International Journal of Geo-Information

MDPI

Review

LLM4GIS: Methods, Applications, and Prospects

Huayi Wu [1,2,3], Lutong Xie [1], Shuyang Hou [1,*], Zhangxiao Shen [1], Haoyue Jiao [1], Shaowen Wu [1], Guanyu Chen [1], Ziqi Liu [1], Donghai Liu [1], Jianyuan Liang [1], Yaxian Qing [1], Rui Li [1], Zhipeng Gui [2,4] and Xuefeng Guan [1]

[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China; wuhuayi@whu.edu.cn (H.W.)
[2] Collaborative Innovation Center of Geospatial Technology, Wuhan 430079, China
[3] Hubei Luojia Laboratory, Wuhan 430079, China
[4] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China
* Correspondence: whuhsy@whu.edu.cn

Abstract

The rapid advancement of large language models (LLMs) offers new avenues for overcoming the limitations of current data-driven GIS analysis methods, giving rise to the emerging field of large language model-driven GIS analysis (LLM4GIS). This study presents a systematic review of 673 relevant research articles published between 2022 and July 2026. By tracing the five-stage evolutionary trajectory of GIS analysis methods and elaborating the framework for LLM4GIS, we conduct a quantitative bibliometric analysis that reveals the field's rapid growth from sporadic exploration to explosive expansion, along with the shifting landscape of its dominant technical trajectories. Building on these analyses, we develop a taxonomy spanning four dimensions—technical approaches, datasets, applications, and evaluation—and systematically analyze the advances and evaluation methodologies across five core application areas: geospatial knowledge question answering, knowledge extraction, spatiotemporal reasoning, geospatial code generation, and analysis modeling. Key challenges and future research directions are further discussed along five dimensions: multimodal data integration, high-quality data development, domain-specific model development, integration of world models with real-world geospatial environments, and interpretability and trustworthiness. This review provides a systematic reference for advancing the bidirectional synergy between LLM4GIS and GIS4LLM and for fostering a new generation of spatiotemporal intelligence ecosystems.

Keywords: large language model; GIS; LLM4GIS; spatiotemporal intelligence

Check for updates

Academic Editors: Wolfgang Kainz and Zhe (Sarina) Zhang

Received: 19 August 2026
Revised: 3 October 2026
Accepted: 13 October 2026
Published: 16 October 2026

Copyright: © 2026 by the authors. Published by MDPI on behalf of the International Society for Photogrammetry and Remote Sensing. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Geographic information systems (GISs) is an interdisciplinary field concerned with the representation, storage, analysis, and use of geographic information, drawing on geography, cartography, computer science, remote sensing, and related disciplines [1,2]. Spatial analysis lies at the core of GIS, transforming spatial data into actionable knowledge by characterizing spatial relationships among geographic features, quantifying spatial patterns, and modeling geographic processes. GIS-based spatial analysis has been widely adopted across domains such as urban and regional planning, transportation, disaster management, and public health [3–5], providing an essential foundation for interpreting geographic phenomena, addressing complex problems, and supporting integrated decision-making.

However, as geospatial data and analytical requirements become increasingly complex, existing GIS analysis still faces several limitations [6]. First, heterogeneous data from multiple sources, including remote sensing imagery, vector data, trajectories, and social sensing text, differ substantially in both structure and semantics, making unified cross-modal understanding and integrated analysis challenging [7]. Second, real-world GIS requirements are often open-ended and compositional, making them difficult to map directly onto predefined analytical tasks [8]. Moreover, spatial heterogeneity and scale effects limit the cross-scenario generalizability of models trained for specific regions or tasks [9]. The substantial expertise required in domain knowledge, GIS tools, and programming further constrains the broader adoption of complex GIS analysis in interdisciplinary applications [10].

Since the release of ChatGPT in November 2022, the rapid development of large language models (LLMs), represented by the GPT, Claude, Qwen, DeepSeek, and Llama families [11–15], has introduced new approaches and opportunities for addressing the aforementioned limitations. Pretrained on massive text corpora, LLMs acquire broad knowledge and strong capabilities in natural language understanding, generation, and instruction following [15,16]. Through targeted architectural design and training, they can further develop capabilities in logical reasoning, multimodal perception, and tool use, enabling them to perform increasingly complex tasks [17,18]. LLMs have been widely applied to general-purpose tasks such as conversational question answering, information extraction, and code generation [19]. Beyond these applications, they are increasingly being explored as research assistants for scientific tasks such as literature understanding, hypothesis generation, and experimental design across disciplines including biomedicine, materials science, and atmospheric science [20–23]. These developments have further stimulated domain-specific efforts to adapt and specialize foundation models for disciplinary problems [24,25]. In GIS spatial analysis, researchers have begun integrating LLMs into specific analytical tasks, leveraging their capabilities in semantic understanding, logical reasoning, and tool use to overcome limitations of conventional GIS analysis workflows. The growing convergence of LLMs and GIS is giving rise to an emerging research field centered on LLM-driven GIS analysis and driving GIS toward a new stage of spatiotemporal intelligence.

Despite the considerable potential of LLMs in GIS, research at the intersection of the two fields remains rapidly evolving and highly fragmented. Although several relevant reviews have emerged, their scope and perspectives remain limited in several respects. First, most focus on individual vertical domains, such as geographical research [26–28], spatial analysis and modeling [29–31], spatial database querying [32], remote sensing image understanding [33,34], or urban planning [35,36], and therefore lack a systematic synthesis from a general GIS perspective encompassing diverse forms of geospatial data and analytical tasks. Second, some reviews primarily examine GeoAI or geospatial foundation models [6,37–39], thereby covering a broad range of data-driven approaches based on conventional deep learning. LLMs, however, differ fundamentally from many other GeoAI techniques: they use natural language as a primary interface, can support a wide range of downstream tasks, and can interact with external tools and computational environments [40,41]. These characteristics give rise to a technical paradigm that differs substantially from conventional deep-learning-based approaches, yet existing reviews have rarely examined these distinctions in a dedicated and systematic manner. Third, the literature still lacks a comprehensive quantitative characterization of the field's overall development. Existing reviews generally emphasize methods and applications, while providing comparatively limited coverage of data resources, benchmarks, and evaluation frameworks. Moreover, the rapid pace of developments during 2025–2026 means that several earlier reviews do not fully capture

recent shifts toward reasoning-oriented and increasingly agentic LLM systems [29,42]. The emergence of reasoning models, agentic reinforcement learning, and the Model Context Protocol (MCP) [43] has begun to transform LLMs from auxiliary tools for GIS analysis into autonomous agents capable of executing complex spatial analysis workflows [44–46], extending beyond the taxonomies adopted in earlier reviews. Consequently, there remains a clear need for a systematic review of the LLM-GIS intersection that integrates technical architectures, data resources and evaluation benchmarks, and the latest advances in applications.

In 2025, a research team at Wuhan University first proposed a conceptual framework for LLM4GIS [47] and reviewed relevant studies published through October 2024. However, owing to constraints in publication timing and scope, the review did not capture subsequent advances in foundational architectures and tokenization mechanisms, emerging training paradigms. To address these gaps, this review extends the literature coverage through July 2026 and provides a systematic synthesis of current LLM4GIS research across four dimensions: technical approaches, datasets, applications, and evaluation.

The main contributions of this review are summarized as follows:

- We conducted a systematic statistical and quantitative analysis of 673 research articles published between 2022 and July 2026, tracing the evolution of LLM4GIS from sporadic early exploration to rapid expansion and identifying the major research hotspots and emerging frontiers in the field.
- We developed a taxonomy spanning four dimensions—technical approaches, datasets, applications, and evaluation—and systematically incorporated emerging directions that were largely absent from previous reviews, including model architecture design, spatial tokenization mechanisms, reinforcement learning, and agent-based methods. This taxonomy provides a more comprehensive view of the technical landscape and application spectrum of LLM4GIS.
- We systematically examined recent advances and representative studies across five core application areas: geospatial knowledge question answering, knowledge extraction, spatiotemporal reasoning, geospatial code generation, and analysis modeling. We further synthesized the evaluation methods and benchmarks used in each application area, providing a structured assessment of how current LLM4GIS capabilities are measured and compared.
- We identified the key bottlenecks constraining the further development of LLM4GIS from five perspectives: unified access to multimodal spatiotemporal data, development of high-quality data resources and standards, exploration of GIS-specific foundation models, integration of world models with real-world geospatial environments, and improvements in interpretability and trustworthiness. Based on these challenges, we outline directions for advancing GIS analysis from specialist-oriented tools toward broadly accessible infrastructure for spatiotemporal intelligence.

The overall structure of this review is illustrated in Figure 1. The remainder of this paper is organized as follows. Section 2 presents the conceptual framework and research trends of LLM4GIS. Section 3 reviews the technical approaches, and Section 4 summarizes the datasets. Section 5 examines the applications across five core scenarios, with Section 6 discussing the corresponding evaluation methods and benchmarks. Section 7 identifies key challenges and future research directions, and Section 8 concludes the paper.

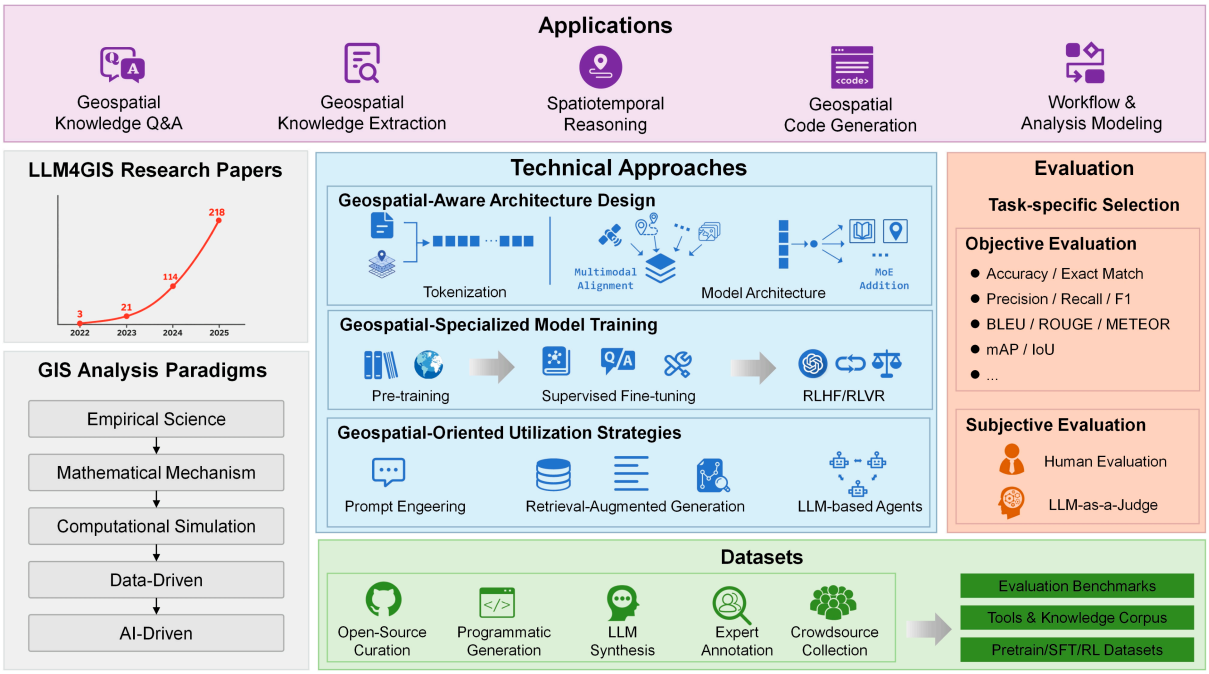


Figure 1. Overview of the Review.

2. LLM4GIS

2.1. Evolution of GIS Analysis Methods

The evolution of GIS analysis methods has been jointly driven by advances in technological tools and the changing demands of geographic research. To better understand the significance of introducing LLMs into GIS analysis and the technological transformations they bring, it is necessary to first trace the historical development of GIS analysis and examine how the technical limitations of each stage have driven successive methodological shifts, ultimately leading to a new generation of intelligent analysis centered on LLMs. As shown in Table 1, advances in information science and technology, together with evolving analytical demands, have progressively reshaped the analytical, data, task, and model characteristics of GIS analysis across five broadly defined stages [48].

Table 1. Evolution of GIS Analysis Methods.

Development Stage	Empirical Science	Mathematical Mechanism	Computational Simulation	Data-Driven	AI-Driven
Indicative Period	Before the 1950s	1950s–1960s	1970s–1990s	2000s–early 2020s	Since 2022
Analysis Methods	Map reading	Quantitative geographic analysis	Tool-based spatial modeling	Spatiotemporal big data analysis	Intelligent and automated analysis
Data	Small-scale observational data	Structured data	Unstructured data	Heterogeneous data	Multimodal big data
Requirements	Qualitative	Quantitative	Location, topology	Complex geographical process modeling	Dynamic decision-making and adaptation
Models	/	Statistical models	Spatial analysis models	Data-driven models	Generative large models

Note: The periods indicate the approximate emergence and growing influence of each paradigm rather than strict boundaries; these paradigms overlap and continue to coexist.

The Empirical Science stage laid the foundations for geospatial analysis through field observation, manual mapping, and experience-based analysis [49]. John Snow's 1854 cholera investigation, which linked the spatial distribution of reported cases to a

contaminated water source, exemplified the early use of spatial overlay to understand geographic patterns and address real-world problems [50].

With the rise of quantitative geography, GIS analysis entered the Mathematical Mechanism stage, applying mathematical and statistical models to investigate quantitative relationships among spatial phenomena [51]. Methods such as spatial interpolation and spatial clustering supported the estimation and identification of geographic patterns [52,53], marking a shift from qualitative description toward quantitative explanation.

Advances in computing enabled the Computational Simulation stage and established a tool-oriented framework for GIS analysis [54]. Classical geometric and topological methods, including buffer and overlay analysis, supported applications such as flood-inundation delineation and public-facility siting [55,56], strengthening the role of digital simulation in spatial decision-making.

In the Data-Driven stage, the accumulation of spatiotemporal big data shifted GIS analysis toward machine learning and deep learning for modeling complex nonlinear patterns [6,57]. Through automated feature learning, these methods have been widely applied to data-intensive tasks such as traffic-flow prediction and urban functional-zone identification [58,59].

As geospatial data types and application requirements become increasingly complex, data-driven approaches centered on task-specific models face growing challenges in multimodal data understanding, adaptation to open-ended tasks, and generalization across scenarios. Meanwhile, complex GIS toolchains and high requirements for domain expertise continue to limit the broader accessibility of spatial analysis. Consequently, GIS analysis is evolving from a Data-Driven paradigm focused primarily on feature learning and predictive modeling toward an AI-Driven paradigm that emphasizes semantic understanding, reasoning, and automated task execution.

2.2. Concept of LLM4GIS

Built on the Transformer architecture, LLMs leverage self-attention mechanisms to capture long-range dependencies and acquire strong semantic understanding and broad knowledge through pretraining on massive corpora. These capabilities enable them to extract structured logic from unstructured information and perform GIS analysis tasks through reasoning and tool use, thereby providing a new pathway for addressing the limitations outlined above.

LLM4GIS is not merely a technical combination of "LLM + GIS"; rather, it represents an AI-Driven paradigm of GIS analysis in which LLMs serve as the core reasoning engine for restructuring the perception, understanding, analysis, and decision-making processes associated with geospatial data. At the data-understanding level, LLMs leverage multimodal alignment and semantic understanding to interpret unstructured text, imagery, and heterogeneous spatial data, while efficiently extracting geographic entities and their implicit relationships. At the intent-understanding level, LLMs can interpret ambiguous user requirements expressed in natural language and translate them into standardized technical analysis frameworks or executable workflows. At the generalization and transfer level, LLMs build on the broad knowledge acquired during pretraining and can adapt rapidly to diverse downstream applications through limited domain-specific knowledge injection or prompt engineering, without requiring fundamental model redesign or large-scale retraining. This substantially improves their adaptability across tasks. At the application level, LLMs can lower the programming barrier in code-based GIS by generating analytical scripts and workflows. When integrated into established GIS platforms, they can further support tool selection, parameter configuration, and workflow execution through natural-

language interaction [10,60], thereby enabling researchers from other disciplines to use GIS tools more readily for integrated problem solving.

In summary, LLM4GIS deeply integrates the cognitive and reasoning capabilities of LLMs with the domain-specific analytical capabilities of GIS, providing a new pathway for overcoming the limitations of existing analysis methods. This integration marks the transition of GIS analysis toward a new AI-Driven stage of spatiotemporal intelligence, characterized by natural-language interaction, logical reasoning, and increasingly automated analytical workflows. Such intelligence is reflected in the capacity of models to autonomously perceive multimodal geospatial data, understand spatiotemporal semantics, reason about spatial relationships and spatiotemporal processes, and support decision-making through spatiotemporal analysis to solve real-world problems.

2.3. Research Development and Trends

To systematically characterize the current research landscape and its evolving trends, we first considered the particular challenges associated with literature retrieval at the intersection of LLMs and GIS. In the broader AI literature, terms such as "spatial" and "geo" are frequently used to describe tasks unrelated to GIS, including indoor robot navigation, video understanding, and medical image analysis. Direct keyword-based retrieval therefore introduces substantial semantic drift and a large volume of irrelevant records. Conversely, restricting the search to domain-specific terms such as "geospatial" risks overlooking emerging studies on GIS analysis and spatiotemporal reasoning published in major AI venues without explicitly using this terminology.

