (Radford *et al.* 2019, Brown *et al.* 2020), such as Generative Pre-trained Transformers (GPT) (Radford *et al.* 2018, Achiam *et al.* 2023), and text-to-image models (Frolov *et al.* 2021), such as DALL-E (Ramesh *et al.* 2021), are specific types of foundation models. Most of these models are generative, i.e., they return novel, synthetic output such as natural language answers or imagery instead of providing answers by (information) retrieval as was common in prior systems, for example, from the field of expert systems. While foundation models may not inherently prescribe a specific interaction style, they can be trained or fine-tuned for various types of interactions by carefully crafting the training dataset for the intended purpose. For example, OpenAI's Codex is trained using paired code examples and comments, enabling natural language instructions to guide code generation effectively (Chen *et al.* 2021). Similarly, Contrastive language-image pre-training (CLIP) facilitates tasks like image search from paired textual descriptions. Reinforcement learning with human feedback (RLHF) is another approach that aligns model outputs with user intent, improving conversational dialog flow, adherence to prompts, and reducing harmful content. The resulting conversational style of interaction is part of their broad appeal but also causes new challenges.

Together, these breakthroughs have opened the door towards conversation-style artificial GIS analysts ('GeoMachina') (Janowicz *et al.* 2020). For instance, ChatGPT-4 can understand instructions for frequent GIS tasks like reading in a dataset (Mooney *et al.* 2023), performing simple spatial analysis steps (by generating PySAL code), or even suggesting appropriate next steps. Consequently, researchers started exploring the capabilities and limits of current AI in representing spatial data (Ji and Gao, 2023), generating maps (Zhang and Kong 2023), extracting place semantics (Hu *et al.* 2023), automating GIS operations (Li and Ning, 2023, Zhang *et al.* 2024), generating code (Gramacki *et al.* 2024), and drawing inferences from such data (Mai *et al.* 2024). Interestingly, the gaps this early research revealed are not unexpected as they have been documented as pain points of prior AI systems before (Janowicz *et al.* 2015). Prominently featured among these shortcomings is the representation of and reasoning with topological spatial relations (Cohn and Renz 2008). Even more, this is true across foundation models, that is, LLMs and text-to-image models struggle similarly. For instance, ChatGPT (OpenAI 2022) will provide a metric distance (e.g., several kilometers) when asked about the border of two neighboring countries. Similarly, DALL-E frequently fails to generate images of regions or parts described using terms such as bordering, adjacent, contained, or specific types of maps (Zhang and Kong 2023).

This is a critical insight as it implies that current work on geo-foundation models (Xie *et al.* 2023), for example, location embeddings (Mai *et al.* 2022b), may benefit the broader AI community across models.

To better understand the limitations of LLMs in handling spatial data and to develop foundation models for advancing geospatial artificial intelligence (GeoAI) (Gao *et al.* 2023), this work aims to explore the potential of representing spatial object geometries in the WKT format to enable LLMs to perform GIS operations and enhance geospatial reasoning. In this work, we present intensive experiments with well-known text (WKT) representation of geometries as inputs for LLMs and with natural language descriptions of (vague) spatial configurations. However, it is important to note that, unlike other types of data, accurate geometries (e.g. points, polylines, and polygons) and their spatial relations, as used in GIS, are not usually expressed in natural language text for such models to consume during training. Without explicitly addressing such structural deficiencies, the proposed approach is not suggested to be directly applicable in practice.

The research contributions (RC) of our work are as follows:

- RC1: We develop a workflow to assess the ability of LLMs to reason with topological spatial relations, more specifically, a subset of topological relations specified according to the Dimensionally Extended 9-Intersection Model (DE-9IM). To do so, we will compare two approaches. First, we will encode the geometries and their topological relations in an embedding space using LLMs. Second, we will use a prompt engineering method to pass WKT format of geometries directly to the LLMs.
- RC2: To test the capabilities of LLMs, we firstly utilize the WKT representation of two geometries to predict the topological spatial relation between them, and then we use one of the geometries and the topological spatial relation to predict the second geometry. To do so, we will utilize the pre-trained text embedding models and also use prompt engineering to elicit the target geometry.
- RC3: Finally, we study the ability of LLMs to extract the formalized topological spatial relations between geographic entities from vernacular descriptions (i.e., everyday language) of the relations between geographic entities, e.g., as found in administrative place descriptions from DBpedia/Wikipedia.

The remaining paper is organized as follows. We first review the literature on spatial relations, parts of the qualitative spatial reasoning and conceptual neighborhoods, large language models, and GeoAI foundation models in Section 2. We then introduce the methodology and workflow used in this research in Section 3, followed by the experiments design and dataset processing in Section 4. After that, we present the experiment results about topological spatial relation qualification and retrieval tasks using LLMs in Section 5. We further discuss the the confusion between the topological predicates with their corresponding conceptual neighborhoods in Section 6. Finally, we conclude this paper and offer insights into future work in Section 7.

2. Related work

2.1. Spatial relations

Spatial relations refer to the connection between spatial objects regarding their geometric properties (Guo 1998), which specify the location of one object related to another one (Carlson and Logan 2001) or more other objects (Majic *et al.* 2021). On the one hand, describing spatial relations in natural language is essential for understanding our surroundings in spatial cognition and navigating through space (Freksa *et al.* 1998). On the other hand, a reverse parsing process, where exact spatial relations are identified from natural language descriptions, is vital to improving the quality of information retrieval and human–computer interaction in tasks such as map reading (Head 1984), geographic question answering (Gao and Goodchild 2013, Mai *et al.* 2020, Scheider *et al.* 2021), spatial query and reasoning (Wang 2000, Du *et al.* 2005, Guo *et al.* 2022), disaster management (Wang *et al.* 2016, Cervone *et al.* 2016), driving and robotics navigation (Wallgrün *et al.* 2014, Tellex *et al.* 2011).

Typically, binary spatial relations use the format of a triplet {*subject, predicate (preposition), object*} to describe the relative positions of objects in space. In this format, the *subject* is an entity being described in relation to another entity, the *predicate (preposition)* is the descriptor between the subject and object, and the *object* is the entity that the subject is being related to in terms of position or location. For example, 'Santa Barbara is situated northwest of Los Angeles' would be expressed as {Santa Barbara, northwest of, Los Angeles} in the format of spatial relations. Even though spatial relations pervade in our daily life conversations, people tend to frequently use a limited number of predicates to describe *topological*, *directional*, and *distance* relations (Mark and Egenhofer 1994, Frank 1992). These expressions are qualitative in nature, offering approximate descriptions of an infinite range of possible spatial configurations. Nevertheless, speakers can convey complex spatial layouts by combining these basic predicates with contextual cues. For example, we might describe the locale of Santa Barbara as 'Santa Barbara is connected via U.S. Highway 101 to Los Angeles about 100 miles to the southeast.', or the position of a person as standing 'in front of the building, facing east.' The ability to combine and modify spatial predicates allows us to express a wide range of spatial relationships with a relatively small vocabulary but increases the difficulty of representing and understanding the meanings of such spatial relation descriptions for computers. The flexibility and ambiguity inherent in natural language often obscure the precise geometry of spatial arrangements, creating a disconnection between semantic interpretation and physical spatial layout. The abundance of web documents containing geographical references offers the opportunity to retrieve spatially-aware information and support qualitative spatial reasoning from natural language texts (Jones *et al.* 2004). To bridge the semantic-physical gap, prior work has focused on extracting spatial relations between named geographic entities by interpreting linguistic cues in text. These efforts include parsing grammatical and spatial semantic structures (Kordjamshidi *et al.* 2011, Loglisci *et al.* 2012, Skoumas *et al.* 2016), as well as applying supervised machine learning models trained on annotated data with spatial linguistic features (Yuan 2011, Wu *et al.* 2023a). The resulting qualitative spatial relations, enriched by contextual narratives (Wallgrün *et al.*

2015), provide a foundation for computational models that link natural language semantics to the structured representations of physical space.

2.2. Formalism of topological relations and conceptual neighborhoods

In the field of GIS, attempts have been made to formalize the conversion between quantitative computational models of spatial relations and qualitative spatial representations from human discourse (Cohn and Hazarika 2001, Chen *et al.* 2015). In Clementini *et al.* (1994), topological relations are defined as spatial relations that are preserved under such transformations as rotation, scaling, and rubber sheeting. For topological spatial relations, region connection calculus (RCC) (Randell *et al.* 1992) and point-set topology intersection models (IM), for example, 4-IM based on intersections of the *boundaries* and *interiors* of two objects (Egenhofer and Franzosa 1991), and 9-IM which also considers the *exteriors* of two objects (Egenhofer and Herring 1991), are widely used approaches. RCC-8 (Cui *et al.* 1993) is a set of eight jointly exhaustive and pairwise disjoint relations defined for regions. The basic relations include topological predicates: equal (EQ), externally connected (EC), disconnected (DC), partially overlaps (PO), tangential (TPP/TPPi) and nontangential (NTPP/NTPPi) relations, which have been shown to be cognitively adequate to be well distinguished by humans (Renz and Nebel 1998).

Point-set topology intersection models analyze whether intersections between the interiors, boundaries, and exteriors of two objects are empty or nonempty point sets. The Dimensionally Extended 9-intersection model (DE-9IM) (Clementini *et al.* 1993) further considered the dimensionality of each geometry in the intersection matrix so that the 9-IM is not a binary operation of intersects. Based on the DE-9IM model, five mutually exclusive relations are identified (Clementini and Di Felice 1996), including {disjoint, touches (meets), crosses, overlaps, within}. the Open Geospatial Consortium (OGC) later added {intersects, contains, equals} to the set for the convenience of GIS software users and included in the *GeoPandas* Python package for programmers. The recent development of RCC*-9 expands the dimensions of RCC-8 and allows for a unified framework to model topological spatial relations (Clementini and Cohn 2014, 2024). However, since DE-9IM predicates were selected for better user interaction and have been implemented by OGC, this work focuses on DE-9IM. In Mark and Egenhofer (1994), human subject testing was conducted to evaluate their model for spatial relations between lines and regions. The participants were presented with pairs of lines or regions and were asked to rate the spatial relation between them using a Likert scale that ranged from 'no relation' to 'strongly related'. The pairs of lines and regions were generated based on the 19 topologically distinct spatial relations defined in the authors' model. The human judgments were then compared to the predicted spatial relations generated by their model. The results showed that the model's predicted topological spatial relations matched the human judgments with a high degree of accuracy, indicating the effectiveness of the model in capturing human perception of topological spatial relations.

In both RCC and IM lineage, the idea of smooth transitions from one topological relation to another has been discussed early on. This means that, for example, if two
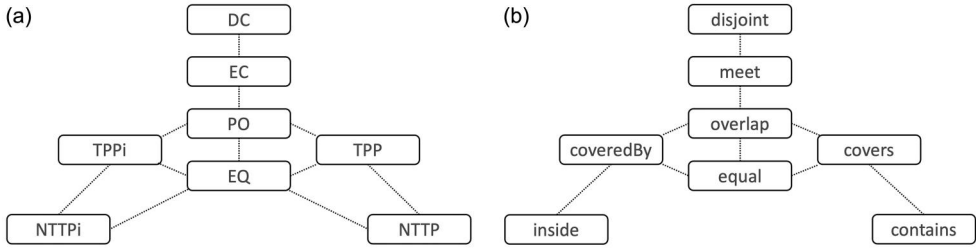
Figure 1. The conceptual neighborhood of topological relations in RCC-8 (Randell *et al.* 1992) on the left (redrawn for comparison) and 9-IM (Egenhofer and Al-Taha, 1992) on the right.

polygon objects are disjoint, they would first require a touch relationship before moving to overlap. In this sense, some relationships are more similar or closer to each other than others, and this is known as the *conceptual neighborhood* of topological relations. Figure 1 shows the neighborhood graphs using the RCC-8 (Figure 1(a)) and 9-IM (Figure 1(b)) nomenclature. Since the DE-9IM example only preserves the connection of a topological relation with its "closest" relation, the inside/contains do not connect with equal in the graph. In addition to the conceptual neighborhood, Egenhofer and Al-Taha (1992) proposed a formula for calculating the topological distance between topological relations using matrix representations, where the smaller distance means more similar between the two topological relations. We adopt the topological distance for evaluation later in this paper to provide a more nuanced perspective on whether LLMs' differentiation of topological relations aligns with human perception.

2.3. Large language models and GeoAI foundation models

The launch of ChatGPT by OpenAI (2022) marked a significant turning point, drawing widespread interest in Large Language Models (LLMs) and conversational AI from the public. Language-based foundation models boast an impressive range of parameters, from 110 million in BERT (Devlin *et al.* 2018) to 1.5 billion in GPT-2 (Radford *et al.* 2019), and up to 137 billion in LaMDA (Google's Bard) (Thoppilan *et al.* 2022) and 175 billion in GPT-3 (Brown *et al.* 2020), demonstrating a significant variation in network architectures, scale, and purposes. Despite these differences, they share a common achievement: they have acquired a sophisticated understanding of language patterns and semantics, setting new performance standards in natural language processing tasks. Other types of foundation models include vision-based (e.g. vision transformer– ViT (Dosovitskiy *et al.* 2020) and segment anything model–SAM (Kirillov *et al.* 2023)) and vision-language multimodal foundation models (e.g., Flamingo with 80 billion parameters (Alayrac *et al.* 2022) and GPT-4 with over 1 trillion parameters (Achiam *et al.* 2023)). These pre-trained foundation models have been applied directly or transferred to a wide range of cross-domain tasks after fine-tuning or few-shot/zero-shot learning, for example, education (Kasneci *et al.* 2023), healthcare (Yang *et al.* 2022), transportation (Zheng *et al.* 2023), etc.