To balance the coverage and relevance of literature identification, this study adopted a multi-source, iterative, manual literature identification and screening process. Literature identification was jointly conducted by five researchers with backgrounds in geographic information science, spatial intelligence, or large language model research, covering the period from 1 January 2022 to 31 July 2026. The research team used ("large language model" OR "LLM") AND ("GIS" OR "geo" OR "spatial") as the core search expression and adjusted it according to the syntax rules of different platforms. Candidate studies were identified through Web of Science Core Collection, Scopus, Google Scholar, and arXiv. Subsequently, forward and backward citation tracking of candidate studies was conducted to further identify potentially missed literature. All candidate records were consolidated and subjected to information normalization, version verification, and deduplication. For records with identical or highly similar titles, overlapping authors, the research team verified their publication status and version history. Preprint and formally published versions of the same study were merged into a single record, with preference given to the formally published version.

Each candidate record was manually assessed by the researchers according to uniform criteria. In this study, LLM4GIS research included in the subsequent statistical analysis was defined as work that simultaneously satisfied the following three conditions: first, an LLM, a multimodal LLM (MLLM), or an LLM-centered agent constituted a key reasoning, generation, or interaction component of the study; second, the research object involved geographically referenced spatial data at the Earth-surface, urban, or regional scale, geographic processes, or explicit GIS analysis tasks; third, the main contribution directly served GIS analysis, geospatial understanding, or geographic decision-making. Studies concerning only non-geospatial tasks, such as general spatial cognition, indoor embodied navigation, robotic manipulation, or planar geometry, as well as spatiotemporal foundation models that did not use an LLM, MLLM, or LLM-centered agent as a key technical foundation, were not included in the subsequent statistical analysis. Each candidate record underwent cross-review by research team members. For records with unclear boundaries,

disputed task attributes, or uncertain version relationships, multiple members discussed the record and reached a consensus in accordance with the operational criteria described above. Reviews, opinion papers, and other non-research publications were used only to identify relevant research leads and were not included in the subsequent statistical analysis. Ultimately, this study included 673 LLM4GIS-related research papers, which formed the basis for the subsequent statistical and content analyses.

Figure 2 presents the annual number of publications and the 20 most prolific publication venues. To avoid double counting the same study as both a preprint and a formal publication, the analysis uses the study as the unit of counting. The 134 studies that were first released on arXiv and subsequently published in a journal or conference are counted under their corresponding formal publication venues, whereas the 118 studies that remained available only as preprints by 31 July 2026 are retained under arXiv. This approach preserves arXiv as an early dissemination channel while ensuring that each study is counted only once when comparing publication venues. The annual publication trend reveals rapid growth in LLM4GIS research, with the number of studies increasing sharply over the review period. In 2022, three relevant studies were identified, indicating that the field remained at an exploratory stage. Following the release of ChatGPT, publication output rose to 21 studies in 2023 and 114 in 2024, accompanied by a shift from basic geospatial capability assessment toward domain-specific model optimization and multimodal processing. In 2025, the number reached 317, while multimodal and reasoning-oriented models such as GPT-4o and DeepSeek-R1 expanded research toward task planning, code generation, tool use, and more complex geospatial applications. By July 2026, 218 studies had already been included, exceeding the total output of 2024 despite covering only the first seven months of the year, confirming the continued rapid growth of the field. Overall, LLMs are evolving from auxiliary tools for GIS analysis toward core components of AI-driven geospatial analysis.

In terms of publication venues, arXiv is the single largest venue category, with 118 studies that remained available only as preprints at the review cutoff. Together with the 134 studies initially released on arXiv but subsequently assigned to their formal journal or conference venues, this pattern indicates that arXiv remains an important channel for the early dissemination of LLM4GIS research. Preprint platforms have therefore become an important channel for the rapid dissemination of emerging methods and findings. Notably, although preprints have not undergone formal peer review, several have already attracted substantial scholarly attention. Despite lacking formal peer review, LLM-Mob [61] and GPT4GEO [62], both released in 2023, as well as ClimateGPT [63], released in 2024, had each received more than 100 citations by 2 August 2026. Among formally published studies, the literature is distributed across both leading computer science conferences and established GIS journals. Top-tier conferences in machine learning and natural language processing, including ACL (23), NeurIPS (12), and EMNLP (10), have contributed numerous AI-oriented solutions to geospatial problems, suggesting that spatial understanding and reasoning are increasingly regarded as important dimensions for evaluating and advancing the general intelligence of LLMs. The presence of CVPR (7) and ICCV (7) further reflects the growing importance of multimodal geospatial research. Within the GIS community, ACM SIGSPATIAL (39) ranks first among formal publication venues, reflecting the close connection between computing and geospatial science. Established GIS journals, including ISPRS IJGI (25), IJGIS (25), and IJDE (21), follow closely and primarily provide venues for more systematic and rigorous theoretical validation and domain-specific applications.

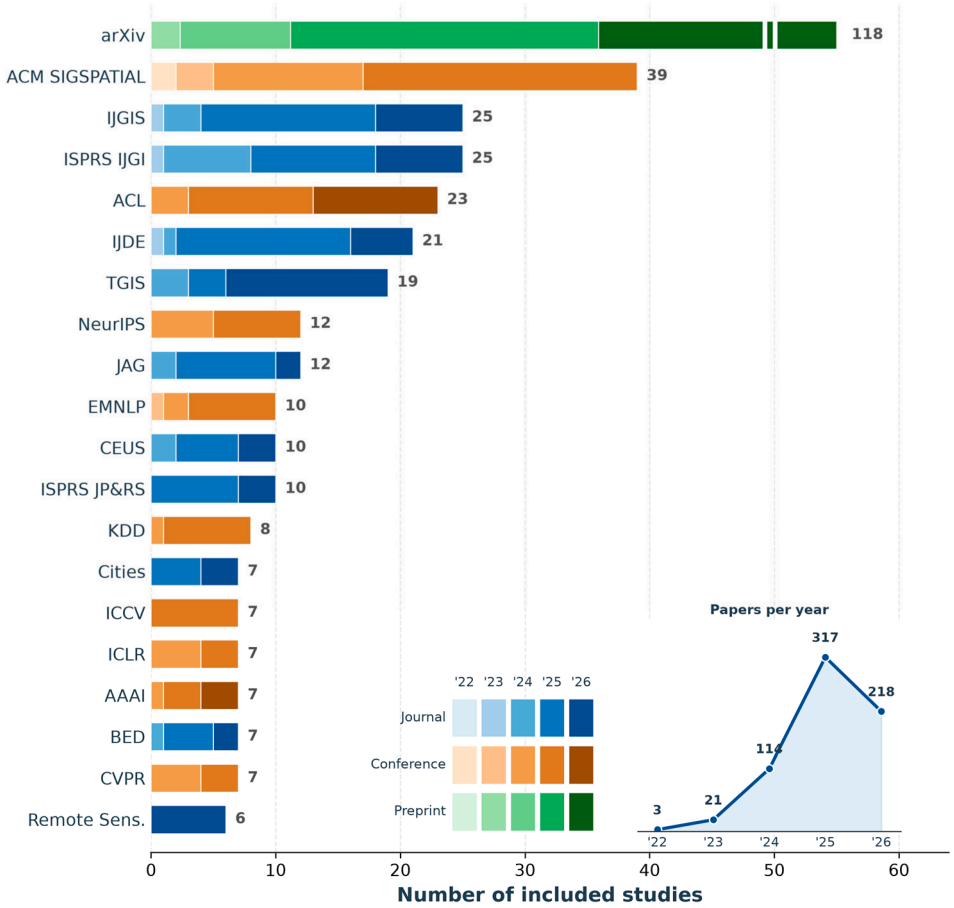


Figure 2. Number of papers by publication year and the top 20 publication venues. The arXiv bar is displayed with an axis break to facilitate comparison with the remaining venues. Color hue indicates publication type, and colour shade indicates publication year. Common venue abbreviations and their full names are provided in Appendix A.

Figure 3 presents the geographical distribution of the included studies according to the country or region of the first author's affiliation. The results show a clear, although not exclusive, geographical concentration in current LLM4GIS research. China ranks first with 317 studies, accounting for 47.1% of the included literature, followed by the United States with 158 studies. Germany, the United Kingdom, Japan, and Australia form a second group of active contributors. Overall, the distribution indicates that the development of LLM4GIS is currently driven mainly by research communities in East Asia and North America, while Europe, Oceania, and other regions also contribute to the continuing expansion of the field.

To identify the technical approaches used in each included study, we used DeepSeek-V4-Pro to examine the full texts of the papers. All extracted results were subsequently checked and verified against the original publications to ensure the accuracy of the final labels. Because a study may employ multiple technical approaches, it may be assigned to more than one methodological category.

Figure 4 presents the resulting distribution of method use across the reviewed studies. Prompt engineering remained the dominant approach during 2024–2026, accounting for 43.7%, 43.1%, and 39.6% of method assignments, respectively. The share of retrieval-augmented generation (RAG) increased from 3.4% in 2023 to 14.5% in 2025, while the share of agents rose from 6.9% in 2023 to 15% in 2025 and 19.0% in the first seven months of 2026. SFT remained relatively stable at approximately 17–18% of method assignments between 2023 and 2026. Architecture-oriented studies grew from 1 paper in 2023 to 11 in 2025, with a further 21 published in the first seven months of 2026. Research on tokenization

mechanisms similarly emerged from virtually none to 11 papers in 2025 and 8 in the first seven months of 2026. These trends reflect growing attention to geospatially adapted representation and model design. The number of studies using reinforcement learning increased from 22 studies in 2025 to 24 studies in the first seven months of 2026. Overall, LLM4GIS is evolving from an application paradigm dominated by prompt engineering toward a broader methodological ecosystem involving retrieval augmentation, agents, model adaptation, architectural innovation, and reinforcement learning.

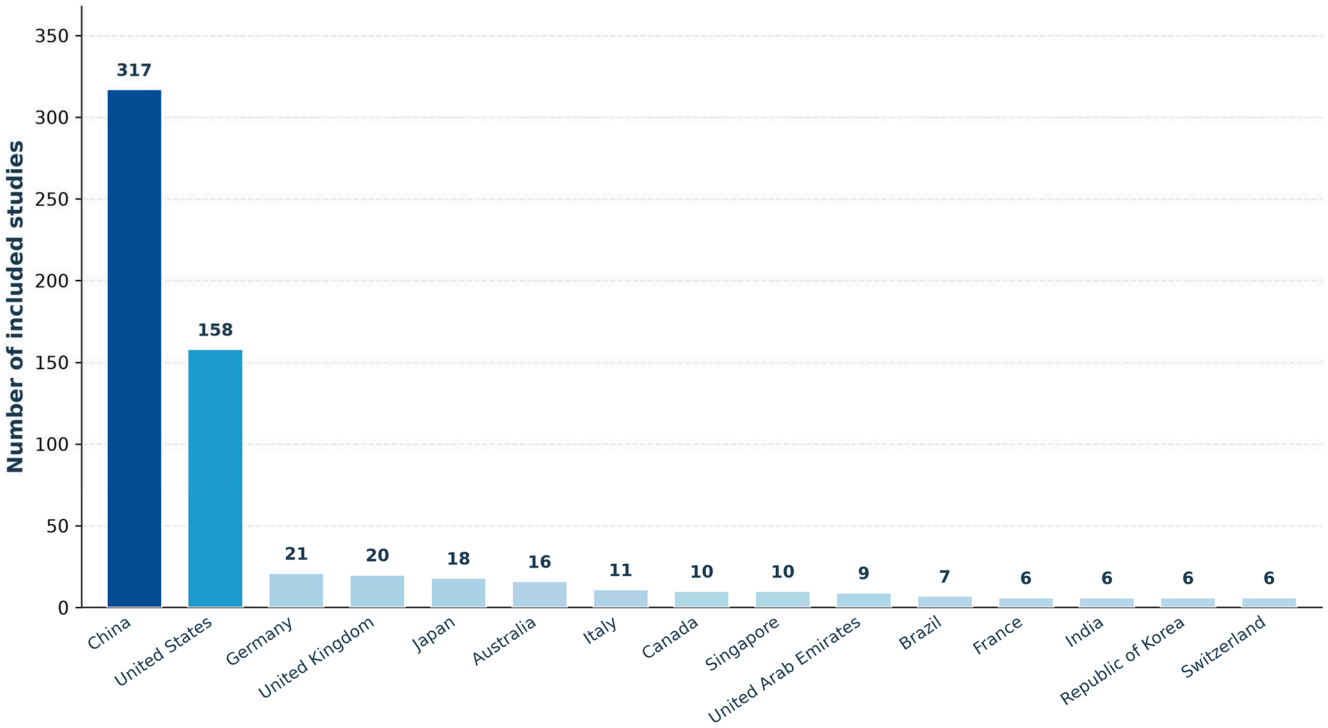


Figure 3. Geographical Distribution of the Top 15 Countries/Regions of Included LLM4GIS papers by First-Author Affiliation.

Figure 5 presents the distribution of keywords across the reviewed literature. Among them, "Spatial Reasoning" occupies a central position, indicating that current research is moving beyond the retrieval or memorization of basic geographic knowledge and factual information toward dynamic reasoning over spatial relationships and geographic processes. The high frequency of "Geospatial Code Generation" further highlights the importance of code as a direct means through which LLMs perform spatial analysis. Compared with conventional GUI-based GIS operations, code offers greater precision, reusability, and capacity for structurally representing complex analytical logic, thereby serving as an effective bridge between natural-language intentions and underlying computational engines. In terms of specific applications, "Human Mobility" "Trajectory Prediction" "POI Recommendation" and "Next-Location Prediction" appear with similar frequencies and form a distinct thematic cluster, suggesting that human mobility modeling is currently one of the most extensively studied application areas. "Image Geolocalization" and "Earth Observation" also occur frequently, indicating another major concentration of research on extracting geographic information from visual data such as satellite imagery and street-view images. This trend demonstrates that LLM4GIS is expanding beyond text-based tasks toward increasingly multimodal applications involving imagery and other geospatial data modalities.

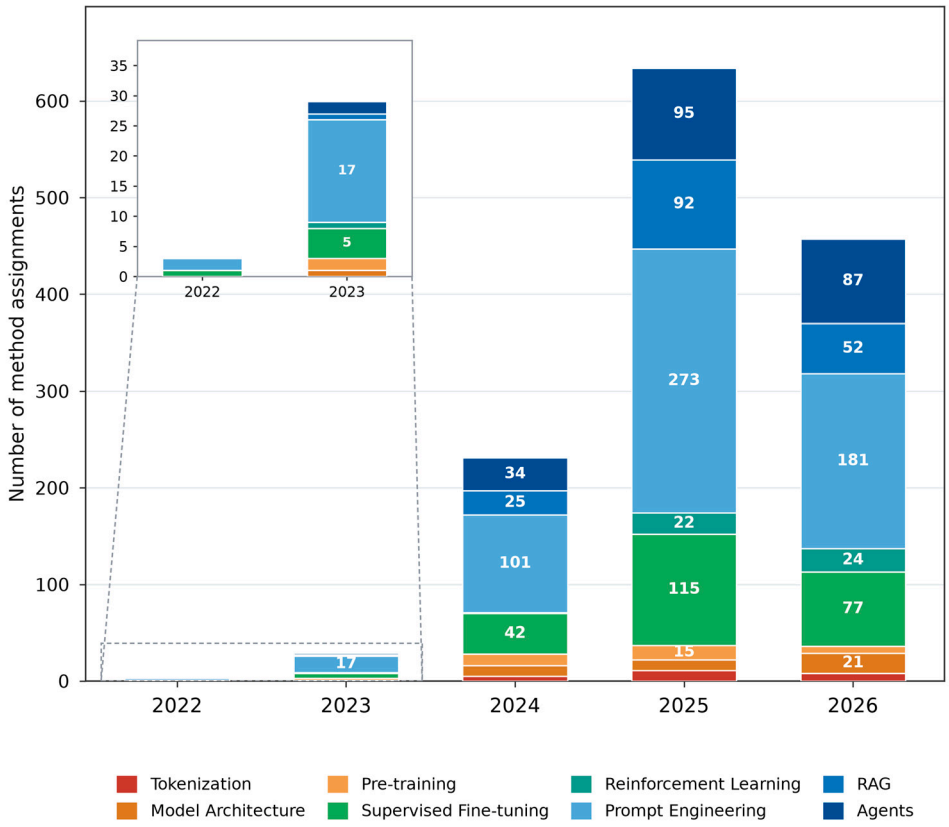


Figure 4. Distribution of Method Usage in LLM4GIS Research. Percentages reported in the text are calculated using the total number of method assignments across the eight categories in each year as the denominator.



Figure 5. Word Cloud of Keywords in LLM4GIS Research Papers.

3. Technical Approaches in LLM4GIS

Unlike general-domain applications, applying LLMs to GIS requires domain-specific knowledge and the ability to process the distinctive forms of geospatial information. On the one hand, GIS analysis involves not only textual information but also diverse spatial data modalities, including trajectories, imagery, and vector data. On the other hand, effective

geospatial analysis depends on domain-specific concepts and properties, such as spatial topological relationships, spatiotemporal dependencies, and spatial heterogeneity. To better adapt LLMs to GIS tasks, existing studies have explored three major technical directions. The first is Geospatial-Aware Architecture Design, which optimizes fundamental mechanisms such as tokenization and model architectures to improve the representation and understanding of geospatial data. The second is Geospatial-Specialized Model Training, which adapts models to specific GIS tasks through approaches including Pretraining, Supervised Fine-tuning, and Reinforcement Learning. The third is Geospatial-Oriented Utilization Strategies, which leave the underlying model unchanged but introduce task-oriented strategies at the application stage to improve its effectiveness for GIS tasks. Accordingly, this section reviews the technical approaches to LLM4GIS from these three perspectives. Figure 6 provides an overview of these technical approaches along with representative studies in each category. Detailed bibliographic information for the studies shown in Figure 6 is provided in Appendix B, Table A2.

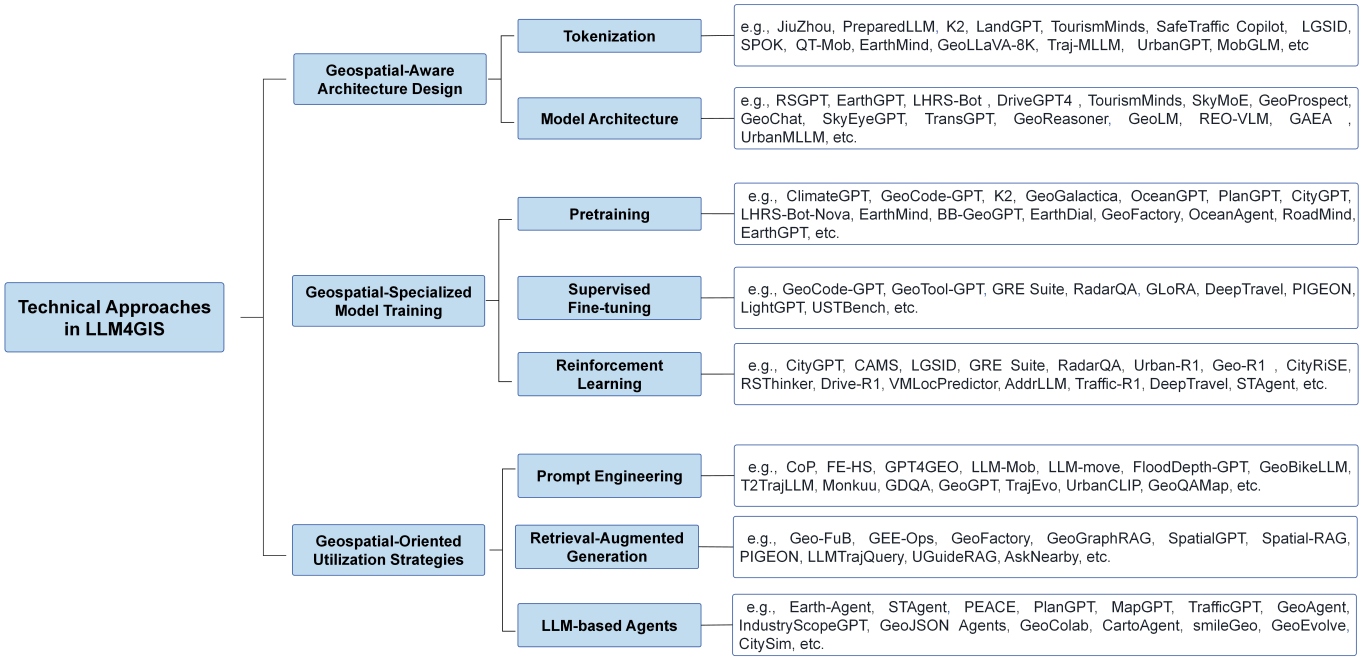


Figure 6. Overview of Technical Approaches.

3.1. Geospatial-Aware Architecture Design

3.1.1. Tokenization

Tokenization is a fundamental step that discretizes continuous natural-language text and geospatial information into tokens that can be processed by LLMs. In LLM4GIS, general-purpose text tokenizers often fragment geospatial entities and relationships, potentially resulting in the loss of spatial semantics. To address this limitation, recent studies have explored GIS-oriented tokenization strategies, primarily along two directions: domain-specific expansion of textual vocabularies and spatial–semantic tokenization of non-textual geospatial entities.

(1) Expansion and Optimization of GIS-Specific Vocabularies.

General-purpose LLM tokenizers, such as those based on Byte Pair Encoding (BPE), are typically trained on general-domain corpora in which GIS-specific content is relatively sparse. Consequently, they often exhibit substantial limitations when processing GIS-related text. For example, continuous coordinate values may be fragmented into semantically meaningless numerical tokens, while domain-specific terminology may be split into unrelated subword units, resulting in the fragmentation and loss of geospatial

semantics. To address this issue, studies such as JiuZhou [64] and PreparedLLM [65] retrained tokenizers on large-scale GIS-specific corpora and merged the resulting domain vocabularies with those of the base models, enabling coordinate values and GIS terminology to be represented more coherently as tokens. In addition, K2 [66] introduced structured tokens such as "[START_FIGURE]" to facilitate the interpretation of complex figures, maps, and other visual elements in the geoscientific literature. GEEToker [67] further developed a plug-and-play tokenizer specifically for Google Earth Engine (GEE) code, with the original study reporting improved performance in geospatial code generation.

(2) Tokenization of Geospatial Features.

Geospatial data contain complex topological, visual, and temporal characteristics, and directly converting them into ordinary text sequences may result in the loss of implicit spatial structure and relationships. To preserve such information, researchers have developed specialized tokens and deep learning-based encoders that transform geospatial features into discrete representations carrying explicit spatial semantics.

For discrete geographic entities and objects, a common strategy is to represent them explicitly using special tokens. To preserve categorical and state-related information associated with geographic entities, LandGPT [68] and TourismMinds [69], for example, introduced special identifiers such as <POI> and <density>, while SafeTraffic Copilot [70] designed customized tokens for accident severity levels. Such explicit token representations provide clear semantic anchors, helping the model maintain attention to and alignment with geographic entity attributes during long-form generation.

For continuous spatial characteristics, such as road-network distance and spatial proximity, the central challenge is to map spatial relationships into discrete token sequences. LGSID [71] and QT-Mob [72], for example, employ Semantic IDs (SIDs) generated through Vector Quantization (VQ) to incorporate location and spatial proximity information into token representations, while SPOK [73] encodes the relative positions of road-delimited parcels within the road network. This enables models to capture spatial proximity and relative positional relationships among locations to a certain extent when processing sequential data. As a methodological reference outside the statistical corpus, UrbanFM [74], a non-LLM spatiotemporal foundation model, proposes the MiniST tokenization strategy, which applies KD-Tree-based greedy capacity-constrained clustering to spatial nodes and combines the resulting clusters with temporal windows to form spatiotemporal patches, thereby unifying heterogeneous sensor and grid data into learnable units analogous to NLP tokens. MoveFM-R [75], in contrast, adopts language-guided semantic tokenization to transform GPS trajectories into mobility-semantic units that can be interpreted by LLMs.

For visual and pixel-level information, tokenization methods focus on efficiently compressing pixel arrays into visual tokens. EarthMind [76] introduced a <SEG> token to support pixel-level localization, while GeoLLaVA-8K [77] employed background pruning and anchor selection to compress ultra-high-resolution remote sensing imagery. Traj-MLLM [78] further transformed trajectory data into both descriptive text tokens and visual tokens. These approaches bridge visual perception and language understanding, enabling LLMs to interpret maps and remote sensing imagery in a token-based manner and thereby support multimodal question answering and spatial reasoning.

Tokenization serves as a bridge between the geographic world and the representation space of language models, and existing research has begun to extend this mapping from text to a broad range of multimodal geospatial features. However, discretizing continuous geospatial information inevitably introduces some information loss; for example, quantization reduces positional precision, and token boundaries may separate spatially adjacent locations. Future work may therefore focus on developing efficient unified geospatial tokenizers capable of accommodating heterogeneous data modalities, or on exploring end-

to-end modality alignment methods that avoid explicit Tokenization altogether, thereby minimizing semantic loss introduced during discretization and quantization.

3.1.2. Model Architecture

Model Architecture selection and optimization constitute a fundamental component of LLM4GIS. Given the multimodal heterogeneity, domain specificity, and logical complexity of GIS data, existing studies have primarily adopted two architectural strategies: external multimodal alignment and internal domain-specific adaptation, as illustrated in Figure 7.

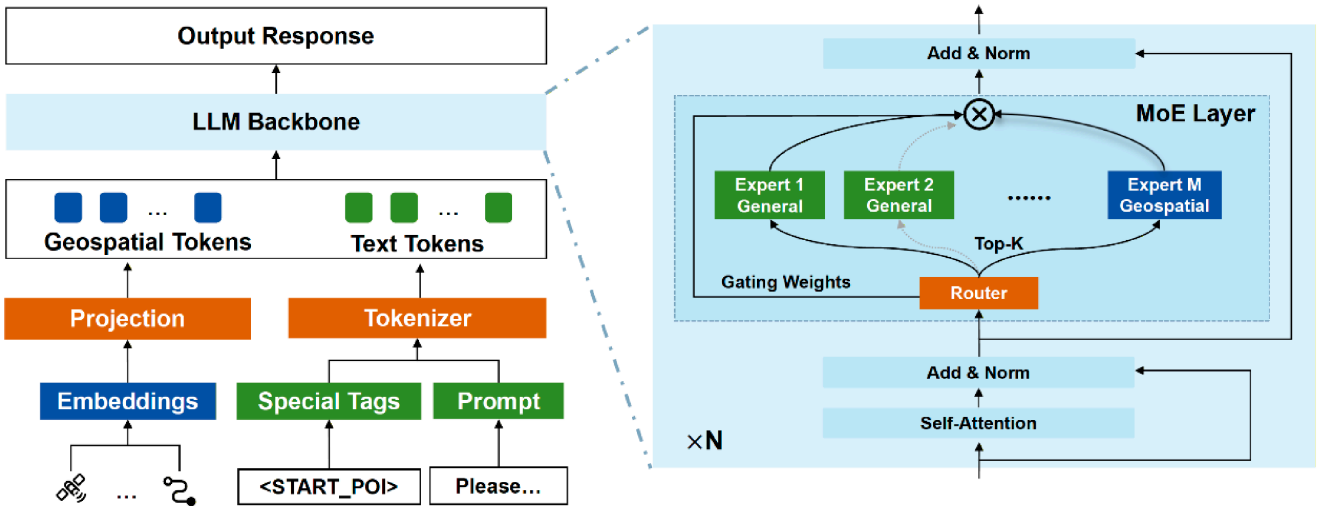


Figure 7. Typical LLM4GIS architecture and tokenization.

At the multimodal adaptation level, GIS tasks frequently involve non-textual modalities such as remote sensing imagery, vector maps, and spatiotemporal trajectories, among which complex spatial correspondences must be preserved. A common architectural strategy is to keep the LLM language backbone frozen while introducing specialized adapters to align geospatial features with the language semantic space. Combined with modality-specific encoders, this design projects heterogeneous geospatial information into a shared representation space that can be jointly processed with text. This architecture has been widely adopted in remote sensing image interpretation, with representative studies such as RSGPT [79], EarthGPT [80], and LHRS-Bot [81], which extend language models with capabilities for understanding and analyzing remote sensing imagery. For dynamic spatiotemporal streams, such as autonomous-driving video and mobility trajectories, DriveGPT4 [82] and TourismMinds [69] designed projection modules that explicitly incorporate temporal information. By transforming continuous spatiotemporal states into sequences enriched with timestamps and location semantics, these models can capture movement trends and spatiotemporal co-occurrence relationships among geographic objects. The main advantage of this external modular architecture is that it enables the perception and understanding of geospatial data at relatively low training cost while preserving the general knowledge encoded in the backbone model.

At the level of internal architectural modification, research focuses on improving domain adaptation to GIS tasks at a fundamental level while balancing GIS-specific expertise against general reasoning capabilities. To mitigate catastrophic forgetting, whereby extensive injection of GIS knowledge or domain-specific fine-tuning degrades a model's general capabilities, several studies have introduced Mixture-of-Experts (MoE) mechanisms in which different experts process tokens associated with different domains. This design enables more fine-grained specialization while retaining general-purpose competence. For example, SkyMoE [83] and GeoProspect [84] replace the conventional feed-forward net-

works in Transformer blocks with MoE modules and employ dynamic routing mechanisms. Depending on the characteristics of each input token, the model can automatically route it to either general-purpose experts or GIS-specific experts. Such structural decoupling at the architectural level allows domain-specific GIS capabilities to coexist with general reasoning ability.

Overall, the evolution of Model Architecture in LLM4GIS shows a clear trend toward multimodal integration and modular design. Existing architectures have made substantial progress in enabling models to perceive and interpret geospatial data, but remain limited in their ability to internalize deeper spatial structures and reasoning principles. Future advances may therefore depend on the development of spatially native model architectures and attention mechanisms that incorporate multimodal fusion and spatial concept modeling directly into the model architecture and training process, rather than treating geospatial information merely as an additional input modality. Such designs could enable models to represent spatial structures, such as topology, scale, and spatial dependence, more explicitly and to reason over them more reliably.

3.2. Geospatial-Specialized Model Training

For general-purpose foundation LLMs or LLMs that have undergone architectural adaptation, existing studies primarily employ three training paradigms to optimize model parameters for GIS-specific tasks: Pretraining, including continual pretraining; Supervised Fine-tuning; and Reinforcement Learning. These paradigms can be adopted individually or in combination depending on task requirements. These training strategies are illustrated in Figure 8.

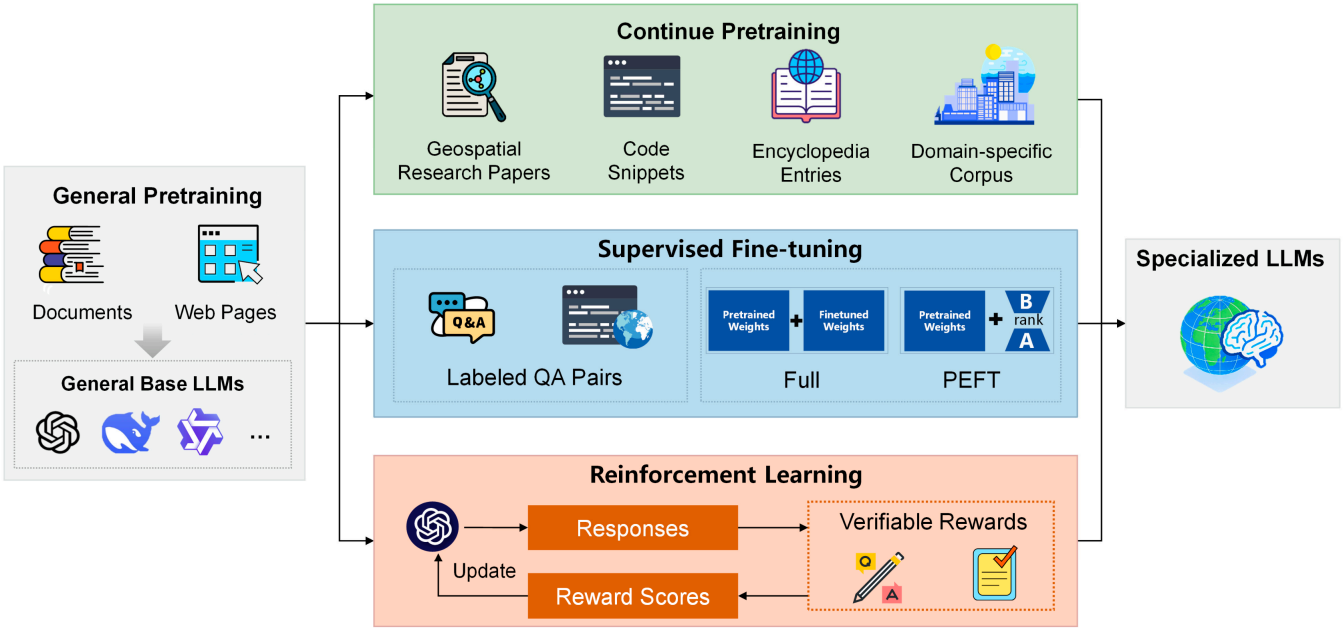


Figure 8. Overview of model pretraining (PT), supervised fine-tuning (SFT), and reinforcement learning (RL). Here, PEFT denotes parameter-efficient fine-tuning.

3.2.1. Pretraining

The Pretraining stage aims to endow models with broad geospatial domain knowledge. Depending on the training starting point, existing approaches can be broadly divided into pretraining from scratch and continual pretraining, which represent different trade-offs between computational cost and the degree of domain adaptation.

Although pretraining from scratch can, in principle, maximize domain adaptation and establish a geospatially oriented knowledge representation from the outset, its practical

adoption is severely constrained by the enormous computational cost and the scarcity of high-quality GIS-specific data. ClimateGPT [63] represents one of the few notable attempts in this direction. It was trained for two months on 256 H100 GPUs using a corpus comprising approximately 300B tokens of general-domain data and 4.2B tokens of climate-related data. However, compared with the more than 2T tokens used to pretrain general-purpose foundation models such as LLaMA-2, the relative scarcity of domain-specific data limits the extent to which such models can be sufficiently trained. Empirical results suggest that, because of the substantial gap in both data scale and quality, models pretrained entirely from scratch may ultimately underperform approaches that adapt stronger general-purpose foundation models to the target domain.

Given the high barriers associated with training from scratch, continual pretraining, also referred to as mid-training, has become the dominant strategy in current research. This approach performs continued unsupervised training on existing foundation models using large volumes of unlabeled GIS-domain data, thereby injecting geospatial knowledge into the model while retaining the capabilities of the original backbone. In practice, GeoCode-GPT-7B [85] builds on Code Llama-7B [86] and uses 286K geospatial code samples to strengthen the model's ability to understand and generate geocomputational code. K2 [66], based on LLaMA-7B, further enhances domain knowledge through continual pretraining on 5.5B tokens of geoscience text collected from encyclopedic documents and academic publications. To mitigate catastrophic forgetting during continual pretraining, some studies adopt mixed-corpus strategies. GeoGalactica [87], for example, uses an 8:2 mixture of general-domain and domain-specific data, with a total training scale of 78B tokens, to balance domain specialization and general capabilities. A growing number of studies have also applied continual pretraining to specific geospatial application domains, including OceanGPT [88] for ocean science, PlanGPT [89] for urban planning, CityGPT [90] for urban scene understanding, and LHRS-Bot-Nova [91] for remote sensing image understanding. PostGISer-7B [92], built on Qwen2.5-7B, performs autoregressive unsupervised pretraining on the official PostGIS documentation and represents one of the relatively few efforts targeting spatial databases rather than imagery or urban text. Some studies further incorporate curriculum learning. EarthMind [76], for instance, adopts a three-stage progressive training strategy that proceeds from visual question answering on natural images, to optical remote sensing adaptation, and finally to multisensor data fusion. By progressively increasing task complexity, this strategy improves the model's perception and reasoning capabilities for heterogeneous multi-source remote sensing data.