These foundation models have been trained on large-scale datasets that also contain geographical knowledge such as descriptions of locations and places in textual

documents as well as spatial elements in maps, geo-referenced photos, and satellite imagery. Recently, researchers and institutions have begun the early exploration of integrating foundation models into GeoAI research and education. For example, Mai et al. (2024) found that task-agnostic LLMs have the capability to surpass fully supervised deep learning models designed for specific tasks in understanding geospatial semantics, including toponym recognition, health data time-series forecasting, urban function, and scene classifications. Hu et al. (2023) fused a few geo-knowledge examples into GPT models to improve the extraction of location descriptions from disaster-related social media messages. Manvi et al. (2023) found that geospatial knowledge can be effectively extracted from LLMs with auxiliary map data from OpenStreetMap. Additionally, spatial-context-aware prompts with pre-trained visual-language models can improve the accuracy of urban land use classification and urban function inference (Wu et al. 2023b, Huang et al. 2024), In GIS, evaluations have been conducted to assess the qualitative spatial reasoning capabilities of LLMs in identifying and reasoning spatial relations using symbolic representations of spatial objects, such as RCC-8 (Cohn 2023, Cohn and Blackwell 2024a) and cardinal directions (Cohn and Blackwell, 2024b). While LLMs perceive the spatial structure through sequences of textual input (Yamada et al. 2023) and leverage commonsense reasoning during their inference process (Cohn and Hernandez-Orallo 2023), they also demonstrate human-like misconceptions and distortions about space (Fulman et al. 2024). Several studies (Mai et al. 2022a, Fernandez and Dube 2023, Tucker 2024) have proposed integrating vector data as a backbone for spatial reasoning. GPT-4 has shown the capability to generate coordinates for outlines of countries, rivers, lakes, and continents that approximate their actual geographic locations (Das 2023). In Ji and Gao (2023), LLM-generated embeddings can preserve geometry types and some coordinate information in the WKT representation of geometries. However, performing qualitative spatial reasoning and executing spatial tasks from implicit textual descriptions involving coordinates remains a significant challenge (Majic et al. 2024). In addition, geospatial analysis workflows and operations can be automated when combing LLMs with spatial analysis tools (Li and Ning 2023, Zhang et al. 2024). ChatGPT can even achieve a promising grade when taking an introduction to GIS examination (Mooney et al. 2023). In the field of Cartography, Tao and Xu (2023) explored the use of ChatGPT-4 for creating thematic maps and mental maps with appropriate prompts. However, Zhang and Kong (2023) pointed out the ethical concerns on AI-generated maps' inaccuracies, misleading information, unanticipated features, and reproducibility. In August 2023, NASA and IBM released their GeoAI Foundation Model–Prithvi, which was trained on NASA's Earth Observation remote sensing imagery (i.e. the harmonized Landsat and Sentinel-2 satellite dataset) (Jakubik et al. 2023) and has been found to have a good performance and transferability on flood inundation mapping (Li et al. 2023). Alongside such remarkable achievements, there are concerns that need to be addressed together with the development and advancement of foundation models for GeoAI and geosciences (i.e., Geo-Foundation Models), such as geographical bias, diversity, spatial heterogeneity, limited human annotations, sustainability, privacy and security risks (Janowicz 2023, Xie et al. 2023, Rao et al. 2023, Hu et al. 2024).
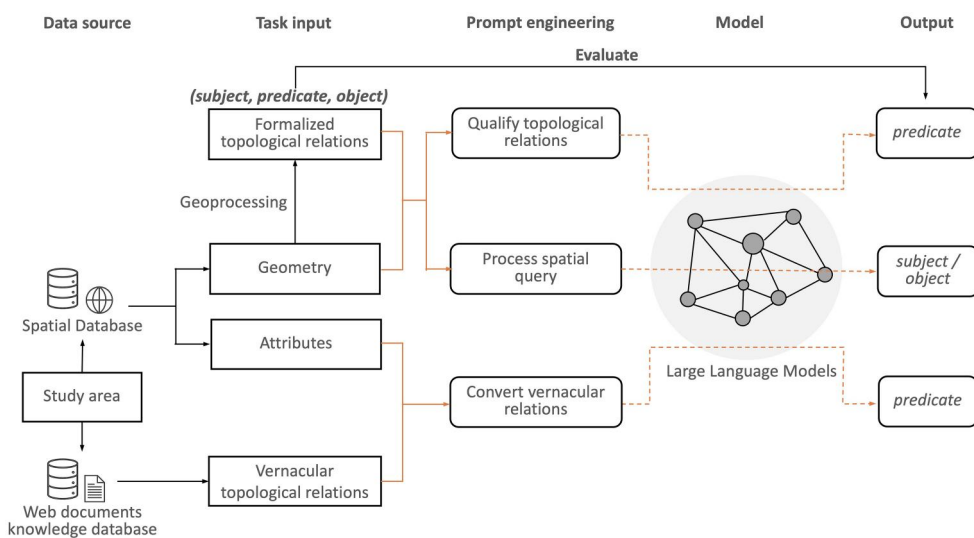
3. Methodology

3.1. Preliminaries and workflow

This research focuses on assessing the ability of LLMs to represent textual descriptions of geometries and understand topological spatial relations between geometric objects. The overall framework of this research is shown in Figure 2. Given a study area, we first retrieve spatial objects from both a spatial database and a textual description about places from a Web document knowledge database (e.g., DBpedia/Wikipedia). When the documents contain vernacular description of topological relations between two places, formalized DE-9IM topological spatial relations will be extracted from the spatial footprints (geometries) in the format of triplets as ground truth. The obtained geometric, attributive, and relational information is used as input for downstream tasks (e.g., qualify topological relations, process spatial query, and convert vernacular description of relations), where task-specific prompts are designed accordingly. The task output from the LLMs is then compared to the ground truth topological relation triplets to evaluate their ability to encode and reason about geometries and topological spatial relations. The following subsections will further provide details on each evaluation task and the corresponding workflow. The definitions and notations used in this paper are listed in Table 1.

3.2. Determining topological spatial relations

In the original work of DE-9IM (Clementini *et al.* 1993), the five defined topological predicates {disjoint, touches (meets), crosses, overlaps, within} were considered mutually exclusive. However, the statement no longer holds with the introduction of 'contains' and 'equals' to the set by the OGC standard. Therefore, to ensure the uniqueness of the topological spatial relations between two objects, we interpret 'within' as 'within (but not equals)' and 'contains' as 'contains (not equals)' in this

Figure 2. Overview of the workflow in this research.

Table 1. Notations.

Notation	Description
A/B	The objectID of spatial objects A or B
g_A	The geometry of A that can be processed in GIS tools
$GeomType(A)$	The geometry type of A, (e.g. Point, LineString, and Polygon when g_A is a simple feature)
$\overset{\circ}{g_A}$	The interior of g_A
$dim(g)$	The dimension of a geometry g. {\tf="OT5c7b8c86_I"{dim}}(g)={\mathopen\sleft \\{{\matrix{−\hfill g=∅\hfill \cr {\curr 0}\hfill g{\rm contains at least one Point without Linestrings or Polygons}\hfill \cr {\curr 1}\hfill g{\rm contains at least one Linestrings without Polygons}\hfill \cr {\curr 2}\hfill g{\rm contains at least one Polygon}\hfill \cr }
$WKT(A)$	The WKT format of g_A
$Enc(A)$	The location encoding of g_A using an LLM model to encode $WKT(A)$
R	The set of predicates to represent the topological spatial relations in this research, i.e. {equals, disjoint, crosses, touches, contains, within, overlaps}, as defined by OGC and implemented in GeoPandas.
rel	A predicate that can be used to represent the topological spatial relation, $rel \in R$
$Rel(A, B)$	The topological spatial relation between the subject A and the object B
$[Enc(A); Enc(B)]$	The concatenation of the embeddings of A and B
$D(rel_1, rel_2)$	The topological distance between two relations rel_1 and rel_2 on the conceptual neighborhood graph (Egenhofer and Al-Taha, 1992)
$S_A(rel, B)$	The relevancy score of a retrieved subject A given the reference object B and the desired topological spatial relation rel.
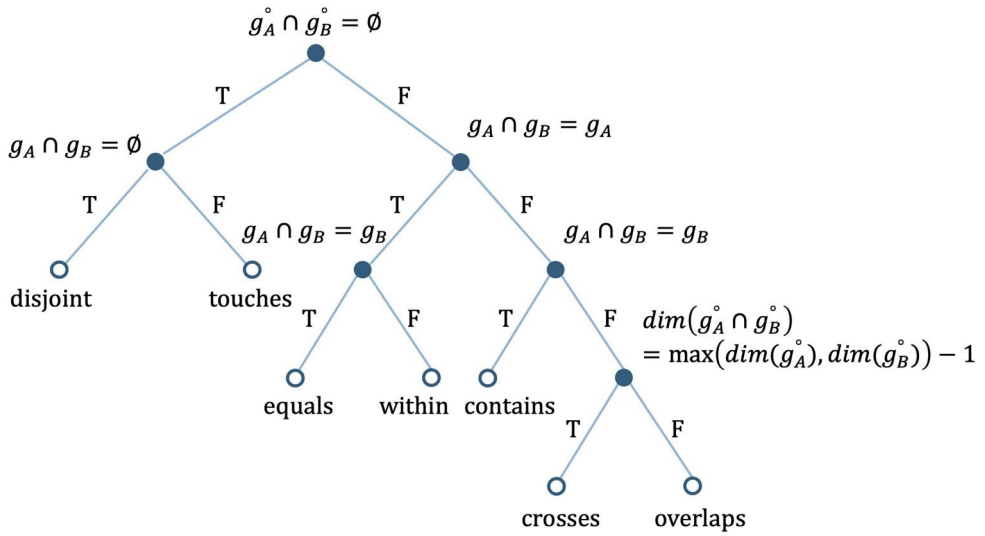
Figure 3. The decision tree for the topological spatial relations.

work. Accordingly, we modify the decision tree in Clementini *et al*. (1993) to do the reasoning about the topological relations between two spatial objects, as illustrated in Figure 3. Based on the decision process, the topological spatial relations do not apply to every combination of geometry types. The definitions and possible geometry type combinations of the seven predicates used in this research are listed in Table 2. Several visual examples of the topological spatial relations between two geometries can be seen in Figure 10.

Table 2. The named topological spatial predicates with the 9-intersection Boolean code (T: true; F: false; *: free value) and corresponding applicable geometry type combinations of a predicate.

Predicate with 9-intersection code	Geometry Type Combination
equals: T*F**FFF*	Point/Point, LineString/LineString, Polygon/Polygon
within: T*F**F***	Point/LineString, Point/Polygon, LineString/LineString, LineString/Polygon, Polygon/Polygon.
contains: T*****FF*	LineString/Point, LineString/LineString, Polygon/Point, Polygon/LineString, Polygon/Polygon.
overlaps: T*T***T**	LineString/LineString, Polygon/Polygon
touches: FT******* or F***T****	Point/LineString, Point/Polygon, LineString/Point, LineString/LineString, LineString/Polygon, Polygon/Point, Polygon/LineString, Polygon/Polygon.
crosses: T*T******	LineString/LineString, LineString/Polygon, Polygon/LineString
disjoint: FF*FF****	Applicable to ALL

3.3. Representing geospatial data as text

An embedding is a multi-dimensional numeric vector representation of objects to capture the complex patterns and relationships in the data. While researchers have explored different approaches to embed geometries using spatially explicit models (Yan *et al.* 2017, Mai *et al.* 2022b, Zhu *et al.* 2022), this study presents a novel perspective by hypothesizing that LLMs can effectively encode the WKT format of geospatial vector data (points, polylines and polygons) and preserve crucial geometric information. We adopt sentence embedding models (Logeswaran and Lee 2018, Reimers and Gurevych 2019, Neelakantan *et al.* 2022) to generate neural embeddings of the input geometry WKT strings, which allows for the comparison and retrieval of spatial information through the semantic search (Muennighoff 2022, Hu *et al.* 2015).

3.4. Evaluation tasks

3.4.1. Topological spatial relation qualification

In Wolter and Wallgrün (2012), spatial relation qualification is defined as the process of inferring qualitative spatial relations from quantitative data. The first task aims to leverage LLMs to *classify* the topological spatial relationships between subject entity A and object entity B into one of seven predefined topological predicates (in Section 3.2), combined with their geometry types. The input and output of **Task 1** are described as follows:

Input: The input for this task is the WKT representations of geometries A and B, denoted as WKT(A) and WKT(B). Example inputs:

- *WKT(A)*: POINT (-89.3551 43.123)
- *WKT(B)*: POLYGON ((-89.3552 43.124, -89.355 43.124, -89.355 43.122, -89.3552 43.122, -89.3552 43.124))

Output: The output is a tuple that describes the topological spatial relationship between the two geometries, in the format of
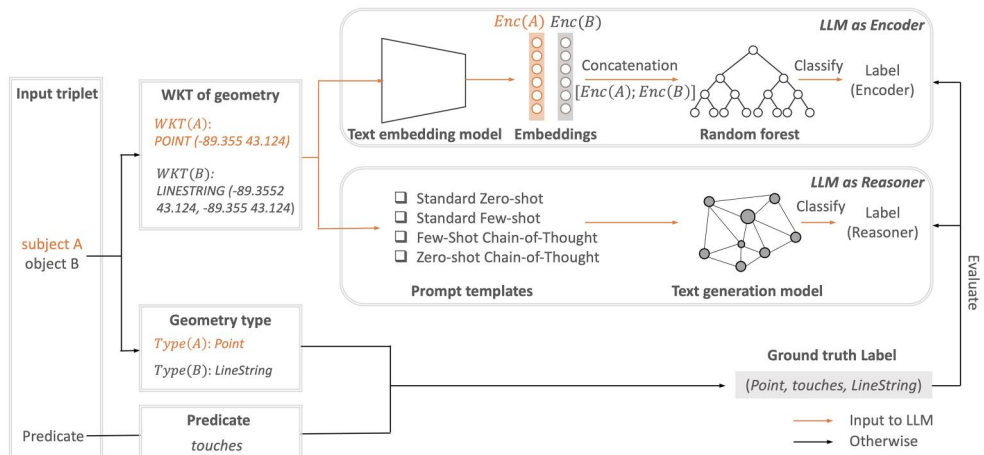
Figure 4. The workflow for the topological spatial relation qualification task.