Pretraining provides the primary foundation for the knowledge encoded in LLMs, and existing studies have demonstrated the effectiveness of continual pretraining on strong general-purpose models for geospatial domain adaptation. Future progress is likely to depend on more targeted data supplementation, including the careful selection or synthesis of high-quality data for underrepresented scenarios, together with more effective corpus-mixing strategies to further balance general capabilities with domain specialization.

3.2.2. Supervised Fine-Tuning

SFT is a key stage for aligning pretrained models with specific GIS tasks. It optimizes the parameters of general-purpose LLMs or continually pretrained models using labeled instruction data, with the aim of reinforcing domain knowledge, standardizing task-solving behavior, and enabling models to execute complex tasks in accordance with GIS-specific analytical logic.

At the methodological level, Parameter-Efficient Fine-tuning (PEFT) has become the dominant approach because of its substantially lower resource requirements than full-parameter fine-tuning. GeoCode-GPT-7B [85] and GeoTool-GPT [93], for example,

employ Low-Rank Adaptation (LoRA), allowing the models to acquire capabilities in geospatial code generation and GIS toolchain invocation by updating only a small fraction of the parameters. For multimodal tasks, GRE [94] and RadarQA [95] also adopt LoRA-based strategies for image localization and weather radar interpretation. Beyond standard PEFT, several studies have proposed domain-adaptive fine-tuning strategies tailored to GIS tasks. GloRA [96] introduces a GIS knowledge-aware hierarchical adaptation mechanism that dynamically allocates parameter weights according to task requirements. CityGPT [97] employs Self-Weighted Fine-Tuning (SWFT) to adjust the loss function dynamically, strengthening urban-domain knowledge while mitigating the forgetting of general knowledge. Study [98], targeting emergency dispatch scenarios, further designs an entropy-constrained loss function to balance the stability of higher model layers with the adaptability of lower layers.

In terms of training objectives, SFT is gradually evolving from simple task-specific mapping toward the joint development of lightweight deployment, deeper reasoning, and agentic capabilities. Knowledge distillation provides an effective means of transferring the capabilities of stronger models to smaller models through imitation-based learning. For example, PIGEON [99], LightGPT [100], and USTBench [101] use synthetic data generated by GPT-4o-mini or DeepSeek-R1 [102] to perform end-to-end fine-tuning of Qwen-family models, substantially improving performance in user interaction, traffic control, and urban spatiotemporal reasoning. In industrial settings, AddrLLM [103] performs Supervised Fine-tuning of Qwen-7B using more than 60 million real-world address logs from JD.com's location-based services system. This study demonstrates the feasibility of SFT in high-throughput and highly standardized scenarios such as e-commerce logistics and represents one of the few large-scale industrial cases based on real operational data. To enhance deep logical reasoning, RSThinker [104] uses Chain-of-Thought (CoT) data structured around "planning–localization–synthesis" to teach GLM-4.1V deeper reasoning patterns for remote sensing interpretation. Drive-R1 [105] employs reflective Chain-of-Thought data to train a "fast–slow thinking" strategy, allowing the model to adaptively select the depth of reasoning according to driving-scene complexity. For agentic capability construction, GeoAgent [106] and STAgent [107] use fine-tuning to significantly improve GIS tool invocation and multi-step reasoning capabilities. Going further, studies such as DeepTravel [108] also use distillation as a cold-start strategy for subsequent Reinforcement Learning.

Existing studies have demonstrated that SFT can effectively equip models with a wide range of downstream GIS capabilities, while PEFT substantially reduces training costs and domain-adaptive fine-tuning further improves performance on specialized GIS tasks. As research moves toward more complex spatiotemporal reasoning and autonomous agent tasks, future work is likely to place greater emphasis on constructing high-quality GIS instruction datasets that include intermediate reasoning steps, error reflection, and iterative correction, thereby fostering deeper and more reliable reasoning capabilities.

3.2.3. Reinforcement Learning

To address two core challenges—human preference alignment and the enhancement of complex spatiotemporal reasoning—Reinforcement Learning (RL) [109] and related preference optimization methods have been introduced into model training. By using reward signals or preference feedback to guide policy optimization, RL provides an effective means of further improving model performance.

At the preference-alignment level, Direct Preference Optimization (DPO) [110] has been widely adopted because of its relatively stable training behavior. Derived from the same KL-constrained reward-maximization objective as reinforcement learning from human feedback (RLHF), DPO optimizes the model directly on offline preference pairs,

without training an explicit reward model or performing online RL. CityGPT [90] and CAMS [111], for example, apply DPO to urban question answering and trajectory generation, respectively, improving the usability and logical coherence of model outputs. More recent studies have further adapted preference optimization to geospatial tasks. LGSID [71], for instance, proposes G-DPO, which converts physical spatial-proximity constraints into preference pairs, thereby translating abstract geographic rules into preference signals that can be directly optimized by the model. This design effectively strengthens the model's ability to capture geospatial regularities during reasoning.

With the emergence of reasoning models such as DeepSeek-R1, Reinforcement Learning with Verifiable Rewards (RLVR) [112], which optimizes the policy through online sampling with rewards computed by verifiable rules, has become an important paradigm for enhancing complex reasoning capabilities in LLMs. Current RL research in LLM4GIS largely builds on established algorithms such as Group Relative Policy Optimization (GRPO) [113], while methodological innovation is increasingly concentrated on reward engineering tailored to the characteristics of geospatial tasks. Basic reward functions typically focus on output correctness and format compliance. GRE [94], RadarQA [95], and Urban-R1 [114], for example, use answer accuracy and response-format validity as primary reward signals. To improve the quality of the reasoning process itself, Geo-R1 [115] introduces a conciseness reward to encourage more efficient reasoning, whereas CityRiSE [116] employs keyword-based rewards to strengthen the model's attention to task-critical information. For highly specialized GIS tasks, more fine-grained domain-specific metrics have been incorporated into reward functions. RSThinker [104], for instance, supplements visual question-answering accuracy with the mean Average Precision (mAP) metric for object detection, while Drive-R1 [105] combines trajectory similarity with rewards for meta-level driving behaviors. To alleviate reward sparsity in complex reasoning tasks, some studies further introduce process-level supervision. VMLocPredictor [117] combines intermediate step rewards with distance-based outcome rewards, whereas AddrLLM [103] constructs stepwise rewards for address rewriting based on semantic similarity, reverse-geocoding semantic scores, and geocoding scores. Notably, a substantial proportion of RL studies published in 2026 use image geolocalization as a primary testbed. Geo-R [118], GeoZero [119], Geo-R1 [115], and Geo-ADAPT [120] all adopt GRPO or its variants to improve geolocation accuracy. Because geolocation tasks naturally provide verifiable geographic distance errors as reward signals, they have become a representative setting for evaluating the applicability of RLVR to GIS. Well-designed reward engineering can therefore effectively shape the optimization direction of LLMs, particularly for GIS tasks with objective and verifiable feedback, such as geolocation and address rewriting. For more open-ended tasks, such as site selection and multi-objective planning, reward design remains challenging, as their outcomes are often evaluated against multiple competing criteria without a unique correct answer.

In scenarios involving tool use and interaction with dynamic environments, Agentic Reinforcement Learning (Agentic RL) [121] is receiving increasing attention as a means of enhancing task planning and execution in complex geospatial environments. Traffic-R1 [122], for example, interacts with the CityFlow traffic simulation environment and optimizes traffic-signal control policies using feedback derived from traffic-flow dynamics. DeepTravel [108] and STAgent [107] adopt an "SFT cold start + RL optimization" strategy, in which multidimensional rewards such as tool-use success rate and information fidelity guide extensive simulated trial-and-error. Through this process, the models learn effective strategies for solving travel-planning tasks under high uncertainty. By moving beyond static datasets, Agentic RL enables models to acquire autonomous problem-solving capabilities through repeated interaction with simulated environments.

Reinforcement Learning enables LLMs to improve through interaction and reward-guided optimization, particularly in planning and decision-making tasks with explicit objectives and constraints. Future advances are likely to depend on optimization algorithms and reward formulations specifically tailored to spatiotemporal reasoning and decision-making, together with geospatial simulation environments that more faithfully reflect real-world geographic settings. Such developments may enable increasingly autonomous spatial decision-making capabilities that approach or potentially exceed expert-level performance in well-defined tasks.

Overall, model development in LLM4GIS has progressed beyond the direct transfer of general-purpose training methods toward a more integrated framework encompassing architectural adaptation, multimodal tokenization, and end-to-end training optimization. Table 2 summarizes representative examples of LLM4GIS models across diverse tasks and compares their technical pathways in terms of backbone model selection, model architecture, tokenization strategy, and training stages, including PT, SFT, and RL.

Table 2. Representative LLM4GIS Models and Methods.

Model Name	Base Model	Arch	Token	PT	SFT	RL	Task
K2 [66]	Llama-7B		√	√	√		Geoscience QA
JiuZhou [64]	Mistral-7B		√	√	√		Geoscience QA
BB-GeoGPT [123]	Llama2-7B			√	√		GIS QA and reasoning
ClimateGPT-FSC [63]	Llama2-7B			√*	√		Climate QA
GeoCode-GPT [85]	Code-Llama-7B			√	√		Geospatial code generation
LandGPT [68]	InternVL2-8B		√	√	√		Land-use classification
UrbanLLaVA [124]	VILA1.5-8B				√		Urban understanding and reasoning
CityGPT [97]	Qwen2.5-7B				√		Urban understanding and reasoning
VMLocPredictor [117]	Qwen2-VL-2B			√	√	√	Trajectory prediction
GRE [94]	Qwen2.5-VL-7B				√	√	Image Geolocalization
SDPO-MLLM [125]	Qwen2-VL-7B				√	√	Flood depth Estimation
LHRS-Bot-Nova [91]	LLaMA3-8B	√		√	√		RS image understanding
SkyMoE [83]	Vicuna-V1.5	√			√		RS image understanding
EarthMind [76]	InternVL2		√	√	√		RS image understanding
RSThinker [104]	GLM-4.1V-9B-Base			√	√	√	RS image understanding
Geo-R1 [115]	Qwen2.5-VL-7B			√	√		Geospatial Reasoning
STAgent [107]	Qwen3-30B-A3B				√	√	Geospatial Agentic Reasoning
DeepTravel [108]	Qwen3-8B, 32B				√	√	Travel planning

Note: √ indicates that the corresponding technical component or training stage was used, whereas a blank cell indicates that it was not used. * indicates that the model was pretrained from scratch.

Although parameter-updating approaches can substantially improve the domain adaptability of LLMs, their high computational costs and demanding data requirements continue to hinder large-scale adoption. Consequently, another critical technical direction is to fully exploit the capabilities of foundation LLMs or domain-trained LLMs without modifying their parameters, thereby enabling more cost-effective adaptation and deployment for GIS applications.

3.3. Geospatial-Oriented Utilization Strategies

Beyond optimizing the underlying Model Architecture and model parameters, another major research direction focuses on aligning general-purpose LLMs with specialized GIS application requirements without modifying model weights. This is typically achieved by improving input interactions, incorporating external knowledge bases, and constructing autonomous agent systems. Representative approaches include Prompt Engineering [126],

RAG [127], and LLM-based Agents, as illustrated in Figure 9. Owing to their low computational cost and high deployment flexibility, these adaptation strategies have become the preferred choice for practical LLM4GIS applications and are often combined with architectural and training techniques to further optimize model performance.

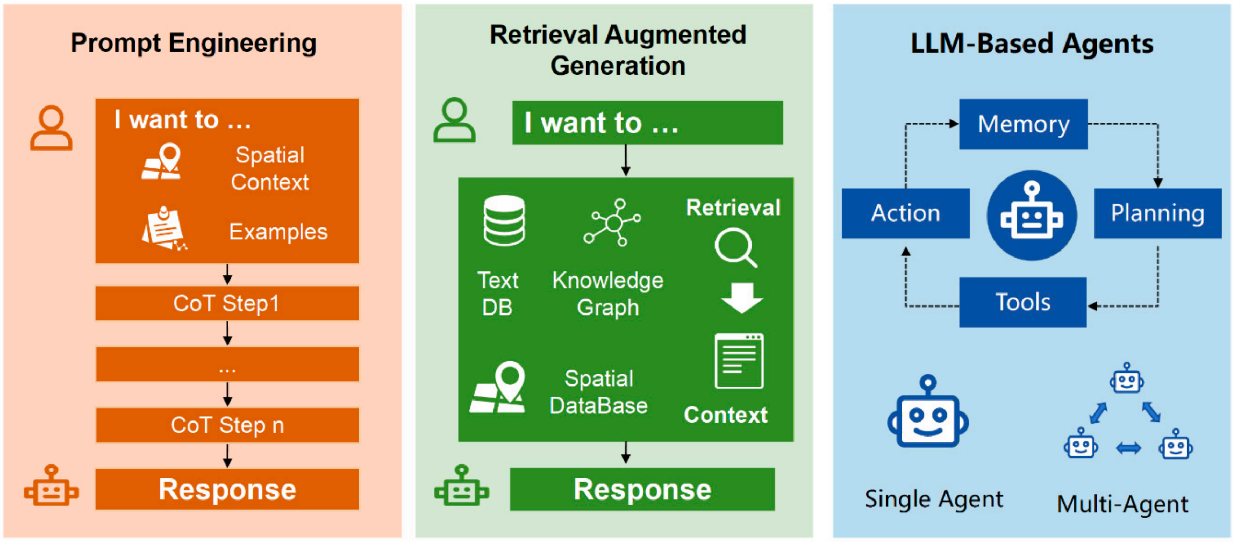


Figure 9. Overview of Prompt Engineering, RAG and LLM-Based Agents.

3.3.1. Prompt Engineering

Prompt Engineering improves task performance by constructing structured instructional contexts that elicit the latent reasoning and problem-solving capabilities of LLMs, guiding them to generate outputs that are more accurate and consistent with GIS-specific requirements. In current LLM4GIS research, general prompting strategies such as Zero-shot prompting, Few-shot prompting, CoT [128], and In-Context Learning (ICL) [129] have been widely adopted as standardized interaction mechanisms to enhance instruction following in geospatial question answering and analytical tasks.

To address the logical complexity and high degree of unstructuredness inherent in many GIS tasks, researchers have developed domain-specific prompting frameworks. For geospatial code generation, Chain-of-Programming (CoP) [130] employs a progressive prompting template comprising requirement interpretation, algorithm design, and code implementation, encouraging the model to perform algorithmic planning before generating code and thereby improving executability and accuracy. For non-standardized texts such as disaster reports, FE-HS [131] introduces an example- and constraint-based prompting framework that explicitly specifies how key elements, including geographic locations and disaster severity levels, should be extracted. This transforms an ambiguous natural-language processing task into a structured information extraction process and improves the accuracy of standardized GIS data extraction. For the highly subjective task of cartographic design, MapMate [132] derives actionable principles for color, symbolization, and layout from 136 cartographic textbooks and tutorials and injects them into the LLM through structured prompts, compensating for the lack of explicit objective criteria in such tasks.

Prompt Engineering enables task adaptation at minimal computational cost, but it fundamentally relies on eliciting knowledge already encoded in the model and therefore cannot fully resolve systematic gaps in GIS-specific knowledge. Consequently, it is often used as a foundational component within more complex LLM4GIS systems rather than as a standalone solution. Future advances may focus on automatically constructing and optimizing prompts through model self-reflection or Reinforcement Learning, as well

as developing prompt-compression strategies capable of accommodating extremely long geospatial contexts.

3.3.2. Retrieval-Augmented Generation

RAG enhances LLMs without modifying the underlying Model Architecture by retrieving and injecting up-to-date geographic data, dynamically evolving domain documents, and personalized user knowledge. This approach can effectively mitigate hallucinations and improve the relevance and accuracy of model outputs for time-sensitive and highly specialized GIS tasks. According to the type of retrieved information, existing RAG methods can be broadly categorized into three levels: text/document retrieval, graph-based retrieval, and spatial data retrieval.

At the text level, RAG primarily retrieves relevant information or documents to provide LLMs with domain knowledge that is absent from their internal parameters. For geospatial code generation, Geo-FuB and GEE-Ops [133,134], for example, construct vector knowledge bases containing GEE operator documentation and mappings of function semantics, thereby assisting models in generating more accurate geospatial code. In GeoSQL generation, studies [135,136] retrieve database metadata and schema information to ensure that generated SQL statements satisfy strict syntactic and structural constraints. To further improve retrieval accuracy, GeoFactory [137] combines the keyword-matching strengths of sparse retrieval with the semantic understanding capabilities of dense retrieval, enabling efficient responses to complex queries over domain knowledge documents. These text-oriented RAG methods can effectively compensate for the limited GIS-specific knowledge and proprietary API awareness of general-purpose LLMs, thereby reducing hallucinations caused by insufficient domain knowledge.

To introduce more structured reasoning capabilities, researchers have integrated graph-structured data, particularly knowledge graphs (KGs), into RAG pipelines, giving rise to Graph RAG. GeoGraphRAG [138], for example, uses semantic code networks as external knowledge sources to improve the logical consistency of code generation, while S-GMKG [139] further explores bidirectional interaction between LLMs and KGs. As a methodological reference, SpatialGPT [140] constructs a spatial knowledge graph as a structured memory module, enabling the model to revisit historical information or explore frontier nodes during route planning. In addition, study [141] proposes a KG-enhanced entity-linking mechanism that maps textual entities to KG nodes, effectively reducing referential ambiguity among geographic entities. The principal value of knowledge graphs lies in their ability to provide explicit relational constraints among entities. Compared with retrieving isolated text fragments, such structured knowledge is better suited to supporting multi-hop spatial reasoning.