$(GeomType(A), predicate, GeomType(B))$. Given the example inputs, the expected output of a correct classification would be:

- (Point, within, Polygon)

Use Case: Task 1 can be relevant to linking the geometries that occur in the same spatial context. For example, suppose one document already provides location, geometry and attribute information on housing resources and public transportation facilities. In that case, the LLM may directly use geographic information and other contexts to suggest affordable and accessible housing by public transportation facilities.

The workflow of task 1 is shown in Figure 4. Given an input triplet that describes the topological spatial relation between subject A and object B, that is, (subject, predicate, object), we first retrieve the WKT strings, and Geometry types of A and B. We then adopt two approaches (embedding-based and prompt-based) to perform the task, utilizing an appropriate LLM, to function as either a text encoder or a reasoner. For encoding, a pre-trained sentence embedding model generates the embeddings of the geometries of A and B. The embeddings are concatenated as the input for a random forest classifier (Breiman 2001). For reasoning, a more powerful generative model, such as GPT-4 and DeepSeek-R1, are employed to perform the task defined in the prompt. Four prompt engineering techniques are adopted to potentially guide the LLMs towards producing a more valid and accurate output of the topological spatial relation, including standard zero-shot learning, standard few-shot learning (Radford *et al*. 2019), few-shot chain-of-thought (CoT) prompting (Wei *et al*. 2022), and zero-shot COT prompting (Kojima *et al*. 2022). In few-shot CoT, we follow the decision tree in Figure 3 to generate the intermediate steps to determine the topological spatial relations as examples. While the identification of topological spatial relations might appear straightforward to the human brain, it involves multi-step reasoning. The DE-9IM framework (Clementini *et al*. 1993) decomposes the problem into intersections of the boundaries, interiors, and exteriors of two geographic entities, with dimensional requirements that map to topological predicates intuitive to users. We hypothesized

Instruction

> You will be given the WKT format of geometries given the subject A and reference object B.
> Classify their spatial relations into one of the following predicates: contains, within, touches, equals,
> crosses, disjoint, overlaps.The predicates are defined by DE-9IM and Open Geospatial Consortium. If A equals to
> B, there is no need to return 'within' or 'contains'.
>
> - (Standard setting)Return in the format (Geometry Type A, [PREDICATE], Geometry Type B), and nothing else.
> Geometry types are Point, LineString, Polygon.
> - (CoT setting) MAKE SURE the output ends with the 'ANSWER: (Geometry Type A, [Predicate], Geometry Type B)
> - MAKE SURE the [PREDICATE] is one of the seven predicates stated.
> - MAKE SURE the [PREDICATE] satisfies the the dimension requirements defined by DE-9IM and OGC given Geometry
> Type A and Geometry Type B.
> - (CoT setting) MAKE SURE the rationale is complete, professional, consistent, coherent, and accurate to derive
> the answer for the classification task.

(a) Zero-shot

Q:
Geometry A: POINT (-89.355 43.124)
Geometry B: LINESTRING (-89.3552 43.124, -89.355 43.124)
A:
......
(Output) (Point, touches, LineString)

(b) Few-shot

[EXAMPLES]
Q:
Geometry A: POINT (-89.463 43.055)
Geometry B: LINESTRING (-89.463 43.055, -89.463 43.055)
A:
(Point, touches, LineString)
…
[END EXAMPLES]
Q:
Geometry A: POINT (-89.355 43.124)
Geometry B: LINESTRING (-89.3552 43.124, -89.355 43.124)
A:
......
(Output) (Point, touches, LineString)

(c) Few-shot CoT

[EXAMPLES]
Q:
…
A:
1. Does the Interior of A intersect the Interior of B? 2.
Does the Boundary of A intersect the Boundary of B?
…
[END EXAMPLE]

Q:
Geometry A: POINT (-89.355 43.124)
Geometry B: LINESTRING (-89.3552 43.124, -89.355 43.124)
A:
......
(Output)
1. Does the Interior of A intersect the Interior of B? No
2. Does the Boundary of A intersect the Boundary of B? Yes
ANSWER: (Point, touches, LineString)

(d) Zero-shot CoT

Q:
Geometry A: POINT (-89.355 43.124)
Geometry B: LINESTRING (-89.3552 43.124, -89.355 43.124)
A:
Let's think step by step!
(Output)
1. Determine the geometry types: …
2. Check if A equals B: …
…
6. Provide the answer:
ANSWER: (Point, equals, LineString)

Figure 5. Topological spatial relation qualification example inputs and outputs with different prompt techniques.

that few-shot prompting and explicit reasoning steps, guided by CoT, could improve the model's performance on this qualification task. The example inputs and outputs of the topological spatial relation qualification task using the above-mentioned different prompt engineering techniques are illustrated in Figure 5.

The metrics for evaluating the topological spatial relation qualification task are as follows.

1. Validity
   a. Valid format of the output: LLMs should follow the instructions to use the given format of the output in (Geometry Type A, Predicate, Geometry Type B).
   b. Valid geometry types: LLMs should preserve the Geometry Type A and Geometry Type B from the given WKT format of geometries.
   c. Valid combinations of geometry types for the topological predicates as shown in Table 2.
2. Accuracy

For valid outputs, we can compute the accuracy when the output topological spatial predicate matches the ground truth.

3. Topological distance in the conceptual neighborhood graph

In this work, we use the shortest path distance between two topological predicates in the conceptual neighborhood graph (Figure 1), where the distance of each edge equals 1. Since Figure 1 was originally proposed for region-to-region (Polygon/

Polygon) relations in 9IM, we mapped their topological predicates to the seven DE-9IM predicates that we use. For other geometry type combinations, we refer to Mark and Egenhofer (1994) and Reis *et al*. (2008) to extract the conceptual neighborhood graphs. With the topological distance measurement, we can further analyze which pairs of predicates can easily confuse LLMs and whether such confusion is directed, by comparing the false-negative and false-positive results.

3.4.2. Spatial query processing

In Sack and Urrutia (1999), a generic spatial query is defined as the retrieval of subjects from a set of candidate geometric entities that are in a specific relation *rel* with the query object *B* on the basis of geometric information only. Our second task aims to evaluate whether LLMs can jointly encode a topological relation and one geometry to capture the feasible geometries that meet the query requirement. The input and output of the **Task 2** are as follows:

Input: The input for this task is the WKT representations of geometry *B*, denoted as $WKT(B)$, and a given predicate of topological spatial relations *rel*. Example input:

- Predicate: within
- $WKT(B)$: POLYGON ((-89.3552 43.124, -89.355 43.124, -89.355 43.122, -89.3552 43.122, -89.3552 43.124))

Output: The output is the identifier of a subject entity *A* whose topological spatial relationship with B is described by the predicate.

Use Case: Task 2 is valuable for retrieving textual reports that involve locations, spatial layouts, and geospatial semantics. This analysis relies on accurate queries using spatial predicates. For instance, it would be beneficial to analyze the selection of a nearby competitor's site report when considering opening a business in the same neighborhood.

The evaluation workflow of Task 2 is shown in Figure 6. Given a query specifying the topological spatial relation *rel* with the query object *B* ($WKT(B)$), we first retrieve the subjects from the study area spatial database as ground truth. We format the query as the input to an LLM using two approaches. First, this query can be directly formulated as a sentence, such as "Retrieve a geometry within POLYGON ((-89.3552 43.124, -89.355 43.124, -89.355 43.122, -89.3552 43.122, -89.3552 43.124)." Alternatively, synthetic geometries can be created using a generative model to expand the query, connecting the query with the search space. The (expanded) query text is inputted into the sentence embedding model to generate the embeddings. The geometries in WKT format for spatial entities are also processed by the same embedding model, to generate the embeddings $(Enc(g_1), Enc(g_2), \ldots, Enc(g_n))$. The most relevant subject geometries are retrieved based on the cosine similarity of their geometry embeddings and the query embeddings. We perform the evaluation as a link prediction task in the "filtered" setting (Bordes *et al*. 2013), which excludes other subjects related to B by the topological predicate *rel* from the database and concentrates on the retrieval of the subject in the triplet. This approach addresses the biases introduced by the significant difference in the number of spatially related subjects across
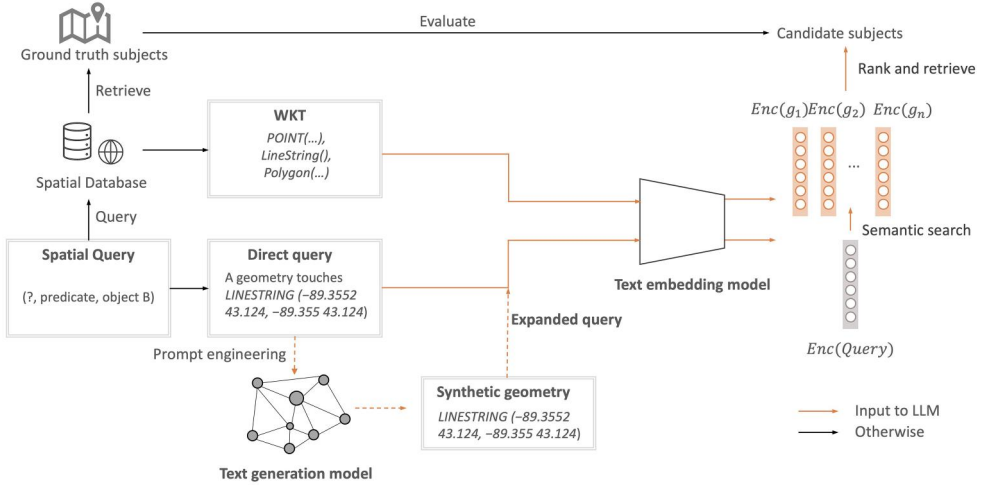
Figure 6. The workflow for the spatial query processing task.

predicates and the objects. Finally, the retrieved subjects are evaluated by their actual topological spatial relation to the reference subject.

In the following, we introduce how to format the direct query and the expanded query with LLM-generated geometries in detail:

1. Direct Query

Given the WKT format of the geometry of a known reference object (e.g. LINESTRING (-89.4534 43.035, -89.454 43.0351)) and a designated topological spatial relation (e.g. 'crosses'), the query formulation is as follows: 'Retrieve a geometry that crosses the LINESTRING (-89.4534 43.035, -89.454 43.0351).' If the search focuses on a specific geometry type, the query can be articulated as 'Retrieve a LINESTRING geometry that crosses the LINESTRING (-89.4534 43.035, -89.454 43.0351).'

2. Expanded query with LLM-generated geometries

In Carpineto and Romano (2012) and Hu *et al.* (2015), (geospatial) query expansion is used to augment the user's original query with new features (e.g., geographic or thematic characteristics) that share a similar meaning as the expected output of semantic search. The method can address the lack of semantic similarity between the query and the desired geometry. We extend the *Query2Doc* model (Wang *et al.* 2023) to the spatial query expansion, where we leverage an LLM to generate a synthetic geometry that can possibly be the response to the query. The prompt template for the generation of geometric objects or subjects is listed in Figure 7. We adopt the following prompting approaches for geometry generation.

- Zero-shot: LLMs generate geometries directly from the given spatial query.
- Zero-shot + Self-check: LLMs are asked to verify the spatial relations before generating the output.

```
Task Description:

(Subject retrieval)You will receive a spatial query containing the spatial relationship PREDICATE with the
reference object B (WKT(B)).Your task is to generate a plausible geometry A in WKT format that could be the
result of the query. The PREDICATE is defined according to the DE-9IM and Open Geospatial Consortium standards.

(Object retrieval) You will receive a spatial query describing the subject A (WKT(A) and a spatial relationship
PREDICATE. Your task is to generate a plausible geometry B in WKT format that could be the result of the query.
The PREDICATE is defined according to the DE-9IM and Open Geospatial Consortium standards.

Instruction:

- Provide only the WKT(A)/WKT(B) as the output, and nothing else.
- MAKE SURE the WKT(A)/WKT(B) represents a valid Point, LineString, or Polygon.
- MAKE SURE the A exhibits the spatial relationship PREDICATE with B based on the WKT(A).
- MAKE SURE the PREDICATE satisfies the the dimension requirements defined by DE-9IM and OGC given Geometry Type
A and Geometry Type B.

Context:

[EXAMPLES]

    …

[END EXAMPLES]

Query:

Query: {query}
Desired output:
```

Figure 7. Prompt template used for geometry generation in the spatial query processing task.