Recent studies have begun extending retrieval beyond text to non-textual spatial data in order to provide richer geospatial context. Spatial-RAG [142] is among the earliest studies to propose a hybrid retrieval strategy for spatial objects, combining SQL-based spatial filtering as a sparse retrieval mechanism with text-based semantic matching as a dense retrieval mechanism. This design establishes a bridge between structured spatial databases and unstructured language models. For trajectory and grid data, PIGEON [99] and LLMTrajQuery [143] use Graph Neural Networks (GNNs) or LLM-generated textual descriptions to encode spatiotemporal characteristics into vector representations for retrieval. To improve indexing efficiency for large-scale spatial datasets, study [144] further introduces H3 grid cells as retrieval units. For result refinement and multimodal fusion, PlanGPT [145] integrates a tailored retrieval module with an urban planning agent framework, enabling efficient retrieval of domain-specific planning documents and spatial data to support context-aware urban analysis, whereas AskNearby [146] integrates text, graph,

and spatial RAG to support local-information retrieval under complex spatiotemporal constraints. In the Earth observation domain, the Land Surface Intelligence system [147] constructs a retrieval database using Facebook AI Similarity Search (FAISS) from 12.1 million environmental embedding vectors generated by the AlphaEarth foundation model. Each record is associated with coordinates, year, and 26 environmental variables, enabling LLMs to answer natural-language environmental queries using physically interpretable embedding features. This represents an emerging direction in which retrieval targets are extended beyond text and knowledge graphs to foundation-model-derived geospatial embeddings.

RAG has become a standard approach for addressing knowledge limitations in LLM4GIS, with current research already extending retrieval from text documents to structured knowledge graphs and spatial objects. Future work is likely to focus increasingly on multimodal spatiotemporal RAG, with the goal of efficiently retrieving and aligning heterogeneous geospatial data such as vectors, rasters, and point clouds and transforming them into contexts that can be interpreted by LLMs. At the same time, developing more efficient data-management and retrieval algorithms for large-scale geospatial repositories, while supporting personalized and context-aware information access, remains an important engineering challenge.

3.3.3. LLM-Based Agents

LLM-based Agents represent an important direction in the evolution of LLM4GIS from static knowledge question answering toward complex problem solving. As autonomous entities capable of environmental perception, task understanding, decision-making, and action execution, agents can extend the depth and scope of GIS task execution by integrating external toolchains, memory mechanisms, and closed-loop environmental feedback. Such agent orchestration does not itself require updating the parameters of the underlying LLM, although it can be further combined with SFT or RL.

As the agent ecosystem continues to expand, standardizing interactions between agents and external environments has become increasingly important. MCP, an emerging interoperability standard, is gradually serving as a standardized bridge between LLMs and external tools. Recent studies such as Earth-Agent [148] and STAgent [107] employ MCP to establish unified tool-invocation interfaces, reducing the technical overhead associated with adapting agents to heterogeneous data sources and analytical tools while improving system interoperability and extensibility.

Single-agent architectures focus on establishing an end-to-end closed loop within a single autonomous entity, covering intent interpretation, planning and reasoning, and tool execution. At the task-execution level, PEACE [149] and PlanGPT [145], targeting geological map understanding and urban planning respectively, adopt modular architectures that integrate information extraction, knowledge injection, and coordinated invocation of multiple models. MapGPT [150] and TrafficGPT [151] decompose complex user requirements into sequences of subtasks and complete specialized cartographic or traffic-analysis tasks through clarification-based interaction and iterative execution with domain-specific foundation models. For more challenging spatial reasoning tasks, GeoAgent [152] and IndustryScopeGPT [153] further introduce Monte Carlo Tree Search (MCTS) to dynamically simulate, evaluate, and refine candidate solutions.

Multi-Agent Systems (MAS) address complex tasks through explicit role specialization, inter-agent communication, and collective intelligence. Hierarchical collaboration is currently the dominant interaction pattern. GeoJSON Agents [154], for example, decomposes tasks into planner and executor roles: the planner formulates high-level strategies, while the executor focuses on implementation and debugging. GeoColab [155] constructs a

multi-role expert team for collaborative geospatial code generation, and similar role-based architectures are adopted in study [136] and CartoAgent [156] for map-style transformation and geospatial database querying. Some studies further incorporate adversarial or evolutionary mechanisms to improve output quality. smileGeo [157] reaches geolocation consensus through multi-round debate between answering and reviewing agents, whereas GeoEvolve [158] employs evolutionary search, iterative code refinement, and structured knowledge feedback to automate the improvement in GIS analysis models. Compared with single-agent architectures, MAS can improve both problem-solving capability and robustness through cross-agent verification, debate, and collaboration.

Large-scale agent-based social simulation and collective intelligence further demonstrate the distinctive value of MAS in urban science. CitySim [159] assigns each agent an independent personality, memory, and belief system, allowing urban dynamics to emerge from simulated large-scale social interactions among individuals. Studies [160,161] construct a triangular interaction framework involving planners, residents, and evaluators, in which resident agents simulate realistic public feedback to support the optimization of urban planning schemes. Such work extends LLM4GIS into the domain of computational social science and provides low-cost, controllable virtual laboratories for population-scale simulation and social governance research.

Agent technologies enable LLMs to interact with real GIS tools and dynamic spatiotemporal environments. Existing studies have already demonstrated effective GIS copilots capable of assisting with a wide range of analytical tasks. Future research is likely to focus on autonomous agents with stronger complex reasoning, self-reflection, and self-optimization capabilities, ultimately enabling end-to-end automation of spatiotemporal modeling and geospatial problem solving.

Prompt Engineering, RAG, and LLM-based Agents exhibit a clear functional progression. Prompt Engineering provides a lightweight interaction interface and serves as the foundation for task adaptation; RAG enables models to access external, updatable, and retrievable information beyond the knowledge encoded in the model; and agents focus on task planning, tool execution, and feedback-driven interaction with external environments. From a broader development perspective, isolated adaptation techniques are increasingly converging into integrated architectures. For next-generation LLM4GIS systems targeting more complex tasks, a likely standard configuration will center on agent-based frameworks that combine RAG, dynamic prompting strategies, and domain-specialized fine-tuned models.

4. Datasets

The geospatial knowledge underpinning LLM4GIS research is distributed across a wide range of sources, including academic publications, technical documentation, encyclopedic entries, remote sensing imagery, and street-view images. Compared with datasets in general-purpose domains, the construction of LLM4GIS datasets faces several challenges that are intrinsic to GIS. First, geospatial data are inherently multimodal and heterogeneous. GIS tasks involve diverse data forms, including textual descriptions, remote sensing imagery, street-view imagery, vector features, raster products, and spatiotemporal trajectories, which often require precise spatial and temporal alignment before they can be jointly utilized. Second, geospatial data exhibit pronounced scale effects. The same geographic entity may receive different labels across spatial scales, such as provincial, municipal, and neighborhood levels, or under different classification schemes, making annotation consistency difficult to maintain. Third, geospatial information is highly time-sensitive. Because geographic environments and their associated information are continuously evolving, datasets can rapidly lose their utility if they are not updated in a timely manner. Fourth, privacy constraints pose an additional challenge. Highly sensitive spatiotemporal data,

such as individual mobility trajectories and residential locations, are subject to stringent privacy regulations, which often prevent high-value real-world data from being directly collected, publicly released, or used for model training. Finally, pronounced spatial heterogeneity results in substantial geographic disparities in data availability and quality. In developed regions, data such as POIs, road networks, and remote sensing products are typically collected at high density, richly attributed, and frequently updated, whereas plateau, desert, and economically underdeveloped regions may remain sparsely covered or even persist as long-term data gaps. These challenges are closely intertwined, making it difficult to directly transfer mature dataset-construction paradigms from general-purpose domains to LLM4GIS. Instead, domain-specific strategies for data acquisition, organization, and curation are required. Accordingly, this section systematically reviews how existing studies have addressed these challenges from two complementary perspectives: dataset construction strategies and functional categories.

4.1. Dataset Construction

From the perspective of dataset construction, LLM4GIS datasets are primarily developed through five approaches: Open-Source Curation, Programmatic Generation, LLM Synthesis, Expert Annotation, and Crowdsourcing.

4.1.1. Open-Source Curation

Open-Source Curation refers to the acquisition and reuse of publicly accessible data resources from websites, online forums, data-service APIs, and other open channels. Existing studies have leveraged different types of open-source data to construct datasets for a wide range of geospatial tasks. One line of research draws on real-world geographic entities and Earth observation data from volunteered geographic information platforms such as OpenStreetMap, including vector features such as roads, hydrographic networks, and points of interest (POIs), as well as raster data such as remote sensing imagery and land-cover products. These resources are commonly used to build datasets for image understanding and spatial reasoning. For example, VL-Chat [162] directly collects OSM features, whereas GeoQAMap [163] further integrates three open knowledge sources—Wikidata, OpenStreetMap, and CityGML—to support more complex geospatial query tasks. Notably, such volunteer-contributed data are themselves a manifestation of spatial heterogeneity: data-rich regions tend to have dense and detailed coverage because of concentrated contributor activity, whereas underdeveloped and remote areas are often only sparsely mapped. This intrinsic geographic imbalance can therefore be directly propagated into downstream training datasets. Another line of research relies on dynamic spatiotemporal data, such as traffic conditions and road-network changes, to construct datasets for spatiotemporal reasoning and decision support. Representative examples include parking-demand simulation in the CityRiSE framework [116] and the multi-task spatiotemporal forecasting benchmark developed by UrbanGPT [164] using open urban data on New York City taxis, bike sharing, public safety, and related phenomena. Because such dynamic data are highly time-sensitive, these datasets often require periodic updates to remain valid and useful. In addition, some studies construct knowledge question-answering or knowledge extraction datasets by curating open-source textual resources, including GIS literature, encyclopedic materials, and statistical reports. Examples include the geoscience encyclopedia documents integrated by K2 [66] and the knowledge corpus developed by CLAImate [165] from publicly available climate data released by NASA, the IPCC, and the Met Office.

Open-Source Curation offers clear advantages in terms of low acquisition cost, large data volume, and broad coverage. However, as discussed above, both data quality and coverage density are spatially uneven, and open-source resources often lack the task-specific

annotations required for model training and evaluation. Consequently, they typically need to be complemented by subsequent filtering, cleaning, and annotation procedures, together with targeted data supplementation for sparsely covered regions.

4.1.2. Programmatic Generation

In this review, Programmatic Generation specifically refers to the automated construction of geospatial task datasets using structured rules, templates, or predefined workflows. It is therefore distinct from Geospatial Code Generation, which refers to the use of LLMs to generate code for geospatial analysis. One line of research relies on rule- and template-based procedures to extract facts and relations from existing knowledge sources or structured data and organize them into question-answering or spatial analysis tasks, with large numbers of task instances and reference procedures generated through systematic parameter enumeration. For example, [135] maps natural-language geospatial questions to executable spatial query templates and constructs datasets by programmatically varying query conditions and constraint combinations. PPNL [166] generates more than 160,000 path-planning samples through rule-based combinations of grid environments and obstacles, which are used to evaluate and fine-tune the spatiotemporal reasoning capabilities of models. For spatial relation reasoning, [167] systematically enumerates directional and topological relation tasks based on the transition tables of classical spatial calculi, including the Cardinal Direction Calculus (CDC) and the Region Connection Calculus with eight base relations (RCC-8), thereby ensuring comprehensive coverage of spatial logic. Another line of work constructs spatiotemporal task data through simulation environments. For instance, Parking Planning Agent [168] programmatically generates vehicle trajectories and parking-demand data using traffic simulation systems such as Simulation of Urban MObility (SUMO). In addition, K2 [66] reorganizes tasks such as geoscience question answering, geospatial entity recognition, and fact verification according to explicit cognitive and reasoning structures to assess model performance on spatiotemporal reasoning tasks.

Programmatic Generation can efficiently scale dataset size while maintaining well-defined task structures. Because its generation process is governed by predefined rules and templates, it can also reduce the scale- and taxonomy-related ambiguities that commonly arise in manual annotation, making it particularly suitable for datasets with explicit spatial logic and reasoning procedures. However, its strong dependence on predefined rules and templates limits its ability to capture real-world geographic heterogeneity and scenario complexity across regions. It therefore often needs to be combined with other dataset construction approaches.

4.1.3. LLM Synthesis

LLM Synthesis refers to the use of large language models to construct datasets from raw data or to expand existing datasets. One line of research rewrites and augments existing geospatial task descriptions to increase linguistic diversity while preserving the underlying task semantics. Other studies use LLMs to directly generate complete answers or solutions from task descriptions, or to supplement samples with multi-step reasoning processes and explanations. LLMs have also been employed to scale up tasks from limited seed data. For example, GeoLLM-Engine [169] begins with a small set of manually designed geospatial user intents and then uses an LLM to automatically generate large-scale, long-form geospatial queries. In addition, some studies use LLMs to generate textual descriptions aligned with geospatial imagery, including remote sensing and street-view images. ChatEarthNet [170], for instance, combines Sentinel-2 imagery with land-cover products and employs ChatGPT to generate more than 170,000 image–text pairs. LLM synthesis has further been explored in privacy-sensitive settings involving individual behavior,

where synthetic population profiles can serve as substitutes for real personal data. In [161], programmatic sampling is first used to match real-world demographic distributions, after which an LLM sequentially generates subjective resident attributes, such as preferences and habits, to construct a dataset of resident agents.

Overall, LLM Synthesis enables rapid dataset scaling at relatively low human cost while increasing the linguistic diversity and complexity of task instances. It also provides an alternative data construction pathway in scenarios where real-world data, such as individual trajectories, are difficult to obtain because of privacy and regulatory constraints. However, LLM-generated data remain subject to uncertainty in spatial constraints, coordinate consistency, and semantic accuracy. Rigorous quality control, typically through expert review or complementary validation procedures, is therefore required to ensure dataset reliability.

4.1.4. Expert Annotation

Expert Annotation refers to the structured organization and labeling of existing data by domain experts using their specialized knowledge. In current LLM4GIS research, Expert Annotation primarily serves three functions. First, it provides reference answers or evaluation criteria for benchmark datasets. For example, the GeoCodes evaluation set in GeoCogent [171] was annotated by experts with correct solutions and failure modes for geospatial coding tasks derived from three types of real-world requirements: explicit, incomplete, and open-ended requests. Reference [172] directly draws on the instructor resources accompanying the GIS textbook GIS Fundamentals, reorganizing assessment questions into a reusable evaluation benchmark. Second, experts participate in the design of geospatial questions and task descriptions to ensure that task settings are well defined and semantically unambiguous. Third, experts review and filter data produced through Programmatic Generation or LLM Synthesis. This role is particularly important in geospatial agent and workflow construction tasks, where professional judgment is required to determine appropriate tool selection and invocation sequences.

Expert Annotation plays an indispensable role in ensuring the domain validity and accuracy of geospatial datasets and is particularly important for resolving annotation ambiguity. By applying consistent classification schemes and scale criteria, experts can assign a unique label, or establish an explicit priority among alternative labels, for the same geographic entity. However, this approach depends heavily on highly qualified domain experts and entails substantial annotation costs. It is therefore better suited to the construction of small-scale, high-quality datasets or to serving as a critical validation mechanism for datasets produced through other construction approaches.

4.1.5. Crowdsourcing

Crowdsourcing primarily relies on non-expert participants for data collection and offers a cost-effective means of scaling dataset size while increasing linguistic diversity. It is often combined with Expert Annotation or Programmatic Generation. For example, GeoLLM-Engine [169] draws on crowdsourced data from Japan's national time-use surveys, travel questionnaires, and well-being surveys to capture public mobility behaviors and subjective perceptions. The CitySim framework [159] collects public safety-perception ratings of street-view images in Chengdu through online crowdsourcing. For natural-language geolocation tasks, HeGeL [173] constructed the first Hebrew geolocation dataset through a two-stage crowdsourcing process: one group of participants wrote location descriptions based on first-hand knowledge without mentioning place names or addresses, while another group verified the descriptions and annotated the corresponding locations on a map. The RVS dataset [174] was collected through Amazon Mechanical Turk, where

one group of annotators wrote non-sequential navigation instructions and another group verified the destination locations, covering Manhattan, Pittsburgh, and Philadelphia.

Crowdsourcing is advantageous in reducing data acquisition costs and introducing greater linguistic and cognitive diversity. Compared with the direct collection of privacy-sensitive data such as individual trajectories, anonymized crowdsourced questionnaires also generally entail lower privacy and compliance risks. However, crowdsourced data tend to be less reliable in terms of accuracy and domain-specific consistency, making them insufficient as a standalone source for tasks that require high reliability or specialized expertise. Crowdsourcing is therefore better suited as a complementary source of information within the broader dataset construction process.

4.2. Dataset Types

According to their functional roles, datasets used in LLM4GIS research have gradually evolved into a differentiated organizational framework and can be broadly classified into six categories: pretraining datasets, fine-tuning datasets, reinforcement learning datasets, evaluation datasets, knowledge corpora, and tool datasets.

Pretraining datasets provide models with foundational geospatial semantic knowledge and typically comprise large-scale textual or multimodal resources, including research articles, technical documentation, encyclopedic entries, and other web-based data. Representative examples include LandGPT-PT [68], which integrates 15 publicly available data sources for remote sensing multimodal tasks; K2 [66], whose geoscience corpus contains 5.5B tokens from publications, metadata, and encyclopedia pages; and ClimateGPT [63], which combines 300B general-domain tokens with 4.2B climate-domain tokens as a mixed-corpus strategy for mitigating catastrophic forgetting. Because pretraining datasets establish broad domain knowledge rather than task-specific capabilities, they generally need to be complemented by fine-tuning datasets.