- Few-shot: Give a few pairs of example queries and corresponding subjects while maintaining spatial relations and object geometry type.
- Few-shot + Negative examples: Apart from the plausible examples, we also include the negative examples that are not the correct responses for the given query. The examples are formatted as "Retrieve a Geometry Type which … Good Response: … Bad Response: … "

We further incorporate the LLM-generated geometries into the spatial queries to assess the usefulness of the expanded queries.

The evaluation includes two parts: First, LLMs' ability to generate valid synthetic geometries as a basis for the expanded queries. Second, query processing performance through semantic search using both direct queries and expanded queries.

1. Validity of the LLM-generated geometries
    a. Valid WKT format of geometries to be successfully parsed by the GIS tool for creating geometry instances.
    b. Correct topological spatial relation *Rel* with the query object *B*.
2. Mean Reciprocal Rank (MRR) and Hits@K of the retrieval performance

We employ two commonly used metrics in geographic information retrieval, Mean Reciprocal Rank (MRR) (Yan *et al.* 2017) and Hits@K, in the 'filtered' setting (Bordes *et al.* 2013). A desirable model is characterized by higher MRR and Hits@K values.

3.4.3. Conversion of vernacular relation descriptions

In Chen *et al.* (2018), a vernacular description of spatial relations between places is an alternative to formal spatial relations in metric space, which occurs in everyday communication in a flexible format of a preposition, verb, phrase, or even implicit text description. The third task aims to evaluate how LLMs can convert the vernacular description (i.e. everyday language) of a topological relationship between two

geographic entities into one of seven predefined topological predicates based on the given context. This task is inspired by LLM's commonsense model of the world and naive geographical knowledge about space, and the domain-specific knowledge of the formalism in calculus to bridge the gap between the vernacular(narrative) descriptions and the formalized topological predicates. For example, ChatGPT is able to provide the rationale behind the statement 'When an island is in the middle of a lake, the island touches the lake if the lake is considered as a separate region (not fully containing the island)' by identifying the lake in this scenario is a double-border object using commonsense knowledge reasoning (rather than precise geometries). It then maps this understanding to the 'touches' topological relation, applying expertise in the GIS domain. The input and output of the **Task 3** are as follows:

Input: The input for this task includes a sentence that describes the topological relationship between two places in everyday language, along with the contextual information of the two places. Example input:

- Sentence: Place *A* is home to Place *B*
- Context: Place *A* is a city. Place *B* is a university

Output: The output will rephrase the sentence using the formalized topological predicates.

- Answer: Place *A* contains Place *B*

Use Case: Parsing vernacular descriptions of spatial relations between places into formal ones can better support the users interacting in natural language and the use of spatial analysis tools that rely on formal topological relations. For example, interpreting vague terms in travel reports to determine if cross-border human behavior exists and interpreting the territorial changes and alignment of contemporary boundaries in the historical context.

We adopt the workflow in Figure 8 to evaluate the capability of LLMs in task 3. The workflow begins with collecting geographic entities from a Web document knowledge database (DBpedia, structured knowledge based on Wikipedia), where named entity recognition is used to extract place names and vernacular spatial relation descriptions.
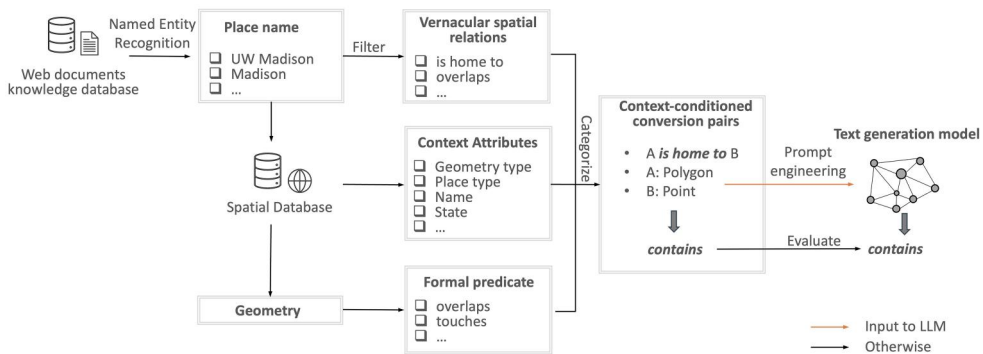
Figure 8. The workflow for the vernacular relations conversion task.

Table 3. The textual description from DBPedia and topological predicate conversion examples.

Description	Context	Predicate	Example
bordered by	No context	touches	Glendora is bordered by Azusa. → A is bordered by B.
along	Geometry type	crosses	Luling is along the San Marcos River. → A is along B. A is Polygon, and B is LineString.
located on	Place type	crosses	Located on Interstate 10, Weimar is a small community. → A is located on B. A is city, B is highway.
on the shore of	Place name	overlaps	Racine is located on the shore of Lake Michigan. → A is on the shore of B. A is Racine in Wisconsin, B is lake Michigan.

These place names, such as 'UW-Madison' and 'Madison, Wisconsin' are then used to retrieve relevant geographic data and corresponding attributes from a spatial database to provide context such as geometry type and place type, with their topological spatial relations identified through GIS tools. The collected spatial relations between two places are formatted as 'A {vernacular topological relation} B' (e.g. A is home to B) for evaluation, where A and B are symbolic placeholders representing two places. The specific locations in geometries are not disclosed, allowing for a generalized discussion of topological spatial relations without actual geographic context. The context will be provided at the end of the text input to support in-context reasoning. The context evaluated in our experiments is in Table 3. The prompts are crafted with the template as shown in Figure 9 and fed into an LLM (e.g. GPT-4) to convert vernacular descriptions to topological spatial relations. We run the model multiple times to identify the possible converted topological predicates and the preference of an LLM. The output topological predicates are then compared with the ground truth predicates calculated by the GIS tool for evaluation. We also compare the performance when no contextual information is provided. This workflow allows us to evaluate the effectiveness of LLMs for analyzing informal topological relations between two entities and to assess the impact of contextual information on performance.

The evaluation metrics for the vernacular relation conversion task are as follows.

1. Frequency: The count of correctly returned predicates across all experiments.
2. Accuracy: The ratio of the frequency of correctly returned predicates to the total number of generated outputs for each conversion pair.
3. Entropy: The information entropy (Shannon 1948) of the returned predicates assesses the level of randomness in converting vernacular descriptions into topological predicates. Smaller entropy values indicate a higher likelihood of certain predicates being preferred over others. The metric is computed as:

$$H = -\sum_{rel \in R} p_{rel} \log (p_{rel}),$$

where p_{rel} represents the probability of a specific topological predicate rel appearing in the outputs for the given context-conditioned conversion pair.

```
Task Description:
The given PREDICATES: contain, within, touches, crosses, disjoint, overlaps, equals.
Given a sentence that include a vernacular term: please give the corresponding PREDICATES.

Instruction:

- Indicate spatial relation from A to B only.
- The vernacular spatial relation term should consider its typical meanings of the terms and the
general geographical relationship it represent. Also pay attention to any other context that you
can get from the sentence.
- MAKE SURE the [PREDICATE] satisfies the the dimension requirements defined by DE-9IM and OGC
given Geometry
Type A and Geometry Type B.
- MAKE SURE that the output includes all plausible predicates and excludes any that are
impossible or irrelevant in the given context.

Output Format:
- Analysis: Provide a detailed, professional, and coherent examination of the spatial relation
from A to B.
- Answer: A [PREDICATE1/PREDICATE2...] B. For example: A [overlaps/within] B

Query:

Query: given the sentence: A in B.
```

Figure 9. Prompt template used in the vernacular relation conversion task.

4. Data and experiments

4.1. Data processing

4.1.1. Extracting topological spatial relations from spatial database

We construct real-world multi-sourced geospatial datasets for our study. The study area for **Task 1 and Task 2** is the city of Madison, Wisconsin, United States. The following datasets are collected.

- OpenStreetMap road network data (including links and intersections) using *OSMnx*.[1]
- Points of interest (POIs) categorized by *SLIPO*.[2]
- Land parcels from *Wisconsin Statewide Parcel Map Initiative*.[3]
- Census block groups from *U.S. Census Bureau*.[4]

Our evaluation tasks focus on the spatial objects with *Point, LineString*, and *Polygon* geometry types, assessing their topological spatial relations. All the computations are performed by using the *GeoPandas* package in Python.

Task 1 and Task 2 share the same dataset of triplets. For each combination of {geometry type *A*, predicate, geometry type *B*}, we obtain 200 triplets. Among these, 160 are allocated for training the random forest classification model, while the remaining 40 triplets are reserved for evaluation. Additionally, we set aside 25 extra triplets as candidate examples to facilitate few-shot learning. Due to the imbalanced distribution of topological spatial relations within the real-world dataset, we employ multiple strategies for sampling a sufficient number of triplets for fair comparisons:

1. For topological spatial relations including 'within', 'contains', 'overlaps', 'touch' and 'crosses', we opt to select a subset of spatial objects and conduct spatial joins to obtain the required triplets.

2. Regarding the 'equals' relationship. we manually created the equivalent spatial entities to preserve the topological spatial relations while making direct identification from geometry coordinate matching challenging.
    a. For *Point*, include only points with identical coordinates.
    b. For *LineString*, interpolate an additional 10% of points along the lines, ensuring that the added points did not alter the original shape.
    c. For *Polygon*, loop the origin point and interpolate additional points along the boundaries.
3. We restrict the occurrence of 'disjoint' to cases where the subject geometry does not touch or overlap but lies within a smaller buffer of the objects (i.e. nearby entities), to avoid easy identification when two spatial entities are far apart, thus enhancing the evaluation on the differentiation of topological predicates.

In **Task 2**, we further exclude the 'disjoint' relations since most real-world geographic entities are disjoint from each other, yielding $40 \times 26 = 1040$ triplets for retrieving the subject or object geographic entities.

4.1.2. Topological spatial relations from DBpedia/Wikipedia

For **Task 3**, we have gathered a total of 1078 unique triplets based on the recognized geographic entities from DBpedia/Wikipedia documents, which we combine with everyday descriptions of topological spatial relationships. We then utilize this extracted data to evaluate GPT-4's capabilities in task 3 as described in Section 3.4.3.

Specifically, we downloaded and refined place descriptions within the States of Wisconsin, Texas, and California from the knowledge base *DBpedia*[5], which is the linked data format of Wikipedia and has been previously used in place name disambiguation task (Hu *et al.* 2014). The data extraction and processing steps are structured as follows:

1. Named entity recognition: From each administrative region's abstract "dbo:abstract"), we extract all place names that can be found in OpenStreetMap, forming the basis for subsequent topological spatial relation identification.
2. Textual spatial relation extraction: For each pair of place names within a DBpedia abstract, we use GPT-4 to extract topological spatial relation terms found between the entities in the text. When hierarchical place relationships are described, our approach only captures direct relations between a subject and each individual object, omitting implicit transitive relations among the objects themselves. For instance, from the sentence 'a city A in a County B, State C,' we extract (A, in, B) but skip (A, in, C) and (B, in, C).
3. Manual verification: We manually review all the extracted spatial relation descriptions to ensure that they indicate topological relations and that the use of the two place names as subjects or objects in the sentence is semantically correct.
4. Description unification: The text descriptions on DBpedia are standardized for consistency. For example, phrases like 'is home to', 'home to' or 'home of', are unified as 'is home to'.

5. Context-conditioned conversion pairs extraction: We identify how vernacular descriptions depend on the following context to convert them to formal topological predicates.

   a. Invariant to context: If a vernacular description consistently corresponds to the same topological predicate, we create a context-conditioned conversion pair (description, predicate, N/A).

   For descriptions that can be converted to multiple formal topological predicates, we associate them with specific contexts for one-on-one conversion.

   b. Place types as context: If grouping by description, place type *A* and place type *B*, results in a unique topological predicate, we create the pair (description, predicate, place type *A*/place type *B*). Place types are extracted from OpenStreetMap data tags.

   c. Geometry types as context: If grouping by description, geometry type *A* and geometry type *B*, results in a unique predicate, we create the pair (description, predicate, geometry type *A*/geometry type *B*).

   d. Place names as context: Each pair of places can have a unique topological relationship. We create the pair (description, predicate, place name *A*/place name *B*), assuming the LLMs have some knowledge about place names.

It is possible that more than one context can assist with one-on-one mapping from a vernacular description to a formal topological predicate. We may retain multiple contexts to compare their effectiveness. For example, the conversion between 'is bordered by' and 'touches' can be identified using place types (is bordered by, town/city, touches), geometry types (is bordered by, Polygon/MultiPolygon, touches) and place names (is bordered by, Aliso Viejo, California/Laguna Beach, California, touches).

6. Data filtering: Only frequently observed context-conditioned conversion pairs are retained for evaluation.

   a. In the cases of invariant to context, place type as context, and geometry types as context, we retain pairs that occur at least 5 times for evaluation.

   b. In the case of place names as context, we first filter (description, predicate) that occur at least 5 times, and then sample 5 pairs for each combination.

Among the 1078 records extracted from DBPedia abstracts of places in the states of Wisconsin, Texas, and California, 212 explicitly refer to directional and distance spatial relations and were thus removed as this research focused on topological relations. The analytical results of task 3 using the remaining records will be presented in Section 5.3.

4.2. Experiment models

In this research, we perform evaluation tasks based on the following models:

4.2.1. Embedding models
We encode WKT geometries into embeddings and process spatial queries using 'text-embedding-ada-002' and 'text-embedding-3-large' provided by OpenAI[6], with output embedding dimensions of 1536 and 3072, respectively.