Unlike pretraining datasets, fine-tuning datasets target specific geospatial tasks and are commonly organized as question–answer pairs constructed through expert annotation or semi-automated procedures. Representative examples include RSICap [175], which contains 2585 high-quality remote sensing image–text pairs for visual question answering; UrbanLLM [176], which uses a self-instruct strategy to generate more than 15,000 urban activity planning samples from 170 seed prompts; and Chatmap [177], which constructs 4111 samples from densely annotated OpenStreetMap regions in Istanbul to transform structured map features into natural-language representations for conversational interaction.

For complex decision-making tasks, reinforcement learning datasets complement SFT by emphasizing environmental interaction and verifiable feedback, particularly in spatiotemporal reasoning, path planning, and urban system optimization. Representative examples include RSThinker-RL [104], which combines RSVA-HR-train, NWPU-Captions-train, and RSTMD-train into 100,049 samples linking remote sensing visual outputs with task-level feedback; the Expert-Collaborative TSC dataset of Traffic-R1 [122], which uses expert-recommended actions validated through SUMO simulation to construct 3000 samples for reinforcement fine-tuning of traffic signal control; and GRE30K [94], which provides complementary chain-of-thought and judgment subsets for cold-start SFT and reward modeling in image geolocation.

Unlike datasets used for model training, knowledge corpora serve as external sources for RAG, providing retrievable and traceable information during inference. They typically comprise structured or unstructured resources such as encyclopedic entries, technical documentation, and domain-specific knowledge bases. Representative examples include ClimateChat-KC [178], which draws on IPCC reports and Wikipedia articles; GeoColab's GeoCode-Base [155], which supports geospatial code generation; and the FEMA-guideline

corpus used by DisasterResponseGPT [179] for emergency-response scenarios. By externalizing specialized knowledge, these corpora can also help mitigate hallucinations without encoding all domain information directly into model parameters.

Tool datasets support geospatial agents in selecting, sequencing, and combining external tools for complex tasks. They typically consist of task descriptions, tool-calling instructions, and execution workflows constructed through expert design and domain documentation. Representative examples include AutoUrbanCI-PT [180], which combines user-uploaded, Web-retrieved, and simulation-generated data for learning tool-selection strategies; Geode [181], which provides agents with access to real-time geospatial and environmental data APIs; and the Tool Space of GeoLLM-Engine [169], which supports Python-based data parsing and geospatial operations. Rather than directly enhancing intrinsic model capabilities, these datasets teach agents how to use external resources and toolchains to solve geospatial problems.

Evaluation datasets assess model performance on defined geospatial tasks using questions paired with reference answers, thereby supporting comparisons across models and methods. Because their reliability depends on the quality of the reference answers, they commonly require substantial expert involvement. Representative examples include LHRS-Bench [81] for remote sensing evaluation, GeoCode-Bench [182] for multi-format geospatial code tasks, and GeoMap-Bench [149] for geological map understanding; their detailed scales and compositions are summarized in Table 3. Although generally smaller than training datasets, evaluation datasets remain essential for model validation, with the corresponding evaluation methodologies discussed further in Section 6.

Datasets in the LLM4GIS domain are designed to enhance the GIS capabilities of large language models and, from a functional perspective, span key stages including foundational geospatial semantic understanding, task-specific capability enhancement, multi-step task execution, and model performance evaluation. Different dataset types serve distinct functional roles and are employed at different stages of model development, making their coordinated use essential for comprehensively supporting capability development and performance evaluation of LLMs on complex GIS tasks. Table 3 summarizes representative datasets reported in LLM4GIS research.

Table 3. Typical Datasets Used in LLM4GIS Research.

Dataset Type	Name	Year	Construction Method	Description
Pretraining Dataset	**K2-PT** [66]	2024	Programmatic Generation	Built from a 5.5B-token geoscience text corpus comprising multiple sources, including geoscience publications, publication metadata, and Wikipedia, and covering domain-specific knowledge in the geosciences.
	EarthGPT-PT [80]	2024	Open-Source Curation	Integrates LAION-400M image–text pairs and COCO Caption data, providing knowledge for cross-modal representation learning and image–text alignment.
	WorldST [74]	2026	Open-Source Curation	Aggregates data from multiple sources, including government open data, academic repositories, and domain-specific APIs, followed by unified standardization.
	PostGIS-PT [92]	2026	Open-Source Curation	Constructed from textual content in the official PostGIS documentation.
	EarthMind-PT [76]	2025	Open-Source Curation	Integrates general vision–language pretraining data with remote-sensing-specific vision–language data, covering cross-modal alignment, remote sensing scene understanding, and cross-sensor representation learning.
	GeoPLACE-PT [183]	2024	LLM Synthesis	Contains curated place-name entries and their canonical names, covering geographic knowledge related to toponym recognition and geographic entity disambiguation, with the aim of improving model understanding and reasoning over geographic entity names.

Table 3. *Cont.*

Dataset Type	Name	Year	Construction Method	Description
Fine-tuning Dataset	**ChatGeoAI-FT** [184]	2024	LLM Synthesis	Built from 1000 prompt–code pairs for geospatial analysis, covering geospatial analysis workflows and GIS programming knowledge, with the aim of improving geospatial code generation capabilities.
	PostGIS-SSFT/TSFT [92]	2026	Programmatic Generation	Constructed from official examples through description generation and data-replacement augmentation; TSFT samples are generated according to three predefined difficulty levels: simple, medium, and complex.
	UrbanGPT-FT [164]	2024	Programmatic Generation	Built from multi-view urban spatiotemporal data, including bicycle and taxi data from New York City and Chicago, and incorporating urban dynamics such as traffic flows and crime events, with the aim of improving model performance in urban computing scenarios.
	Delta-QA [185]	2026	Open-Source Curation, Programmatic Generation	Built from open-source remote sensing datasets such as SECOND, Landsat-SCD, and WUSU, with question–answer pairs generated through an automated pipeline.
	UrbanLLM-FT [176]	2024	LLM Synthesis	Comprises 15,249 training samples covering urban activity planning and management query tasks, with the aim of improving spatiotemporal reasoning capabilities for urban activity planning.
	ChangeChat-105k [186]	2026	Open-Source Curation, LLM Synthesis	Uses image pairs from the open-source LEVIR-CC and LEVIR-MCI datasets as the foundation, while open-ended question answering and multi-turn dialogue samples are generated with ChatGPT.
	GeoChat-FT [187]	2024	LLM Synthesis	Contains 308,000 multimodal instruction-following samples constructed from multiple open-source remote sensing datasets, covering remote sensing scene understanding, object attribute representation, and related multimodal knowledge, with the aim of enhancing instruction-following capabilities in remote sensing large models.
Reinforcement Learning Dataset	**ST-RL** [188]	2026	LLM Synthesis, Programmatic Generation	Built on 1064 spatiotemporal–text pairs synthesized through a multi-agent pipeline, with 6593 reinforcement learning samples retained through rejection sampling.
	MP16-Reason-RL [189]	2025	Programmatic Generation	Comprises 33,000 multi-view image samples extended from MP16, covering visual-cue reasoning and geolocation discrimination, with the aim of improving visual geographic reasoning and localization in complex real-world scenarios.
	RQA-70K-RL [95]	2025	Expert Annotation	Built from 70,000 multimodal samples derived from SEVIR radar storm events, covering storm-evolution perception, numerical feature interpretation, and radar nowcasting reasoning, with the aim of enabling the model to achieve near-expert judgment in weather-radar forecast quality assessment.
	SpotRL [190]	2026	Programmatic Generation	Applies a dynamic filtering strategy to SFT data, progressively selecting samples according to spatial difficulty in a curriculum-learning manner.
	MP16-Hard-200K [118]	2026	Open-Source Curation, Programmatic Generation	Programmatically selects remote and visually ambiguous samples located more than 200 km from popular regions from the open-source MP16-Pro dataset.
	Geo-R1-RL [191]	2025	Open-Source Curation	Built from few-shot remote sensing reinforcement learning datasets, including VRSBench-FS, NWPU-FS, and EarthReason-FS, covering decision-feedback knowledge for few-shot object recognition, open-set detection, and referring-expression segmentation, with the aim of improving spatiotemporal reasoning over combined textual and remote sensing inputs in few-shot scenarios.
Evaluation Dataset	**MapGPT-Bench** [150]	2024	Expert Annotation	Built from multiple map-generation cases produced by MapGPT for regions including Guangdong, Wuhan, and Henan, covering GIS-specific criteria such as map generation quality, spatial representation accuracy, and cartographic aesthetics, with the aim of evaluating the overall cartographic capability of large language models in map generation tasks.
	GPSBench [192]	2026	Open-Source Curation, Programmatic Generation	Built on place data from the GeoNames database, with 17 evaluation tasks generated through predefined rules.
	LLMmove-Bench [193]	2024	Open-Source Curation	Built from 103,941 chronologically ordered user check-in records in New York City and 405,300 in Tokyo, together with their corresponding trajectory information, with the aim of evaluating model performance in predicting human mobility behavior.
	DORA [194]	2026	Expert Annotation	Contains 515 expert-authored tasks with gold-standard trajectories validated by domain experts; the underlying imagery and vector data are drawn from public datasets including xBD, DisasterM3, and OpenStreetMap.
	ShapefileGPT-Bench [195]	2025	Expert Annotation	Built from 42 standardized Shapefile-based spatial analysis tasks, covering GIS spatial analysis workflows and analytical functions, with the aim of evaluating model capabilities in Shapefile-driven geospatial analysis and task execution.
	UrbanCLIP-Bench [196]	2024	Expert Annotation	Built from street-view evaluation data and urban classification schemes for Shenzhen, Singapore, and London, covering urban ground-level functional recognition tasks, with the aim of evaluating cross-city street-view understanding and urban function classification capabilities.

Table 3. *Cont.*

Dataset Type	Name	Year	Construction Method	Description
Knowledge Corpus	**GeoKnowRAG** [158]	2025	Open-Source Curation	Constructed from 141 geospatial knowledge documents, covering systematic knowledge of geospatial theories and methodologies.
	[147]	2026	Programmatic Generation	Built through a custom pipeline that extracts embeddings from the AlphaEarth foundation model and indexes them using FAISS.
	QIV-CSSTD [197]	2024	Expert Annotation	Defines a hierarchical quality assessment framework for six types of crowdsourced spatiotemporal data, comprising primary quality dimensions and secondary quantifiable indicators, and covering knowledge related to spatiotemporal data quality assessment and standardized representation.
	MapMate [132]	2026	Open-Source Curation, LLM Synthesis	Draws on open-source cartographic textbooks, tutorials, and blogs, from which actionable principles are extracted and synthesized by an LLM using structured prompts.
	SpatialKG [198]	2025	Programmatic Generation	Constructed from OpenStreetMap and Foursquare data, representing geospatial knowledge of urban spatial entities and their relationships.
	AskNearby-KC [146]	2025	Open-Source Curation	Built from approximately 20,000 localized user-generated content records from RedNote, including text, timestamps, and geographic coordinates, and covering neighborhood-scale spatiotemporal semantics and local knowledge.
Tool Dataset	**GeoGPT GIS Tool Pool** [8]	2024	Expert Annotation	Built around 50 geospatial tasks and 22 callable GIS tools, covering task-level geospatial reasoning and coordinated tool use.
	Remote Sensing ChatGPT [199]	2024	Expert Annotation	Constructed as a visual-model toolset based on six representative remote sensing image task datasets, covering remote sensing task understanding, tool selection, and workflow planning.
	Climate Environment [200]	2026	Open-Source Curation	Integrates 150 specialized tools and 30 databases, covering operations such as data preprocessing and simulation algorithms, and providing climate agents with an executable environment and supporting physical knowledge.
	MapCopilot workflow/tool environment [201]	2025	Expert Annotation	Constructed from historical topographic maps of Taiwan spanning 1904–1989 and equipped with callable spatiotemporal place-name indexing tools.
	iwmi-api-plugin [202]	2026	Open-Source Curation	Encapsulates a real-time hydrological database for the Limpopo River Basin, providing dynamic query interfaces for environmental flows, rainfall, reservoir water levels, water accounting, irrigation, and related variables.
	CrimeMind-TP [203]	2025	Expert Annotation	Built from multiple tool-oriented data sources, including SafeGraph demographic data, Google Street View imagery, human-annotated safety perception data, and official crime records, covering urban environment modeling, perception alignment, and crime risk assessment.

5. Applications

LLM4GIS has been widely applied across diverse GIS tasks, which can be broadly categorized into five major types according to their primary objectives: Geospatial Knowledge Q&A, Geospatial Knowledge Extraction, Spatiotemporal Reasoning, Geospatial Code Generation, and Geospatial Analysis Modeling, as illustrated in Figure 10. Detailed bibliographic information for the studies shown in Figure 10 is provided in Appendix B, Table A3.

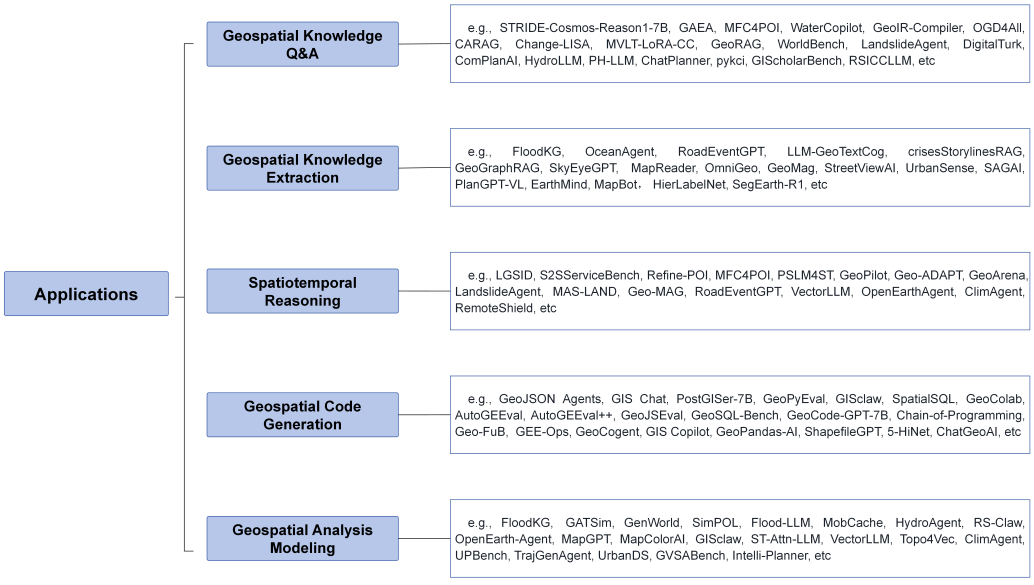


Figure 10. Overview of LLM4GIS Applications.

5.1. Geospatial Knowledge Q&A

Traditional GIS knowledge question answering primarily relies on rule-based methods that generate answers by matching and retrieving information from predefined knowledge bases, making it difficult to address complex questions when knowledge coverage is incomplete or when integrated reasoning is required. In contrast, large language models can perform semantic understanding, inference, and generation even when relevant knowledge is not explicitly encoded, thereby enabling multi-turn, context-consistent Geospatial Knowledge Q&A. Against this background, a variety of domain-specific question-answering models have been developed, including ClimateGPT [63] for climate science, OceanGPT [88] for ocean science, GeoProspect [84] for geology, and TrafficGPT [151] for transportation, as well as models that incorporate multimodal information, such as CityGPT [97] for urban spatial question answering, TraveLLaMA [204] for travel scenarios, RSThinker [104] for remote sensing image understanding, and EarthMind [76] for Earth observation. These models typically employ continued pretraining and instruction tuning to equip LLMs with systematic domain-specific GIS knowledge, enabling them to provide high-quality natural-language responses for analysis and decision support. In parallel, external knowledge bases and retrieval mechanisms have been introduced to improve question-answering accuracy. Representative examples include GeoQA2 [205], which is built on the YAGO2 and YAGO2geo knowledge graphs; HybridRAG [206], which incorporates an earthquake emergency knowledge graph; a constrained question-answering framework augmented with a flood-risk knowledge graph [207]; and Spatial-RAG [142], which integrates spatial database queries to effectively answer complex geographic questions involving locations, routes, and spatial constraints.

The application of LLMs to Geospatial Knowledge Q&A not only improves the efficiency and accuracy of information acquisition, but also provides an effective knowledge-access interface for downstream geospatial applications.

5.2. Geospatial Knowledge Extraction

Traditional GIS knowledge extraction typically relies on task-specific datasets and dedicated extraction models, limiting its flexibility across heterogeneous applications. In contrast, LLM-based Geospatial Knowledge Extraction can generalize under zero-shot or few-shot settings and adapt to multimodal inputs and diverse task contexts. Existing studies can be broadly grouped into three categories.

For code and workflow knowledge extraction, existing studies mine large-scale GEE scripts to extract function semantics, composition patterns, and relational chains, thereby constructing function-oriented knowledge bases such as Geo-FuB and GEE-Ops [133,134]. Another line of work, represented by S-GMKG and Geo-CLASS, employs LLMs to identify modeling tasks, processing workflows, and data entities, and subsequently generates standardized geographic modeling knowledge graph templates to support cross-platform reuse and automated code generation [139,208].

For textual knowledge extraction, LLMs are used to identify geographic events and associated attributes from unstructured text. For example, FE-HS combines toponym-tree search with prompt engineering to extract fine-grained information such as flood depth, the number of stranded people, and addresses from social media posts [131]. LLM-MPE transforms descriptions in online event reports into structured event features to support subsequent spatiotemporal modeling [209]. UrbanKGent extracts entity–relation triples from urban planning documents and constructs and completes urban knowledge graphs to enable integrated querying of urban knowledge [210].

For multimodal knowledge extraction, the integration of vision models and LLMs has been increasingly applied to remote sensing imagery and street-view image analysis, with a

particular focus on automatically extracting fundamental geographic knowledge elements, including geographic entities and their attributes, spatial locations and relations, as well as scene and region types. Models such as CityVLM and SightSense transform visual scenes, objects, and urban elements into structured or interpretable geospatial semantic knowledge, thereby supporting urban environmental understanding and sustainability analysis [91,211,212].

Collectively, these studies demonstrate the capability of LLMs to extract knowledge from heterogeneous GIS data sources, including code, text, and imagery. This capability provides a new pathway toward constructing reusable and reasoning-oriented structured geospatial knowledge bases, while also supporting complex GIS analysis and intelligent decision-making.

5.3. Spatiotemporal Reasoning

Spatiotemporal Reasoning aims to identify patterns and understand relationships within complex spatiotemporal data, thereby supporting prediction and decision-making. Compared with conventional deep learning approaches, LLM-based Spatiotemporal Reasoning is better positioned to integrate homogeneous and heterogeneous data, incorporate commonsense and external knowledge, and improve generalizability and interpretability through chain-of-thought reasoning and tool use. Existing studies can be broadly grouped into four categories.

Multimodal geographic reasoning integrates visual and linguistic information across tasks such as geolocation, cross-view understanding, and spatial-relation inference. Representative studies include visual-cue-based geolocation with GeoGuess and GEO-Detective [211,213], socioeconomic and environmental inference from satellite and street-view imagery with UrbanMLLM [214] and CityVLM [212], and evaluation of directional and distance relations with SpatialRGPT [215].

To address the limited availability of spatiotemporal data caused by privacy constraints and high acquisition costs, some studies use LLMs to generate realistic synthetic data for planning and simulation. LLMob [216] and MobGLM [217] learn mobility patterns from historical trajectories and generate synthetic movement trajectories with statistical properties close to those of real-world data. The Urban Generative Intelligence framework employs multiple agents representing mobility, economic, and social behaviors to simulate individual actions and interactions for evaluating urban planning scenarios [90]. Other studies use multi-agent negotiation to simulate interactions between planners and residents, automatically generating optimized land-use plans and reproducing the behavioral dynamics of complex urban systems [160].

In spatiotemporal prediction and reasoning, applications include traffic-state forecasting [218], trajectory behavior modeling [219], urban attribute estimation [220], daily demand generation [108], chain-of-thought-based traffic-signal control with LLMLight [100], reinforcement-learning-based next-location prediction with VMLocPredictor [117], and socioeconomic and health indicator inference with GeoLLM [221].

For geographic decision support, LLMs are increasingly used to generate and optimize decision plans subject to geographic constraints, including travel planning, site selection, and route navigation. Multimodal models such as TraveLLaMA automatically generate travel itineraries based on user preferences and map information [204], while AgentTravel [222] and DeepTravel [108] further assess route accessibility and temporal feasibility. For urban facility siting and layout optimization, agents such as GeoGPT [8] and IndustryScopeGPT [153] can invoke GIS spatial analysis tools to identify locations that satisfy specified constraints. AgentSense [223] iteratively adjusts sensor inspection routes during emergency events, supporting more efficient and accurate disaster response.

Collectively, these four lines of research demonstrate the broad applicability of LLMs to understanding spatiotemporal patterns, modeling complex behaviors, generating future states, and supporting spatial decision-making. However, as reasoning chains become longer, errors may accumulate across multiple reasoning steps, while the generalization capability of these models across regions and task domains remains insufficiently validated. These limitations constitute important challenges for future research in Spatiotemporal Reasoning.

5.4. Geospatial Code Generation

GIS spatial analysis programming relies heavily on platform- and language-specific expertise. Whether using the JavaScript/Python interfaces of GEE, Python-based extensions for desktop GIS software such as ArcPy and PyQGIS, or spatial database query languages such as PostGIS SQL and SPARQL, users are required to master both geoscientific analytical logic and implementation-level programming details. This dual requirement constitutes a major barrier to the interdisciplinary adoption of GIS tools. With their capabilities in code generation and semantic understanding, LLMs can directly translate analytical intents expressed in natural language into executable code, substantially lowering this barrier. Existing studies on Geospatial Code Generation can be broadly grouped into three categories.

For platform- and language-specific code generation, GEE has become one of the most intensively studied environments because of its widespread use in Earth observation data processing. GeoCode-GPT-7B [85], the first code generation model specifically designed for geospatial tasks, was built upon Code Llama and further trained on a GEE code corpus. Geo-FuB [133] and GEE-Ops [134] construct operator–function knowledge bases that provide retrieval support for generating syntactically correct GEE scripts, whereas AutoGEEval [224] and AutoGEEval++ [225] establish multilevel and multimodal automated evaluation frameworks for systematically assessing the executability and accuracy of GEE code generated by LLMs. Within the Python ecosystem, ChatGeoAI [184] translates natural-language queries into PyQGIS code snippets, enabling non-expert users to perform spatial analysis without mastering GIS software operations. GeoPandas-AI [226] embeds LLMs into GeoPandas workflows as state-aware intelligent coding assistants that support common geospatial programming tasks such as coordinate reference system transformation and topology processing. For JavaScript code generation, GeoJSEval [227] develops a unit-level automated evaluation framework for widely used mapping and spatial computing libraries, including Turf.js, Leaflet, and OpenLayers, to assess the ability of LLMs to generate front-end map interaction code. In addition, Chain-of-Programming [130] introduces a chained prompting framework that progresses from requirement interpretation to algorithm design and code implementation, effectively mitigating code hallucinations and providing a methodology applicable across platforms and programming languages.

For natural-language querying and database retrieval, LLMs have been widely used to automatically translate user queries into structured query statements, thereby lowering the barrier to accessing spatial databases. Existing studies cover the generation of multiple query languages, including NL2GeoSQL, OverpassQL, and SPARQL. Representative tasks include geographic entity and relation extraction, execution of spatial analysis operations, PostGIS query generation, and cross-database knowledge graph construction, collectively improving the practical utility and generalizability of LLMs for spatial data interaction and automated geospatial modeling [228–231]. PostGISer-7B [92] extends this paradigm to spatial databases through PostGIS-PT pretraining followed by a two-stage chained fine-tuning strategy comprising PostGIS-SSFT and PostGIS-TSFT, making it the first specialized end-to-end model for generating GeoSQL queries.

In agent-based and autonomous programming, researchers have further encapsulated code generation capabilities into agents equipped with task planning and environment-feedback mechanisms. GeoCogent [171] integrates task planning, tool invocation, environmental feedback, and memory to improve execution stability in complex geospatial code generation tasks. GeoColab [155] adopts a multi-agent collaboration architecture consisting of "product manager–algorithm engineer–programmer" roles, emulating real-world software development workflows to progressively complete task interpretation and code implementation. GeoAgent [152] incorporates search engines, code interpreters, and Monte Carlo tree search to dynamically optimize and correct code execution processes. ShapefileGPT [195], targeting the Shapefile format that is widely used in GIS vector data management, constructs a multi-agent collaborative framework to automate specialized operations such as format parsing and attribute-field processing. In terms of tool integration, GIS Copilot embeds LLMs directly into QGIS to support conversational spatial analysis [60], while GeoJSON Agents [154] combines function calling with code generation to automate modeling operations such as format conversion, buffer analysis, spatial joins, and aggregation, facilitating the transition of GIS programming toward lower technical barriers and greater autonomy.

Collectively, these studies indicate that LLMs are evolving from auxiliary generators of isolated code snippets into intelligent agents capable of understanding geoscientific analytical intents and autonomously completing complex spatial programming tasks across platforms. This evolution substantially lowers the technical barriers to GIS spatial analysis.

5.5. Geospatial Analysis Modeling

Geospatial Analysis Modeling refers to the design and construction of GIS-based spatial analysis models tailored to specific analytical requirements. Compared with traditional approaches that rely heavily on expert knowledge for requirement interpretation and model design, LLM-based analysis modeling can automate multiple stages of the modeling process, including requirement analysis, model design, and model construction. This automation has the potential to lower the technical barriers to interdisciplinary GIS applications and improve modeling efficiency. Existing studies mainly focus on the following aspects.

For geoprocessing workflow generation, LLMs are used to automatically orchestrate sequences of analytical tools. GeoTool-GPT [93] is fine-tuned on spatial analysis tasks to enable LLMs to automatically select and order appropriate tools. GeoGPT [8] further integrates LLMs with a GIS execution environment and, after interpreting user intent, directly generates end-to-end workflows covering data acquisition, preprocessing, spatial analysis, and cartographic visualization, thereby improving both the degree of automation and the execution efficiency of geospatial analysis.

For thematic analysis automation, LLMs are applied to specific GIS research questions and management needs, enabling customized modeling from data processing and analytical output to decision recommendation. At the scientific analysis level, GeoLLM-Squad [232] employs a multi-agent system to manage remote sensing workflows and supports modeling tasks in domains such as agricultural monitoring, climate analysis, urban studies, and forest conservation. In hydrology, HydroAgent [233] formalizes forecasting expertise into composable skill modules and constructs an agent-based system covering the full hydrological modeling workflow. For Earth observation tasks in open environments, OpenEarth-Agent [234] goes beyond merely invoking predefined tools by enabling coding agents to dynamically create new tools; its adaptive workflow planning capability was validated on 596 real-world cases spanning seven domains, including urban areas, agriculture, and water bodies. RegionDefiner [235] focuses on regionalization, a classic spatial analysis

task, and develops an end-to-end agent workflow from task-intent recognition to clustering with the PySAL algorithm library. GeoEvolve [158] combines a multi-agent framework with evolutionary search to automatically discover new variogram families that outperform conventional kriging approaches in fundamental geospatial modeling tasks such as spatial interpolation, highlighting the potential of LLM-assisted model discovery. At the decision-output level, Remote Sensing ChatGPT [199] integrates ChatGPT with specialized remote sensing models through a workflow comprising prompt-template generation, task planning, task execution, and response generation, thereby supporting multiple interpretation tasks such as scene classification and object detection. TrafficGPT [151] performs traffic data extraction, visualization, and simulation-system operations through natural-language interaction, providing decision support for traffic-signal optimization and road-network performance analysis. The Instructor-Worker system [236] uses a code-execution module to access air-quality data through Google Maps APIs and then delegates tasks such as anomaly detection to a Worker LLM, supporting environmental decision-making during emergencies such as the 2025 Los Angeles wildfires.

For automated cartography, LLMs are used to transform natural-language descriptions into maps. Frameworks such as MapGPT [150] interpret user intent, automatically invoke mapping tools to generate maps, and support conversational style refinement. Multimodal models such as LandGPT [68] integrate imagery and population data to perform land-use classification and produce corresponding map outputs, demonstrating capabilities in thematic mapping. Other studies identify dashboard wireframes in presentation slides and automatically generate front-end code for web applications containing map components, further improving the intelligence and efficiency of map design and development workflows [237].

LLMs are progressively being integrated into all stages of Geospatial Analysis Modeling, from query understanding and workflow execution to code generation and map production. This progression is shifting geospatial modeling from basic functional automation toward more accessible, efficient, and intelligent workflows, laying the foundation for increasingly autonomous GISs.

Overall, existing studies provide task-specific evidence for the potential advantages of LLM4GIS, but the strength of this evidence varies across applications. Most reported improvements are established through within-study comparisons under particular datasets, baselines, and testing conditions. Differences in evaluation settings limit direct comparison across studies, while computational costs, cross-region generalization, and failure cases are not consistently reported. Therefore, the reported gains in accuracy, generalizability, and accessibility should be interpreted as context-dependent findings rather than uniformly demonstrated advantages across all LLM4GIS applications.

6. Evaluation

Existing evaluation methods for LLM4GIS can be broadly classified into objective and subjective evaluation. According to the strictness of answer matching, objective evaluation can be further divided into three categories: exact matching, similarity-based metrics, and execution-based evaluation. Subjective evaluation encompasses both human evaluation, including scoring, ranking, and compliance assessment, and LLM-based evaluation (LLM-as-Judge), which can be further categorized into reference-guided and criterion-guided approaches. These methods are not mutually exclusive; rather, their combinations and relative emphasis vary across the five applications depending on the form of the expected output, including whether a unique ground-truth answer exists, whether execution-based validation is required, and whether subjective judgments are involved. Compared with general-purpose LLM evaluation, the evaluation of GIS tasks must additionally assess

models' ability to understand diverse spatial data types, such as vector data, raster data, and cartographic symbols, while accounting for the substantial variations that the same problem may exhibit across different spatial scales. Consequently, general-purpose evaluation frameworks cannot be directly transferred to GIS, necessitating specialized evaluation systems that reflect the distinctive characteristics of the discipline. Figure 11 provides an overview of the evaluation methods and representative benchmarks, while Table 4 summarizes representative examples of benchmarks across the five applications.

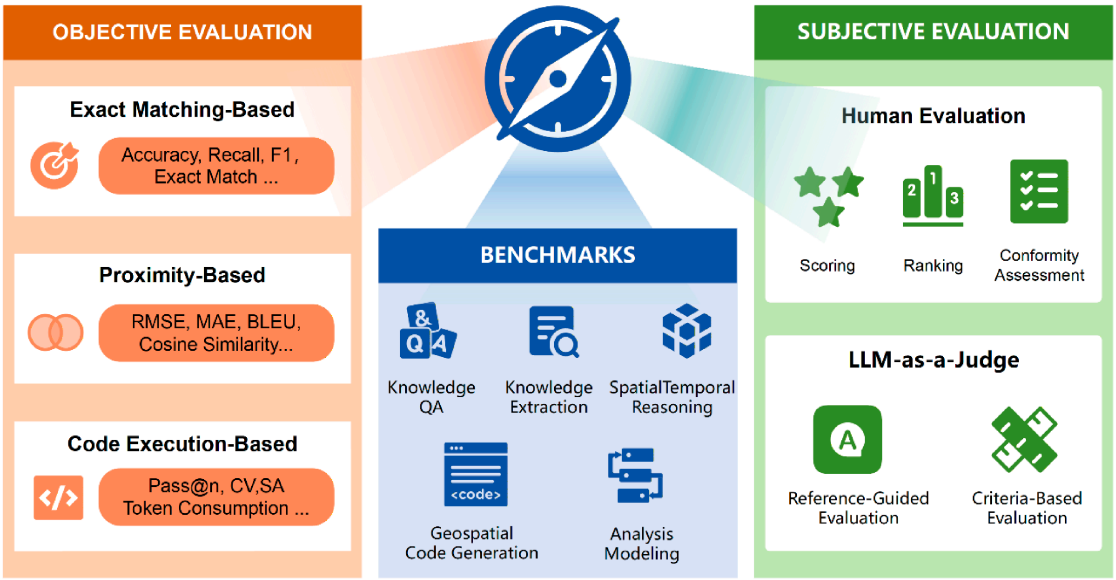


Figure 11. Overview of Evaluation Methods and Benchmarks.

Table 4. Representative LLM4GIS Benchmarks.

Benchmark	Evaluated Task	Evaluation Method	Dataset Scale
ClimaQA [238]	Climate science knowledge Q&A	Objective evaluation, human evaluation, LLM-as-Judge	566 climate science question–answer pairs
GeoBench [66]	Geoscience knowledge Q&A	Objective evaluation, human evaluation, LLM-as-Judge	1578 objective geoscience questions and 939 open-ended questions
VRSBench [239]	Remote sensing image captioning, visual grounding, and visual question answering	Objective evaluation, LLM-as-Judge	29,614 remote sensing images, 29,614 detailed captions, 52,472 referring expressions, and 123,221 question–answer pairs
GeoSQL-Bench [229]	Spatial query knowledge Q&A and GeoSQL generation	Objective evaluation	2380 multiple-choice and true/false questions on spatial query fundamentals, 7488 SQL generation tasks, and 4310 schema retrieval tasks
GeoCorpus-3 [240]	Geological named entity recognition	Objective evaluation	8923 instances in total, with 20% used as the test set
MapIQ [241]	Map understanding	Objective evaluation	14,706 map-understanding question–answer pairs
ClimateIQA [242]	Weather map image understanding	Objective evaluation, LLM-as-Judge	26,280 heatmaps and 762,120 question–answer pairs in total, with 20% used for testing
GEOBench-VLM [243]	Image understanding	Objective evaluation	More than 10,000 image-understanding instances across 31 task types, including scene understanding and building counting
MapEval [244]	Map understanding and geospatial reasoning	Objective evaluation, human evaluation, LLM-as-Judge	700 map-based visual reasoning questions, including 300 text/map-API-based questions and 400 visual questions
GeoArena [245]	Image geolocation	Human evaluation, LLM-as-Judge	Thousands of voting records, including user-uploaded images, model responses, and user judgments recorded as win/tie/loss outcomes

Table 4. *Cont.*

Benchmark	Evaluated Task	Evaluation Method	Dataset Scale
[246]	Toponym resolution and spatial relation reasoning	Objective evaluation	115 questions on toponym resolution and spatial relation reasoning
PPNL [166]	Path planning in grid environments	Objective evaluation	82,524 path-planning tasks with ground-truth optimal paths
GeoCode-Bench [182]	Geospatial code generation and code knowledge extraction	Objective evaluation, human evaluation, LLM-as-Judge	8000 question–answer items and 1000 programming tasks
AutoGEEval++ [225]	GEE code generation	Objective evaluation	6365 code-generation test cases
GeoAnalystBench [247]	Workflow generation and code generation	Objective evaluation, human evaluation	50 geoprocessing tasks
[248]	Geospatial code generation	Objective evaluation	20 geospatial tasks and 77 code-generation samples

Geospatial Knowledge Q&A, Geospatial Knowledge Extraction, Spatiotemporal Reasoning, Geospatial Code Generation, and Geospatial Analysis Modeling differ fundamentally in both their underlying capabilities and output forms, making it difficult to assess them using a unified set of metrics. Therefore, task-specific evaluation methods are required to capture the core capabilities of each application. Accordingly, this section examines the evaluation framework from the perspective of these five applications.