4.2.2. Reasoning models

In our evaluation tasks, we employ GPT-3.5-turbo, GPT-4 and DeepSeek-R1-14B as the LLM-based reasoning models. While performance varies by task, these models have demonstrated potential in commonsense reasoning and in-context learning on certain benchmarks. GPT-3.5-turbo and GPT-4 are primarily optimized for few-shot learning, whereas DeepSeek-R1-14B emphasizes zero-shot capabilities and may experience a decline in performance when few-shot prompting is applied (Guo *et al*. 2025).

4.2.3. Model settings

1. Random Forest classifier: The number of estimators (trees) in the Task 1 classifier is set to 100.
2. Temperature settings for GPT-3.5-turbo and GPT-4: For the topological relation qualification task, we set the temperature to 0 to encourage more deterministic outputs. However, achieving full reproducibility remains challenging, even with a temperature of 0, as discussed by Blackwell *et al*. (2024). Conversely, generating synthetic geometries to support semantic search in Task 2, employs a higher temperature of 0.7 for greater creativity.
3. Temperature settings for DeepSeek-R1-14B: The temperature of the topological relation qualification task is set to be 0.6 to better exploit the reasoning ability of DeepSeek, given its emphasis on deeper, more deliberate thinking.

5. Results

5.1. Topological spatial relation qualification

5.1.1. Validity of the output

Before diving into the effectiveness of using LLMs to qualify spatial relationships, a validity check is necessary because of the inherent nondeterministic nature of generative AI models. Furthermore, beyond validating the output as a valid format of {Geometry type *A*, predicate, Geometry type *B*}, it is essential to focus on grounding the qualitative spatial reasoning in the matched geometry types and topological relations.

The validity results of the output are shown in Table 4. The random forest classifier using the LLM-generated embeddings consistently produced valid output on the test dataset. This highlights that the sentence embedding models can effectively preserve geometry types in the WKT format of geometries, aligning with previous research which encoded WKT by aggregating the token embeddings from GPT-2 and BERT (Ji and Gao 2023). While GPT-4 and GPT-3.5-turbo largely adhere to the instructions in the desired format, even with the CoT generation, it is more challenging for DeepSeek-R1-14B to strictly output the desired format (but still achieved over 0.9 validity accuracy). When tested with few-shot prompting, the DeepSeek-R1-14B largely ignored the provided examples and adhered to its typical reasoning patterns. As a result, we did not include these results in our evaluation. The highest validity of GPT-4 model suggests that a language model that is characterized by a larger number of parameters, broader training data, and stronger alignment with human instructions, may also possess a better understanding of the definitions of the DE-9IM topological predicates.

Table 4. The validity accuracy of the outputs (N/A: not available).

Approach	LLM	Prompt	Format	Geometry type	Predicate
Random Forest	text-embedding-ada-002	N/A	1	1	1
	text-embedding-3-large	N/A	1	1	1
Question Answering	GPT-3.5-turbo	Zero-shot	0.959	1	0.911
		Zero-shot-dim	0.999	1	0.927
		Few-shot	1	1	0.901
		Zero-shot-CoT	0.944	1	0.944
		Few-shot-CoT	0.998	1	0.894
	GPT-4	Zero-shot	1	0.996	0.997
		Zero-shot-dim	1	0.999	0.999
		Few-shot	1	0.999	0.992
		Zero-shot-CoT	0.984	0.990	0.968
		Few-shot-CoT	1	0.999	0.999
	DeepSeek-R1-14B	Zero-shot	0.936	0.996	0.913
		Zero-shot-dim	0.919	0.998	0.913

Table 5. The topological predicate classification metrics using topological distance.

Approach	LLM	Prompt	Accuracy	Distance in conceptual neighborhood
Random Forest	text-embedding-ada-002	N/A	0.633	1.449
	text-embedding-3-large	N/A	0.632	1.419
Question Answering	GPT-3.5-turbo	Zero-shot	0.423	1.331
		Zero-shot-dim	0.408	1.360
		Few-shot	0.479	1.595
		Zero-shot-CoT	0.443	1.370
		Few-shot-CoT	0.465	1.174
	GPT-4	Zero-shot	0.632	1.238
		Zero-shot-dim	0.635	1.212
		Few-shot	0.666	1.272
		Zero-shot-CoT	0.610	1.256
		Few-shot-CoT	0.627	1.225
	DeepSeek-R1-14B	Zero-shot	0.534	1.257
		Zero-shot-dim	0.557	1.260

5.1.2. Classification metrics

Table 5 presents the results of the topological predicate classification task. Both the embedding-based random forest and geospatial question-answering with GPT models can achieve an accuracy of over 0.6. This suggests that identifying topological spatial relationships from the WKT format of geometries with LLMs is promising but remains challenging. Failure to recover the topological spatial relations from embeddings suggests a potential information loss through text tokenization. Incorrectly classified topological relations often cluster within the conceptual neighborhoods or resemble each other (with a small distance), while confusion may also arise from the diverse semantics of topological spatial predicates.

For GPT-3.5-turbo and GPT-4, among the four types of prompts (introduced in Section 3.4.1), few-shot learning achieved the best performance with pairs of geometries and their topological relationships for LLMs to learn in context. GPT-4 with few-short promoting achieved 0.66 accuracy. The findings highlight the importance of prompt engineering in the use of LLMs and the critical role of understanding spatial contexts in improving geospatial query processing accuracy and reliability. However, chain-of-thought (CoT) prompts, which have demonstrated improvement in many

other tasks (Wei *et al.* 2022), did not yield the expected benefit in our spatial reasoning evaluation experiments. As mentioned by Yang *et al.* (2024), CoT reasoning can sometimes induce unreliable or counterproductive outputs in spatial reasoning tasks. Upon analyzing the generated rationale, we observed that when LLMs are prompted with 'Let's think step by step', they attempt to check the topological spatial predicates one by one based on their respective definitions from the OGC standard. Few-shot-CoT prompts, on the other hand, were explicitly designed with examples grounded in scientific definitions and logical decision processes proposed in Clementini *et al.* (1993), aiming to 'teach' the models to reason about topological spatial relations from analysis on interiors, boundaries, and exteriors. Despite this structured approach, the accuracy declined due to cascading errors in intermediate steps, such as failing to determine whether the interiors of two geometries intersect at the very beginning. With an explicit reasoning process, DeepSeek-R1-14B outperformed GPT-3.5-turbo (20B parameters). Analysis of its thought generation reveals that, rather than always iteratively checking candidate answers, the model often employed more intuitive reasoning strategies, such as mental mapping (e.g. 'Let me plot them mentally') and self-verification (e.g. 'In WKT, a LINESTRING is just a sequence of points connected by straight lines. If it starts and ends at the same point, it doesn't automatically become a polygon'). Although these reasoning patterns may appear convincing when interpreting individual geometries, they often fall short when reasoning about spatial relations between two geometries. This is mainly due to an overreliance on superficial, linear interpretations of coordinate information, rather than a holistic understanding of topological spatial relationships across the plane.

5.2. Spatial query processing

Based on the superior performance in task 1, the experiments in task 2 only used GPT-4 as the geometry generator and text-embedding-3-large as the embedding model.

5.2.1. Direct query

We first identified an effective query format for geospatial semantic search (Hu *et al.* 2015), which is the foundation for applying query expansion in understanding geospatial semantics. As shown in Table 6, specifying the subject geometry type would achieved higher performance due to a narrowed mapping space to the same geometry type. In the following experiments, we assumed that the user query with the

Table 6. Spatial query performance comparison results.

Target		Query Format	MRR	Hits@5	Hits@10	Hits@20
Subject	Direct query	Abstract as 'geometry'	0.081	0.131	0.161	0.194
		Specify the subject geometry type	0.152	0.212	0.26	0.29
	Expanded query	Direct query + one LLM-generated geometry.	0.18	0.238	0.278	0.328
		Direct query + three LLM-generated geometry.	0.169	0.232	0.28	0.32
Object	Direct query	Original predicate	0.105	0.131	0.17	0.211
		Reversed predicate	0.152	0.219	0.256	0.297
	Expanded query	Original predicate + one LLM-generated geometry.	0.15	0.215	0.261	0.302
		Reversed predicate + one LLM-generated geometry.	0.179	0.248	0.294	0.333

Table 7. Validity of LLM-generated geometries using different prompts.

Prompt	Valid_WKT	Geometry Type	Predicate	Topological Distance
Zero-shot	0.999	1	0.763	1.142
Zero-shot-Check	0.998	1	0.755	1.075
Few-shot	0.996	1	0.728	1.177
Few-shot-Negative	0.997	1	0.754	1.212
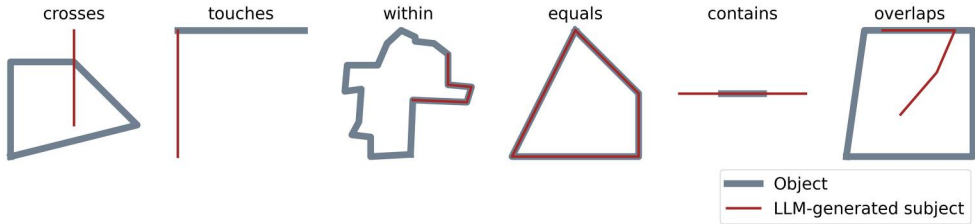
Figure 10. Synthetic geometries generated by GPT-4 for LineString/LineString relations.

geometry type (e.g. retrieving a street from the spatial database implies LineString), and further investigated the factors that may impact the effectiveness of query expansion using LLMs.

5.2.2. Synthetic geometry generation

An effective LLM-generated geometry is expected to maintain the same topological spatial relation with the given object as the subject entity while being close to the subject entity in the embedding space. Table 7 compares the validity of the LLM-generated geometries produced by different prompting approaches. We find that 1) GPT-4 effectively comprehends spatial queries and generates geometries in a valid WKT format, and 2) GPT-4 demonstrates a notable level of spatial reasoning regarding the reference object, even in the zero-shot setting, as indicated by the high relation-preserving accuracy (over 0.72) and the low topological distance in the conceptual neighborhood graph (Reis *et al.* 2008). Figure 10 presents examples of the LLM-generated geometries generated by GPT-4. In the following section, we will check the usefulness of such synthetic geometries generated by zero-shot prompts in enriching spatial query processing.

5.2.3. Expanded query on subject retrieval

As shown in Table 6, retrieving a subject based on the embeddings encoded from the expanded spatial queries remains challenging. However, including an LLM-generated geometry enhanced the probability of ranking the target subject higher among all candidates. Over 23% of the subjects were ranked within the top 5 candidates. But adding additional synthetic geometries did not appear to provide further improvements.

5.2.4. Performance comparison on object retrieval

While the above experiment primarily focuses on retrieving the subject in a triplet, we proceed to evaluate the performance of object retrieval. For a given triplet, we tested the queries formulated with either the original predicate describing the spatial

relationship from the subject to the object (e.g. 'Retrieve a Point which A contains'), or the reversed predicate referring to the subject (e.g. 'Retrieve a Point which is within A'). The results of object retrieval are also summarized in Table 6.

Among the object-retrieval query formats, the queries with the original predicate that maintained the subject-to-object directionality yielded worse performance. When we manually reversed the topological spatial relation and treated the object as the subject, the performance matched its counterpart for subject retrieval in Table 6, high-lighting the importance of structuring spatial queries to align with everyday language patterns commonly used for spatial reference.

5.3. Vernacular topological relation conversion

In Task 3, we collected textual descriptions of topological spatial relations between two places and attempted to identify mapping patterns between these descriptions and the corresponding context-conditioned topological spatial relations. These map-pings were then used as input to GPT-4 (giving its superior performance in previous tasks) to evaluate their ability to convert textual descriptions into formal topological spatial relations.

5.3.1. Conversion pairs invariant to context

Table 8 lists the six descriptions that consistently map to the same topological rela-tionship in our dataset. However, the results show varying levels of conversion accur-acy from vernacular descriptions to preferred formal topological predicates. While the ground truth topological relationships were likely to be implied from 'share border with' and 'is the location of', 'is an enclave of' was interpreted as *within* instead of *touches* or *disjoint*. Even though GPT-4 could infer an *overlaps* relation from 'has part of the population in' in all 10 experiments, the model might be unsure about its answer and would provide multiple topological predicate alternatives. Despite the sub-tle difference between 'midway' and 'halfway', a higher entropy of 'halfway' indicates greater randomness in the conversion.

5.3.2. Conversion pairs conditioned on place types or geometry types

The comparisons between scenarios with and without place-type/geometry-type con-text are illustrated in Tables 9 and 10, respectively. Our initial hypothesis was that including contextual information in the prompt would reduce ambiguity, resulting in a higher frequency of correct predicate predictions, improved accuracy, and lower entropy using LLMs. However, these improvements were highly instance-dependent and not consistently observed across all the conversion pairs evaluated in our

Table 8. Result of topological relation conversion pairs invariant to context.

Description	Predicate	Frequency	Accuracy	Entropy
share border with	touches	10	1.000	0.000
has part of the population in	overlaps	10	0.588	0.435
is the location of	contains	9	0.818	0.244
is midway between C and	disjoint	6	0.600	0.560
is halfway between C and	disjoint	6	0.500	0.817
is an enclave of	touches	1	0.100	0.167

Table 9. Result of topological relation conversions using place type as context.

Description	Predicate	Spatial Context	Frequency	Accuracy	Accuracy without Context	Entropy	Entropy without Context
is home to	contains	city/amenity	10	1	1	0	0
borders	touches	city/municipality	10	1	1	0	0
is located in	within	town/county	10	1	1	0	0
is located in	within	city/state	10	1	1	0	0
is bordered by	touches	town/city	10	1	1	0	0
is adjacent to	touches	city/municipality	10	1	0.909	0	0.157
borders	touches	city/city	10	1	1	0	0
is in	within	village/county	10	0.909	0.909	0.157	0.157
is located in	within	village/county	10	0.909	1	0.157	0
is partly in	overlaps	city/county	10	0.833	1	0.232	0
is bounded by	touches	city/city	7	0.7	0.333	0.314	0.327
connect C and	crosses	industrial/city	8	0.4	0.474	0.52	0.355
extend into	overlaps	city/county	5	0.357	0.421	0.561	0.491
is surrounded by	touches	city/city	2	0.167	0	0.232	0
is between C and	touches	town/town	0	0	0.3	0.211	0.773
is surrounded by	touches	town/city	0	0	0	0	0
is within	touches	city/municipality	0	0	0	0	0

Table 10. Result of topological relation conversions using geometry type as context.

Description	Predicate	Spatial Context	Frequency	Accuracy	Accuracy without Context	Entropy	Entropy without Context
is in	within	Polygon/MultiPolygon	10	1	0.909	0	0.157
is neighboring	touches	Polygon/Polygon	10	1	1	0	0
is bordered by	touches	Polygon/MultiPolygon	10	1	1	0	0
is the county seat of	within	Polygon/Polygon	10	1	1	0	0
extend into	overlaps	Polygon/Polygon	8	0.714	0.421	0.307	0.491
connect C and	crosses	LineString/MultiPolygon	9	0.412	0.474	0.348	0.355
is surrounded by	touches	Polygon/MultiPolygon	0	0	0	0	0
is on	crosses	Polygon/LineString	0	0	0	0	0.327

experiments, indicating GPT-4's limitation in considering all possible interpretations of the given vernacular description. This limitation was also evident in pairs with 0 accuracy, where the model consistently outputted the same incorrect answer. In other instances, GPT-4 still struggled to determine the appropriate topological predicates for certain vernacular descriptions.

5.3.3. Conversion pairs with place names

The accuracy and entropy of the topological relation conversions with place names were also compared to the metrics obtained without the context. As shown in Table 11, mentioning place names did not necessarily improve the accuracy of the conversion or guide the LLM to a preferred answer. GPT-4's explanation indicates that (1) It focuses on the topological relationships between general geographic locations or boundaries rather than leveraging specific knowledge about each place; (2) The approach tends to exclude predicates possibly with inaccurate and abstract geometries. For instance, in analysis 'A is along B' when A is Brazos Bend, Texas, and B is Brazos River, the reasoning begins with 'This suggests a specific geographical relationship between a place (A) and a river (B). The term "along" typically indicates that A is situated in a linear arrangement adjacent to B, but not necessarily crossing it or being contained within it.'

**Table 11.** Accuracy and Entropy changes for conversion pairs with place names.

Accuracy	Topological relation conversion pairs (Order by the absolute values in change,* with entropy reduction).	
Improves	1) is bounded by → touches*	2) is surrounded by → touches
	3) is suburb of → touches	4) is on → crosses
	5) is part of → within*	6) is between C and → touches*
	7) is partly in → touches	8) is suburb of → disjoint
	9) is between C and → disjoint*.	10) is near → touches.
Unchanged	Remains 1:	
	1) includes → contains	2) borders → touches
	3) is bordered by → touches	4) is neighboring → touches
	5) is located in → within	6) is the county seat of → within
	Remains 0:	8) is in → overlaps*
	7) is bordered by → disjoint	10) is within → touches.
	9) on the shore of → overlaps.	
Declines	1) is mostly in → overlaps	2) is near → disjoint
	3) is along → crosses*	4) is partly in → overlaps
	5) is situated on → overlaps	6) is adjacent to → touches
	7) extend into → overlaps	8) is in → within
	9) is home to → contains.	10) connect C and → crosses*.

6. Discussion

In this section, we would like to further discuss whether the confusion between the topological predicates aligns with the corresponding conceptual neighborhood of topological spatial relations (Egenhofer and Al-Taha 1992, Egenhofer and Mark 1995, Formica *et al.* 2018), the confusion in geometry generation, and the confusion in vernacular description conversion.

6.1. Confusion between topological predicates in topological spatial relation qualification

When using GPT-4 (zero-shot learning) for topological spatial relation qualification, the confusion matrices for all the geometry type combinations are drawn in Figure 11. We compare the topological predicate pairs that may confuse GPT models with the classic conceptual neighborhood graphs in Figure 1. The observations are twofold: (1) The most frequently confused topological spatial relation for a given predicate depends on the geometry types involved. For example, consider the predicate 'overlaps.' In a Linestring/Linestring relationship, it is rarely classified correctly and is often confused with 'crosses', 'equals' or 'touches'. However, in a Polygon/Polygon relationship, 'overlaps' is more likely to be correctly identified, though it may occasionally be confused with 'contains' or 'disjoint'. Another illustrative example involves the predicate 'touches.' A Point 'touches' a Linestring or Polygon, or a Linestring 'touches' a Polygon is frequently mistaken as 'within' while such confusion is less between two geometries of the same dimension, such as two Polygons or two Linestrings. These examples suggest the varied degree to which an LLM understands formal geometry boundaries associated with geometry types, particularly their dimensions, which is crucial in identifying formal topological spatial relations. However, the constraint of formal definitions may contradict common conceptual interpretations, such as excluding a polygon from containing its boundary, leading to fewer occurrences in GPT-4's response. (2) For the same geometry type combination, distinguishing certain pairs of topological
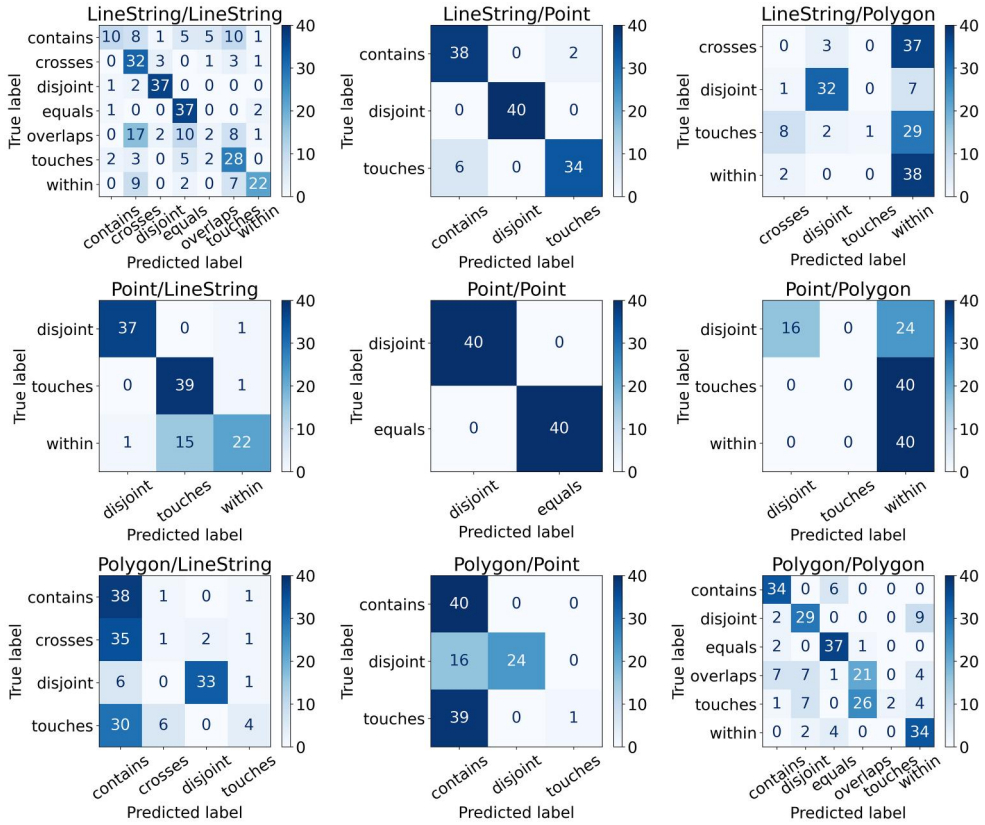
Figure 11. Confusion matrices between topological predicates in relation qualification.

spatial relations is more challenging than others. These pairs mostly fall within the conceptual neighborhood, though exceptions exist. Take 'Linestring/Linestring' as an example. GPT-4 can identify 'crosses', 'disjoint', and 'equals' more accurately. However, it struggles with predicates like 'contains' and 'overlaps,' frequently confusing them with 'crosses' or 'touches'. The four topological spatial relations all require that the two geometries share elements like points or line segments and might be interchangeable in daily use. This challenge highlights that the ambiguous semantics of these predicates can encompass scenarios broader than their strict formal definitions. Overall, the issues observed in Task 1 actually reflect an alignment with everyday spatial reasoning. While formal definitions are precise and dimension-contingent, everyday language and intuitive reasoning often blur the distinctions.

6.2. Confusion between topological predicates in geometry generation

The confusion pattern changes when leveraging an LLM to generate geometries given a spatial query and the required geometry type, as shown in Figure 12. The findings can be summarized as follows: (1) Directionality in describing the topological relation between two geometry types matters. For example, generating a Polygon that 'crosses' a LineString proves challenging for GPT-4, while the reversed query—
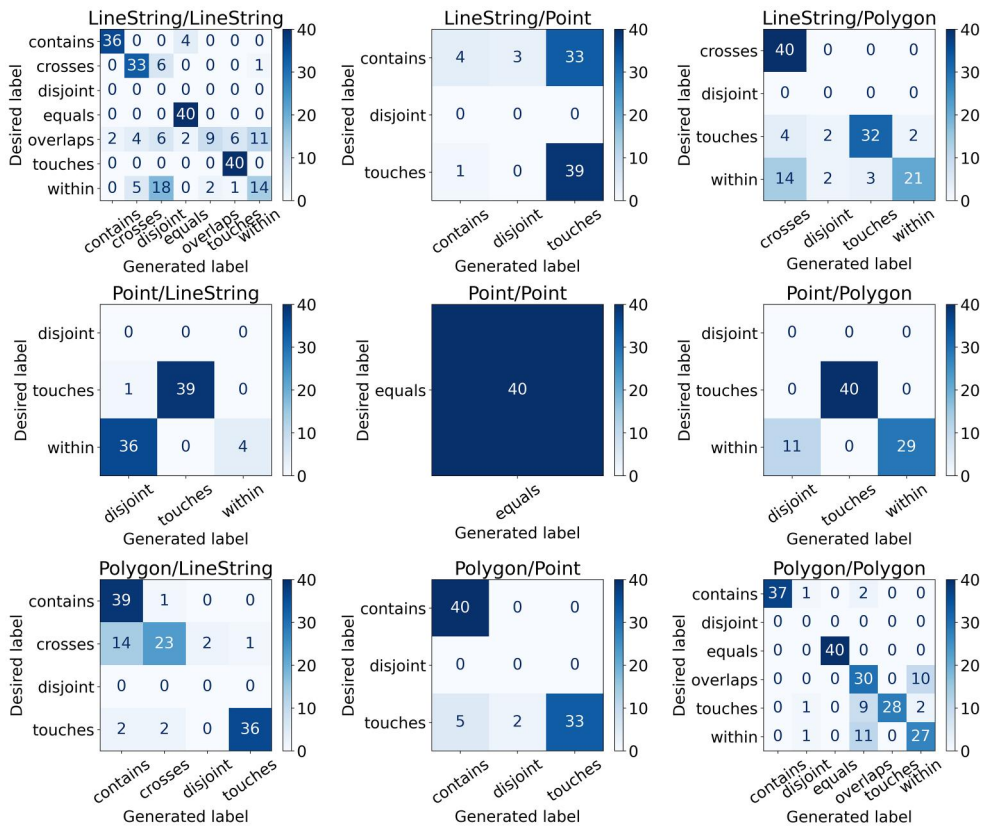
Figure 12. Confusion matrices between spatial predicates in geometry generation.

generating a LineString that 'crosses' a Polygon—is handled more effectively. Similarly, the model is more successful in generating a Polygon that 'contains' a Point, LineString, or another Polygon but struggles to produce a Point, LineString, or Polygon that is 'within' a Polygon. This asymmetry can be attributed to the model's approach of extracting coordinates from the query geometry to construct the second geometry. This reliance limits the model's ability to conceptualize spatial relationships beyond the provided coordinates. (2) The geometry type of the reference object affects the results. Figure 13 provides examples of LineString/LineString topological spatial relations where the generated topological spatial relations were different from the predicate in the spatial query. As observed from these examples, even if the spatial queries specify the reference object geometry type as LineString, the model sometimes applies definitions for Polygons when a line forms a closed shape. In this case, when we manually changed the reference object geometry type into Polygon and recompute its topological relations with the LLM-generated geometry, 223 out of 391 queries (across all prompts) with closed geometries were found to exhibit the desired topological relationship. This observation suggests GPT-4's perception based on the provided coordinates over the geometry type specified in the text, inspiring us to further explore the cognition potential of the LLMs.
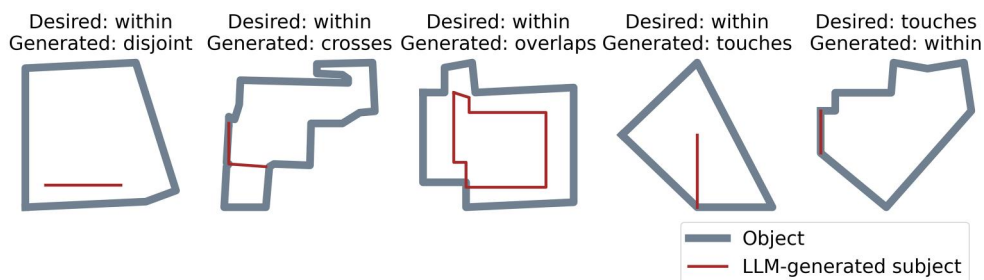
Figure 13. Invalid synthetic geometries generated by GPT-4 for LineString/LineString relations with close-shape objects.

6.3. Confusion between topological spatial relations in conversion

While GPT-4 can accurately convert several vernacular descriptions to corresponding formal predicates, there are instances where its performance falls short. This can be attributed to the mismatch between crispy geometry and the vague human perception of place boundaries. We can further divide it into three categories. (1) The abstraction of spatial entities' shapes in the spatial dataset may differ from those used in descriptions. For example, when converting 'is along' for Brazos Bend, Texas, and Brazos River, Texas, GPT-4 returned 'touches' when considering Brazos Bend as a region (Polygon) and 'within' when considering Brazos Bend as a Point. (2) The computed topological spatial relations can be sensitive to the marked shape points, while human perception can tolerate such systematic errors, yielding the description of the relations in the conceptual neighborhood of the ground truth. A typical example is when the ground truth label of 'is suburb of' is 'disjoint', but the two cities look like they 'touch' each other on the map. (3) Official geographic boundaries might differ from people's perception of a place (Gao *et al.* 2017). In our dataset, 'is within' and 'is an enclave of' can map to 'touches'. But the LLM would constantly output 'within'. For instance, the City of Shullsburg, Wisconsin, and the Town of Shullsburg, Wisconsin, illustrate this discrepancy[7]. Although the City of Shullsburg is enclosed by the Town of Shullsburg, the city boundary is separated from the town boundary, creating a hole in the town boundary. In summary, even though GPT-4's responses can be partly interpreted from the conceptual neighborhood of topological spatial relations, challenges remain due to the vagueness of real-world geographic entity boundaries and human perception of shapes and places.

7. Conclusion and future work

This study focuses on the evaluation of the ability of LLMs including GPT-3.5, GPT-4 and DeepSeek-R1-14B to process, represent, and reason with topological spatial relations. Consequently, we designed a workflow to assess the efficacy of LLMs in addressing three typical problems on topological spatial relations. The core idea involves converting geometric objects into textual strings (WKT), which can then be decoded and utilized for spatial reasoning. The first task, topological spatial relation

qualification, focuses on determining if such textual representation can retain the necessary geometric information for deriving named topological predicates. The second task explores the feasibility of conducting geospatial queries through semantic search, where LLMs can generate a geometry to augment the query and also generate embeddings. The third task presents an everyday scenario where an LLM serves as a translator to convert vernacular descriptions of spatial relations into formalized topological predicates based on their capability to understand linguistic patterns.

From the multi-source geospatial datasets, we extract triplets to represent topological spatial relations in real-world spatial configurations. Using the triplets as input, we have compared the performance on the three evaluation tasks with ground truth data. In Task 1, both the random forest and GPT-based reasoning models can identify most relations correctly (over 0.6 accuracy on average), while some relations can be confounding. For GPT-3.5-turbo and GPT-4, few-shot prompt engineering is essential to improve the performance while CoT prompting strategy had a negative impact on our topological spatial relation inference task. The thought generation process and the self-verification allow DeepSeek-R1-14B to perform spatial reasoning more intuitively and outperformed GPT-3.5-turbo in accuracy. Further comparison with the conceptual neighborhood allows for a more quantitative understanding of the errors. Even though task 2 further verifies the challenge of replacing spatial queries with semantic search. However, improvements can be observed when we customize the query and augment it with LLM-generated geometries. The LLM-generated geometries are not only valid WKT but also have high accuracy (up to 0.76) in preserving topological spatial relations (or within their conceptual neighbors). In task 3, the improvement of LLMs to reduce ambiguity in spatial queries is relatively limited. However, in most cases, the generated outputs fall into the conceptual neighborhood of the ground-truth topological predicate. Moreover, given various contexts, the changes in the preferred response show the ability of the LLM to reason about it using commonsense knowledge and the typical spatial configurations. Interestingly, adding the geometry-type context in prompts has derived more performance improvement compared to the cases with adding the place-type context, but the performance of adding context or without context varies by instance.

In conclusion, through the three tasks and intensive experiments, we systematically approach the overarching question of LLM's ability in understanding geometry information and their topological spatial relations, moving from the broader challenge to more targeted strategies involving spatial context, tailored prompting techniques, and specialized domain knowledge in GIScience.

However, it is essential to acknowledge the limitations of our work. First, our focus was primarily on in-context learning, and we did not explore fine-tuning approaches, which could potentially yield further performance improvements. Retrieval-Augmented Generation (RAG) (Lewis et al. 2020) presents a promising approach for enhancing the qualitative spatial reasoning capabilities LLMs by integrating external spatial databases, GIS tools and domain-specific knowledge from GIScience. Unlike in-context learning, which allows for intuitive qualitative spatial reasoning, the effective implementation of RAG relies on the precise generation of formalism-based spatial queries from natural

language input along with reliable high-resolution datasets. Improving the translation from natural language to into symbolic form and logic also opens the door to neuro-symbolic approaches (Sheth *et al.* 2023), in which LLMs serve as translators that convert user text input into symbolic representations, which are then processed by symbolic engines with strong capabilities in explanation, verification, and formal reasoning. Realizing this potential requires addressing the challenges identified in this work, such as ambiguities in linguistic spatial descriptions, conceptual neighborhood relationships in topological reasoning, and the correct use of available GIS functions and analytical workflows. Additionally, our dataset is currently limited to the city and state levels, and further investigation into multi-scale spatial relations is still needed to fully capture the complexity of spatial interactions across different geographical scales and heterogeneous datasets. Moreover, the scope of this work is limited to topological relations in natural languages. While we can handcraft datasets with ground truth labels of formalized predicates for evaluation, the mathematical computation makes LLMs less competent than spatial databases. Directional and distance relations introduce more vagueness in language use due to factors such as shape and scale, especially when two places cannot be viewed as points. In future work, we plan to explore other spatial relations using datasets such as the Geograph project[8], which provides rich expressions of various spatial relations associated with geometries, text descriptions and photos (M Hall *et al.* 2011). This will enable us to evaluate LLMs' or multimodal foundation models' capabilities (e.g. vision-language geo-foundation models) in geospatial reasoning from a more comprehensive perspective (Mai *et al.* 2024). Lastly, we rephrased our own text instead of directly using paragraphs from DBpedia to allow for flexibility in introducing different context information. However, this approach may result in some loss of authenticity in language use, such as anaphora, which is prevalent in original text documents and worth exploring in future research.

In summary, this research demonstrates the promise and limitations of using state-of-the-art LLMs to analyze topological spatial relations, while offering insights for future research of advancing LLMs with geographical knowledge, aiming to develop GeoAI foundation models capable of qualitative spatial reasoning and other spatial intelligence tasks.

Notes

1. http://osmnx.readthedocs.io/.
2. http://slipo.eu/.
3. https://www.sco.wisc.edu/parcels/data/.
4. https://www2.census.gov/geo/tiger/TIGER2020PL/LAYER/BG/.
5. https://www.dbpedia.org/.
6. https://openai.com/blog/new-embedding-models-and-api-updates.
7. https://en.wikipedia.org/wiki/Shullsburg_(town),_Wisconsin.
8. http://www.geograph.org.uk.

Acknowledgments

the Environment (ICICLE). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funder(s).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Yuhan Ji is a PhD student GIScience at the Department of Geography, University of Wisconsin-Madison. Her main research interests include transportation, geospatial data science, and GeoAI approaches to human mobility.

Dr. Song Gao is an associate professor in GIScience at the Department of Geography, University of Wisconsin-Madison. He holds a Ph.D. in Geography at the University of California, Santa Barbara. His main research interests include GeoAI, geospatial data science, spatial networks, human mobility and social sensing.

Ying Nie is an undergraduate student at the Department of Computer Sciences, University of Wisconsin-Madison. Her main research interests include geospatial data science and GeoAI.

Dr. Ivan Majic is a Postdoc researcher in Geoinformatics at the Department of Geography and Regional Research, University of Vienna. His research interests include Spatial Data Science, GeoAI, and Qualitative Spatial Reasoning.

Dr. Krzysztof Janowicz is a Professor in Geoinformatics at the Department of Geography and Regional Research, University of Vienna. His research interests are knowledge representation and reasoning as they apply to spatial and geographic data, e.g. in the form of knowledge graphs.

ORCID

Song Gao ⓘD http://orcid.org/0000-0003-4359-6302

Data and codes availability statement

The data and codes supporting the main findings of this study are available at Figshare: https://doi.org/10.6084/m9.figshare.25127135.v1 and the GitHub repository at https://github.com/GeoDS/GeoFM-TopologicalRelations.

References

Achiam, J., *et al.*, 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alayrac, J.-B., *et al.*, 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35, 23716–23736.

Blackwell, R.E., Barry, J., and Cohn, A.G., 2024. Towards reproducible llm evaluation: quantifying uncertainty in llm benchmark scores. *arXiv preprint arXiv:2410.03492*.

Bordes, A., *et al.*, 2013. Translating embeddings for modeling multi-relational data. *Advances in Neural Information Processing Systems*, 26, 1–9.

Breiman, L., 2001. Random forests. *Machine Learning*, 45 (1), 5–32.

Brown, T., *et al.*, 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

Carlson, L.A., and Logan, G.D., 2001. Using spatial terms to select an object. *Memory & Cognition*, 29 (6), 883–892.

Carpineto, C., and Romano, G., 2012. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys*, 44 (1), 1–50.

Cervone, G., *et al.*, 2016. Using Twitter for tasking remote-sensing data collection and damage assessment: 2013 boulder flood case study. *International Journal of Remote Sensing*, 37 (1), 100–124.

Chen, H., *et al.*, 2018. A graph database model for knowledge extracted from place descriptions. *ISPRS International Journal of Geo-Information*, 7 (6), 221.

Chen, J., *et al.*, 2015. A survey of qualitative spatial representations. *The Knowledge Engineering Review*, 30 (1), 106–136.

Chen, M., *et al.*, 2021. Evaluating large language models trained on code. *arXiv preprint arXiv: 2107.03374*.

Clementini, E., and Cohn, A.G., 2014. RCC*-9 and CBM. In *International Conference on Geographic Information Science*, p. 349–365. Springer.

Clementini, E., and Cohn, A.G., 2024. Extension of RCC*-9 to complex and three-dimensional features and its reasoning system. *ISPRS International Journal of Geo-Information*, 13 (1), 25.

Clementini, E., and Di Felice, P., 1996. A model for representing topological relationships between complex geometric features in spatial databases. *Information Sciences*, 90 (1-4), 121–136.

Clementini, E., Di Felice, P., and Van Oosterom, P., 1993. A small set of formal topological relationships suitable for end-user interaction. In *International symposium on spatial databases*, p. 277–295. Springer.

Clementini, E., Sharma, J., and Egenhofer, M.J., 1994. Modelling topological spatial relations: Strategies for query processing. *Computers & Graphics*, 18 (6), 815–822.

Cohn, A.G., 2023. An evaluation of ChatGPT-4's qualitative spatial reasoning capabilities in RCC-8. *arXiv preprint arXiv:2309.15577*.

Cohn, A.G., and Blackwell, R.E., 2024a. Can large language models reason about the region connection calculus? *arXiv preprint arXiv:2411.19589*.

Cohn, A.G., and Blackwell, R.E., 2024b. Evaluating the Ability of Large Language Models to Reason About Cardinal Directions. In Adams, B., Griffin, A. L., Scheider, S., and McKenzie, G., editors, *16th International Conference on Spatial Information Theory (COSIT 2024), volume 315 of Leibniz International Proceedings in Informatics (LIPIcs)*, pages 28:1–28:9, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Cohn, A.G., and Hazarika, S.M., 2001. Qualitative spatial representation and reasoning: An overview. *Fundamenta Informaticae*, 46, (1-2):1–29.

Cohn, A.G., and Hernandez-Orallo, J., 2023. Dialectical language model evaluation: An initial appraisal of the commonsense spatial reasoning abilities of llms. *arXiv preprint arXiv: 2304.11164*.

Cohn, A.G., and Renz, J., 2008. Qualitative spatial representation and reasoning. *Foundations of Artificial Intelligence*, 3, 551–596.

Cui, Z., Cohn, A.G., and Randell, D.A., 1993. Qualitative and topological relationships in spatial databases. In *Advances in Spatial Databases: Third International Symposium, SSD'93 Singapore, June 23–25, 1993 Proceedings 3*, p. 296–315. Springer.

Das, S., 2023. *Evaluating the capabilities of large language models for spatial and situational understanding*. PhD thesis, (MA). University of Cambridge.

Devlin, J., *et al.*, 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A., *et al.*, 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, S., *et al.*, 2005. Spatial data query based on natural language spatial relations. In *Proceedings. 2005 IEEE International Geoscience and Remote Sensing Symposium, 2005. IGARSS'05.*, volume 2, p. 1210–1213. IEEE.

Egenhofer, M.J., and Al-Taha, K.K., 1992. Reasoning about gradual changes of topological relationships. In Goos, G., Hartmanis, J., Frank, A. U., Campari, I., and Formentini, U., editors, *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space, volume 639 of Lecture Notes in Computer Science*, p. 196–219. Springer: Berlin, Heidelberg.

Egenhofer, M.J., and Franzosa, R.D., 1991. Point-set topological spatial relations. *International Journal of Geographical Information Systems*, 5 (2), 161–174.

Egenhofer, M.J., and Herring, J.R., 1991. Categorizing binary topological relations between regions, lines, and points in geographic databases. Technical report, Department of Surveying Engineering, University of Maine, Orono, Maine, USA.

Egenhofer, M.J., and Mark, D.M., 1995. Modelling conceptual neighbourhoods of topological line-region relations. *International Journal of Geographical Information Systems*, 9 (5), 555–565.

Fernandez, A., and Dube, S., 2023. Core building blocks: Next gen geo spatial GPT application. *arXiv preprint arXiv:2310.11029*.

Formica, A., *et al.*, 2018. Approximate answering of queries involving polyline–polyline topological relationships. *Information Visualization*, 17 (2), 128–145.

Frank, A.U., 1992. Qualitative spatial reasoning about distances and directions in geographic space. *Journal of Visual Languages & Computing*, 3 (4), 343–371.

Freksa, C., Habel, C., and Wender, K.F., 1998. *Spatial cognition: An interdisciplinary approach to representing and processing spatial knowledge*, vol. 1404. Berlin, Germany: Springer Science & Business Media.

Frolov, S., *et al.*, 2021. Adversarial text-to-image synthesis: a review. *Neural Networks: The Official Journal of the International Neural Network Society*, 144, 187–209.

Fulman, N., Memduhoğlu, A., and Zipf, A., 2024. Distortions in judged spatial relations in large language models. *The Professional Geographer*, 76 (6), 703–711.

Gao, S., and Goodchild, M.F., 2013. Asking spatial questions to identify GIS functionality. In *2013 Fourth International Conference on Computing for Geospatial Research and Application*, p. 106–110. IEEE.

Gao, S., Hu, Y., and Li, W., 2023. *Handbook of Geospatial Artificial Intelligence*. New York, USA: CRC Press.

Gao, S., *et al.*, 2017. A data-synthesis-driven method for detecting and extracting vague cognitive regions. *International Journal of Geographical Information Science*, 31 (6), 1–27.

Gramacki, P., Martins, B., and Szymański, P., 2024. Evaluation of code llms on geospatial code generation. In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, 54–62.

Guo, D., *et al.*, 2022. DeepSSN: A deep convolutional neural network to assess spatial scene similarity. *Transactions in GIS*, 26 (4), 1914–1938.

Guo, D., *et al.*, 2025. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Guo, R., 1998. Spatial objects and spatial relationships. *Geo-Spatial Information Science*, 1 (1), 38–42.

Head, C.G., 1984. The map as natural language: a paradigm for understanding. *Cartographica*, 21 (1), 1–32.

Hu, Y., *et al.*, 2024. A five-year milestone: reflections on advances and limitations in GeoAI research. *Annals of GIS*, 30 (1), 1–14.

Hu, Y., Janowicz, K., and Prasad, S., 2014. Improving wikipedia-based place name disambiguation in short texts using structured data from dbpedia. In *Proceedings of the 8th workshop on geographic information retrieval*, p. 1–8.

Hu, Y., *et al.*, 2015. Metadata topic harmonization and semantic search for linked-data-driven geoportals: A case study using ArcGIS Online. *Transactions in GIS*, 19 (3), 398–416.

Hu, Y., *et al.*, 2023. Geo-knowledge-guided gpt models improve the extraction of location descriptions from disaster-related social media messages. *International Journal of Geographical Information Science*, 37 (11), 2289–2318.

Huang, W., Wang, J., and Cong, G., 2024. Zero-shot urban function inference with street view images through prompting a pretrained vision-language model. *International Journal of Geographical Information Science*, 38 (7), 1414–1442.

Jakubik, J., *et al.*, 2023. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*.

Janowicz, K., 2023. Philosophical foundations of GeoAI: Exploring sustainability, diversity, and bias in GeoAI and spatial data science. *arXiv preprint arXiv:2304.06508*, p. 1–17.

Janowicz, K., *et al.*, 2020. GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. *International Journal of Geographical Information Science*, 34 (4), 625–636.

Janowicz, K., *et al.*, 2015. Why the data train needs semantic rails. *AI Magazine*, 36 (1), 5–14.

Ji, Y., and Gao, S., 2023. Evaluating the effectiveness of large language models in representing textual descriptions of geometry and spatial relations (short paper). In *12th International Conference on Geographic Information Science (GIScience 2023)*, 43, 1–6. Schloss-Dagstuhl-Leibniz Zentrum für Informatik.

Jones, C.B., *et al.*, 2004. The spirit spatial search engine: Architecture, ontologies and spatial indexing. In *Geographic Information Science: Third International Conference, GIScience 2004, Adelphi, MD, USA, October 20-23, 2004. Proceedings 3*, 125–139. Springer.

Kasneci, E., *et al.*, 2023. ChatGPT for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274.

Kirillov, A., *et al.*, 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Kojima, T., *et al.*, 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35, 22199–22213.

Kordjamshidi, P., Van Otterlo, M., and Moens, M.-F., 2011. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing*, 8 (3), 1–36.

Lewis, P., *et al.*, 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.

Li, W., *et al.*, 2023. Assessment of a new GeoAI foundation model for flood inundation mapping. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, 102–109.

Li, Z., and Ning, H., 2023. Autonomous GIS: the next-generation AI-powered GIS. *International Journal of Digital Earth*, 16 (2), 4668–4686.

Logeswaran, L., and Lee, H., 2018. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*.

Loglisci, C., *et al.*, 2012. An unsupervised framework for topological relations extraction from geographic documents. In *Database and Expert Systems Applications: 23rd International Conference, DEXA 2012, Vienna, Austria, September 3–6, 2012. Proceedings, Part II 23*, 48–55. Springer.

Hall, M.M., Smart, P.D., and Jones, C.B., 2011. Interpreting spatial language in image captions. *Cognitive Processing*, 12 (1), 67–94.

Mai, G., *et al.*, 2022a. Towards a foundation model for geospatial artificial intelligence (vision paper). In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, 1–4.

Mai, G., *et al.*, 2024. On the opportunities and challenges of foundation models for GeoAI (vision paper). *ACM Transactions on Spatial Algorithms and Systems*, 10 (2), 1–46.

Mai, G., *et al.*, 2020. SE-KGE: A location-aware knowledge graph embedding model for geographic question answering and spatial semantic lifting. *Transactions in GIS*, 24 (3), 623–655.

Mai, G., *et al.*, 2022b. A review of location encoding for GeoAI: methods and applications. *International Journal of Geographical Information Science*, 36 (4), 639–673.

Majic, I., *et al.*, 2021. RIM: a ray intersection model for the analysis of the between relationship of spatial objects in a 2D plane. *International Journal of Geographical Information Science*, 35 (5), 893–918.

Majic, I., *et al.*, 2024. 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*, 99–105.

Manvi, R., *et al.*, 2023. GeoLLM: Extracting geospatial knowledge from large language models. *arXiv preprint arXiv:2310.06213*.

Mark, D.M., and Egenhofer, M.J., 1994. Modeling spatial relations between lines and regions: combining formal mathematical models and human subjects testing. *Cartography and Geographic Information Systems*, 21 (4), 195–212.

Mooney, P., Cui, W., Guan, B., and J., L., 2023. Towards understanding the geospatial skills of ChatGPT: Taking a geographic information systems (GIS) exam. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, 85–94.

Muennighoff, N., 2022. Sgpt: GPT sentence embeddings for semantic search. *arXiv preprint arXiv: 2202.08904*.

Neelakantan, A., *et al.*, 2022. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*.

OpenAI 2022. Introducing ChatGPT.

Radford, A., *et al.*, 2018. Improving language understanding by generative pre-training. *OpenAI Blog*, 0 (1), 12.

Radford, A., *et al.*, 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1 (8), 24.

Ramesh, A., *et al.*, 2021. DALL·E: Creating images from text.

Randell, D.A., Cui, Z., and Cohn, A.G., 1992. A spatial logic based on regions and connection. *Principles of Knowledge Representation and Reasoning*, 92, 165–176.

Rao, J., *et al.*, 2023. Building privacy-preserving and secure geospatial artificial intelligence foundation models (vision paper). In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pages 1–4.

Reimers, N., and Gurevych, I., 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *arXiv preprint arXiv:1908.10084*.

Reis, R.M., Egenhofer, M.J., and Matos, J.L., 2008. Conceptual neighborhoods of topological relations between lines. In *Headway in Spatial Data Handling: 13th International Symposium on Spatial Data Handling*, 557–574. Springer.

Renz, J., and Nebel, B., 1998. Spatial reasoning with topological information. In *Spatial Cognition: An Interdisciplinary Approach to Representing and Processing Spatial Knowledge*, 351–371. Berlin, Germany: Springer.

Sack, J.-R., and Urrutia, J., 1999. *Handbook of computational geometry*. Amsterdam, Netherlands: Elsevier.

Scheider, S., *et al.*, 2021. Geo-analytical question-answering with GIS. *International Journal of Digital Earth*, 14 (1), 1–14.

Shannon, C.E., 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27 (3), 379–423.

Sheth, A., Roy, K., and Gaur, M., 2023. Neurosymbolic artificial intelligence (why, what, and how). *IEEE Intelligent Systems*, 38 (3), 56–62.

Skoumas, G., *et al.*, 2016. Location estimation using crowdsourced spatial relations. *ACM Transactions on Spatial Algorithms and Systems*, 2 (2), 1–23.

Tao, R., and Xu, J., 2023. Mapping with ChatGPT. *ISPRS International Journal of Geo-Information*, 12 (7), 284.

Tellex, S., *et al.*, 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 1507–1514.

Thoppilan, R., *et al.*, 2022. LaMDA: Language models for dialog applications. *arXiv preprint arXiv: 2201.08239*.

Tucker, S., 2024. A systematic review of geospatial location embedding approaches in large language models: A path to spatial ai systems. *arXiv preprint arXiv:2401.10279*.

Wallgrün, J.O., Klippel, A., and Baldwin, T., 2014. Building a corpus of spatial relational expressions extracted from web documents. In *Proceedings of the 8th workshop on geographic information retrieval*, 1–8.

Wallgrün, J.O., Klippel, A., and Karimzadeh, M., 2015. Towards contextualized models of spatial relations. In *Proceedings of the 9th Workshop on Geographic Information Retrieval*, 1–2.

Wang, F., 2000. A fuzzy grammar and possibility theory-based natural language user interface for spatial queries. *Fuzzy Sets and Systems*, 113 (1), 147–159.

Wang, L., Yang, N., and Wei, F., 2023. Query2doc: Query expansion with large language models. *arXiv preprint arXiv:2303.07678*.

Wang, Z., Ye, X., and Tsou, M.-H., 2016. Spatial, temporal, and content analysis of Twitter for wildfire hazards. *Natural Hazards*, 83 (1), 523–540.

Wei, J., *et al.*, 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.

Wolter, D., and Wallgrün, J.O., 2012. Qualitative spatial reasoning for applications: New challenges and the sparq toolbox. *Qualitative Spatio-Temporal Representation and Reasoning: Trends and Future Directions: Trends and Future Directions*, 336.

Wu, K., *et al.*, 2023a. Deep learning models for spatial relation extraction in text. *Geo-Spatial Information Science*, 26 (1), 58–70.

Wu, M., *et al.*, 2023b. Mixed land use measurement and mapping with street view images and spatial context-aware prompts via zero-shot multimodal learning. *International Journal of Applied Earth Observation and Geoinformation*, 125, 103591.

Xie, Y., *et al.*, 2023. Geo-foundation models: Reality, gaps and opportunities. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, 1–4.

Yamada, Y., *et al.*, 2023. Evaluating spatial understanding of large language models. *arXiv preprint arXiv:2310.14540*.

Yan, B., *et al.*, 2017. From ITDL to Place2Vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In *Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems*, 1–10.

Yang, J., *et al.*, 2024. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*.

Yang, X., *et al.*, 2022. A large language model for electronic health records. *NPJ Digital Medicine*, 5 (1), 194.

Yuan, Y., 2011. Extracting spatial relations from document for geographic information retrieval. In *2011 19th International Conference on Geoinformatics*, 1–5. IEEE.

Zhang, Q., and Kang, Y., R., R., 2023. The Ethics of AI-Generated Maps: DALLE-2 and AI's Implications for Cartography (Short Paper). In *12th International conference on geographic information science (GIScience 2023)*, 93–98. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Zhang, Y., *et al.*, 2024. GeoGPT: An assistant for understanding and processing geospatial tasks. *International Journal of Applied Earth Observation and Geoinformation*, 131, 103976.

Zheng, O., *et al.*, 2023. ChatGPT is on the horizon: Could a large language model be all we need for intelligent transportation? *arXiv preprint arXiv:2303.05382*.

Zhu, R., *et al.*, 2022. Reasoning over higher-order qualitative spatial relations via spatially explicit neural networks. *International Journal of Geographical Information Science*, 36 (11), 2194–2225.