6.1. Evaluation for Geospatial Knowledge Q&A

Evaluation for Geospatial Knowledge Q&A primarily focuses on the accuracy of model responses and the depth of domain knowledge coverage, examining whether a model can correctly retrieve and generate geoscientific knowledge without explicit knowledge injection. A distinctive feature of geographic knowledge is that the "correct answer" is often not static: the administrative affiliation of a geographic entity may change following boundary adjustments, while climate or population statistics for the same region may vary with statistical definitions and temporal windows. Therefore, reference answers should explicitly specify spatiotemporal attributes, spatial scale, and other relevant contextual factors, rather than treating geographic knowledge as a collection of immutable facts.

For closed-ended questions with well-defined ground-truth answers, exact-match accuracy is widely used, as in K2 [66], while weighted accuracy can assign partial credit to responses that are not fully correct [249]. Examination-based evaluations may further combine letter grades with chi-square analysis to identify whether errors are concentrated in particular GIS question types or knowledge topics [172]. GeoGalactica [87] and CityGPT [97] additionally compare GIS-specific performance with general-purpose benchmarks such as MMLU [250], C-Eval [251], and GAOKAO-Bench [252] to distinguish domain-adaptation gains from underlying general capabilities. GPSBench [192] evaluates GPS-coordinate understanding using 57,800 samples across 17 tasks. Its results show that fine-tuning can improve geometric tasks such as polygon-area calculation while degrading performance on knowledge-intensive tasks such as boundary analysis and name disambiguation, revealing a trade-off between domain specialization and general knowledge.

For open-ended descriptive responses, study [253] applies BLEU [254] and ROUGE [255], while Typhoon-T5 [256] combines embedding-based cosine similarity, Jaccard similarity, and ROUGE to compare generated responses with reference answers. Human evaluation can complement these metrics through multidimensional criteria; for example, GeoGalactica [87] invited ten geoscience experts to rate scientific validity, correctness, and coherence on a 1–3 scale. For LLM-as-Judge evaluation, OceanGPT [88] mitigates positional bias by swapping the presentation order of competing outputs before calculating task- and instance-level win rates, illustrating a calibration practice that remains uncommon in GIS evaluation.

Benchmarks for Geospatial Knowledge Q&A have gradually evolved from general knowledge tests toward multidisciplinary and domain-specific scientific question answering, covering vertical domains such as climate science (ClimaQA [238]), Earth science (EarthSE [257]), ocean science (OceanBench [88]), geology (GeoEval [84]), and remote sensing visual question answering (VRSBench [239]). Geospatial Knowledge Q&A capability therefore serves as a foundational criterion for determining whether a model can support more complex downstream GIS tasks.

6.2. Evaluation for Geospatial Knowledge Extraction

Evaluation for Geospatial Knowledge Extraction primarily assesses the accuracy and completeness with which models extract structured geographic information from sources such as text, imagery, and maps. Compared with general information extraction tasks, GIS-oriented extraction presents two distinctive challenges. First, the same geographic feature may correspond to different valid labels under different classification schemes or spatial scales, making annotation consistency inherently difficult to guarantee. Second, the semantics of cartographic symbols are highly dependent on legend context, meaning that symbol recognition results cannot be meaningfully evaluated in isolation from the corresponding legend information.

For entity and relation extraction, the standard evaluation metrics are Precision, Recall, and F1-score. Both S-GMKG [139] and GEE-Ops [134] use these metrics to assess the quality of entity–relation extraction in geographic modeling knowledge graphs. However, exact matching is often overly restrictive for structured knowledge, because the same functional semantics can be expressed in multiple valid forms. To address this issue, Geo-FuB [133] introduces a composite metric that integrates semantic and structural correctness through a weighted combination, in which semantic correctness receives a higher weight than structural correctness. Semantic correctness is measured using Word2Vec [258] cosine similarity, and structural correctness is determined through hierarchical label matching. GEE-Ops [134] further introduces multiple structural similarity metrics for knowledge-chain extraction, including longest common subsequence (LCS), N-gram similarity, Siamese-based similarity, and BERT-based semantic similarity. These designs reflect an explicit trade-off between strict exact matching and tolerance for semantically equivalent outputs.

For spatial localization and extraction from imagery and maps, Intersection over Union (IoU) is the primary metric, although its specific formulation varies across tasks. Visual grounding tasks commonly regard predictions with IoU > 0.5 as correct, as in LHRS-Bot [81], or report threshold-based metrics such as acc@0.5 and Pr@0.6, as in GeoChat [187] and EarthGPT [80]. Object detection tasks instead adopt standard detection protocols such as AP@40 and AP@50 [80]. Address and toponym extraction tasks have also developed specialized composite error metrics. For example, Ref. [259] proposes Accuracy@161km, defined as the proportion of geocoding predictions whose errors fall within 161 km, together with Mean Error and the area under the error curve (AUC). These metrics characterize geocoding performance from the perspectives of tolerance, central tendency, and overall error distribution, representing a distinctly geographic evaluation design that is rarely encountered in general NLP entity-linking tasks. Multi-label attribute extraction, such as simultaneously identifying disaster type, affected population, and other fields from disaster-related text, is additionally susceptible to evaluation bias caused by class imbalance. To address this issue, FE-HS [131] combines Hamming Loss with macro-, micro-, and weighted-average F1 scores.

Existing benchmarks now span the full spectrum of extraction tasks, from unstructured text to complex cartographic symbols and remote sensing imagery. Ref. [260] evaluates models' ability to determine topological relations between spatial entities repre-

sented in Well-Known Text (WKT) format, MapIQ [241] assesses map understanding, and GeoMap-Bench [149] is designed for geological map understanding. GeoLLM-QA [261] and GEOBench-VLM [243] further evaluate the ability of models to perform visual extraction tasks such as object detection from remote sensing imagery. Overall, performance on Geospatial Knowledge Extraction tasks provides a direct measure of a model's capacity for structured understanding of complex geographic scenes.

6.3. Evaluation for Spatiotemporal Reasoning

Evaluation for Spatiotemporal Reasoning essentially examines whether a model can correctly reason about spatial relationships and spatiotemporal processes. Compared with general reasoning tasks, spatial reasoning evaluation must account for GIS-specific properties such as topological invariance, scale transformation, and coordinate system conversion. For predictive tasks, outputs must also satisfy basic physical constraints; for example, planned routes should not cross obstacles, and predicted population or flow values should not be negative. For decision-oriented tasks such as site selection and route navigation, however, solution quality is often highly context-dependent, making it difficult to define a unique ground-truth answer independent of the specific task. Consequently, human evaluation and LLM-as-Judge play particularly important roles in this class of tasks.

For tasks with deterministic answers, such as geolocation and directional judgment, distance-based error provides a more spatially meaningful measure than Accuracy alone. Reference [243] computes the mean distance between predicted and ground-truth coordinates using the Haversine formula, evaluating localization performance directly in terms of physical distance rather than categorical correctness. For continuous numerical prediction, mean absolute error (MAE), root mean square error (RMSE), and R^2 are among the most commonly used metrics. Orca [262] employs these measures for significant wave height estimation, while GeoLLM [221] uses Pearson's r^2 to quantify the linear association between predicted and observed socioeconomic indicators, including population density and asset wealth. SPOK [73] further combines RMSE, MAE, and R^2 to capture complementary aspects of spatiotemporal prediction accuracy. For location recommendation, the evaluation focus shifts from simple correctness to ranking quality. Ref. [263] adopts accuracy at k (ACC@k), mean reciprocal rank (MRR), and normalized discounted cumulative gain at 5 (NDCG@5) to assess the position of the true location within the model-generated candidate list, providing a finer-grained evaluation dimension that conventional classification accuracy cannot capture.

For spatiotemporal data generation and simulation, the evaluation objective shifts from whether a prediction matches a single ground truth to whether the generated outputs statistically resemble the real-world distribution. Chain-of-Planned-Behaviour [264] uses Jensen–Shannon divergence to measure distributional differences between generated and observed trajectories across statistics such as radius of gyration and the number of daily activity locations. This distribution-level perspective represents a key distinction between the evaluation of generative spatiotemporal tasks and that of discriminative prediction tasks. For agent tasks involving tool invocation, GeoLLM-Engine [169] employs a model checker to verify the functional correctness of task execution and uses Correctness Rate, defined as the proportion of correct function calls, together with Success Rate, defined as the proportion of successfully completed tasks, to assess the integrity of the decision-execution chain. In human evaluation, LAMP [265] invites GIS experts to score POI recommendations along three dimensions: factual validity, referring to whether the recommended place actually exists; spatial awareness, referring to whether it is located near the user; and semantic relevance, referring to whether it satisfies the query intent. This multidimensional

framework also provides a reusable evaluation paradigm for other geographic decision-making tasks.

In terms of benchmarks, CityBench [266] targets urban-scene tasks and includes evaluations such as image geolocation and trajectory prediction. SpatiaLab [267] further decomposes spatial reasoning into six major categories—relative localization, depth and occlusion, direction, scale, navigation, and 3D geometry—comprising 30 subcategories, making it one of the most fine-grained benchmarks currently available for spatial reasoning. These benchmarks evaluate model performance across multiple dimensions of spatial reasoning capability.

These benchmark results should nevertheless be interpreted in relation to their geographic coverage and data-partitioning protocols. CityBench evaluates eight urban tasks across 13 cities worldwide and reports performance across these city-specific settings; however, its overall evaluation is not organized as a unified held-out-city protocol specifically designed to test transfer to unseen regions. SpatiaLab evaluates fine-grained spatial reasoning using visual question–answer pairs collected from real-world scenes, rather than geographic or temporal transfer across regions. More broadly, spatial or temporal separation between model-development data and evaluation data, as well as explicit checks for overlap with model pretraining or instruction-tuning corpora, is not consistently reported across current benchmarks. Such contamination is particularly difficult to verify for proprietary models whose training data are undisclosed. Therefore, only evaluations using explicitly held-out regions or time periods can provide direct evidence of cross-region or cross-temporal generalization; otherwise, the reported results should be understood as performance within the existing benchmark settings.

6.4. Evaluation for Geospatial Code Generation

Evaluation for Geospatial Code Generation primarily examines whether generated GIS code or spatial query statements can be executed successfully and produce the expected results. The correctness of GIS code depends not only on syntactic validity but, more importantly, on whether spatial semantics are preserved, including the correctness of coordinate reference systems, the invariance of spatial relations during transformation, and the appropriateness of map projections. Moreover, the same geospatial task can often be implemented through multiple valid coding strategies, making simple text-level code matching insufficient for assessing code quality. Execution-based evaluation is therefore essential for determining whether the generated code is functionally correct in practice.

At the code-similarity level, ChatGeoAI [184] combines CodeBLEU [268], ChrF [269], ROUGE-L, and BERTScore [270] to assess lexical, structural, and semantic similarity between generated and reference code. For execution-based evaluation, AutoGEEval++ [225] combines pass@n (n = 1/3/5), the coefficient of variation, stability-adjusted accuracy, token consumption, and inference time to assess correctness, stability, and efficiency. CoP [130] uses Matchability, Executability, and Accuracy to diagnose requirement understanding, code execution, and output correctness, respectively. ChatGeoAI [184] further measures first-attempt success rate, average number of attempts, and permanent failure rate to evaluate interactive reliability.

At the human-evaluation level, both CoP [130] and GeoCode-GPT-7B [85] adopt blind review to assess code readability, with five experts providing scores and the highest and lowest ratings removed before averaging. GeoColab [155] further decomposes readability into four subdimensions—conciseness, clarity, comment accuracy, and logical consistency—and evaluates each separately. For LLM-as-Judge evaluation, GeoCode-Bench [182] uses DeepSeek-R1 as an automated evaluator for open-ended programming tasks, substantially reducing the cost of human assessment while maintaining a high level of agreement.

Benchmarks for Geospatial Code Generation now span multiple platforms and programming paradigms, including GeoSQL query generation (GeoSQL-Bench [229]), GEE code generation (AutoGEEval++ [225]), geospatial python code generation (GeoPyEval [271]), JavaScript-based mapping and spatial computing code generation (GeoJSEval [227]), and code-driven automated discovery of geospatial models (GeoEvolve [158]). Performance on these benchmarks directly reflects the practical value of LLMs in completing real-world GIS programming tasks.

6.5. Evaluation for Geospatial Analysis Modeling

Evaluation for Geospatial Analysis Modeling primarily focuses on the completeness of end-to-end workflows, the appropriateness of tool invocation, and the quality of final analytical outputs. Among the five applications, this category is the most open-ended, as task objectives often lack a unique ground-truth solution and objective metrics alone are insufficient to capture overall performance. Subjective evaluation therefore plays a dominant role. Cartographic tasks introduce additional aesthetic and professional criteria, including color schemes, legend design, and scale appropriateness, and the quality of an analytical solution often requires domain experts to judge it within a specific application context. Moreover, because such workflows involve coordinated use of multiple tools, evaluation must cover both the execution process and the final output.

At the quantifiable execution level, GeoLLM-Engine [169] employs a model checker to programmatically verify agent function calls and task execution results, reporting Correctness Rate and Success Rate. This provides an automated and reproducible evaluation pathway for otherwise highly open-ended workflow tasks. In human evaluation, GeoGPT [8] invited five GIS master's-level experts to inspect framework outputs individually and determine whether they accurately satisfied the query requirements. GeoTool-GPT [93] similarly adopts compliance-based assessment to examine whether generated tool-invocation plans meet task requirements. MapGPT [150] focuses specifically on cartographic output and applies a scoring scheme to quantify the aesthetic quality of generated maps. At the LLM-as-Judge level, GeoGraphRAG [138] uses an LLM to jointly assess the semantic similarity and structural consistency between generated modeling solutions and reference results, providing a scalable evaluation mechanism for analytical tasks without unique answers. However, LLM-as-Judge is particularly prevalent at this level, and its probabilistic generation mechanism also introduces evaluation instability, as repeated judgments on the same input may yield inconsistent conclusions.

In this highly open-ended task category, some studies, such as 5-HiNet [272] and selected experiments in GeoCogent [171], adopt case studies as either the primary or a complementary evaluation approach. Rather than relying on a single quantitative metric, these studies demonstrate the feasibility of a method by presenting complete analytical workflows and outputs in concrete tasks. Although such case-based evaluation is limited in terms of comparability and large-scale validation, it remains a widely used and practically persuasive supplementary approach at a stage when unified benchmarks for Geospatial Analysis Modeling are still lacking.

In terms of benchmarks, GeoAnalystBench [247] evaluates analytical modeling capability using 50 expert-validated Python geoprocessing tasks derived from real-world GIS problems, assessing performance from both workflow-generation and code-semantics perspectives. GeoBenchX [273] targets the multi-step problem-solving capability of geospatial agents and contains more than 200 tasks across four categories, including data integration and visualization, spatial operations, heatmap generation, and contour mapping. GeoJSON Agents [154] organizes 70 tasks into basic, intermediate, and advanced levels according to task complexity. Drawn from authoritative GIS textbooks and recent research, the bench-

mark evaluates hierarchical capabilities in areas such as data storage and transformation, spatial relation analysis, and spatial query and retrieval. For cartography specifically, [274] constructs a dedicated cartographic capability benchmark spanning six dimensions: map design, map projection, cartographic generalization, spatial databases, map visualization, and computational cartography. This benchmark helps address a gap in existing workflow-oriented evaluations, which tend to emphasize code and process correctness while giving limited attention to cartographic aesthetics and professional standards. In addition, UPBench [275] directly compares LLM-generated urban planning solutions with outputs produced by professional planners, introducing an expert-referenced evaluation paradigm rather than a conventional ground-truth-based one. This helps compensate for the long-standing reliance on case studies and the lack of quantitative comparative benchmarks in this task category. Collectively, these benchmarks reflect the integrated capability of models to solve realistic and open-ended GIS analysis tasks.

Although existing LLM4GIS evaluation frameworks have made substantial progress across the major applications, they remain limited in coverage, scale, and evaluation reliability. The field still lacks large-scale, comprehensive, and widely recognized benchmarks comparable to MMLU [250] and GPQA [276] in the general-purpose LLM domain. Future evaluation research should therefore focus on constructing larger datasets that cover broader geographic regions and more complex task types, while establishing widely accepted evaluation standards and automated assessment frameworks. In particular, as geospatial agents become increasingly autonomous, future benchmarks should more closely simulate real-world GIS analysis tasks and assess autonomous decision-making and execution capabilities in complex task planning, error correction, and multi-turn interaction.

7. Challenges and Prospects

LLM4GIS is currently at a critical stage of transition from early application-oriented exploration toward systematic methodological innovation. LLMs have demonstrated broad potential across core applications, including Geospatial Knowledge Q&A, Geospatial Knowledge Extraction, Spatiotemporal Reasoning, Geospatial Code Generation, and Geospatial Analysis Modeling. Nevertheless, substantial challenges remain in areas such as data foundations, model capabilities, and the reliability of real-world deployment. This section examines the major research bottlenecks and future directions from five perspectives: multimodal data integration, high-quality data construction, domain-specific model development, integration with world models, and interpretability and trustworthiness, with the aim of informing the development of a new generation of spatiotemporal intelligence ecosystems.

Unified integration of multimodal spatiotemporal data. GIS data encompass diverse modalities, including text, remote sensing imagery, vector data, raster data, and spatiotemporal trajectories. These modalities differ fundamentally in data structure, coordinate reference systems, spatiotemporal resolution, and semantic granularity. Moreover, multimodal alignment in geospatial contexts must simultaneously address spatial reference harmonization, timestamp alignment, and scale effects. Existing studies typically employ modality-specific encoders and adaptation layers; however, such architectures do not fundamentally resolve the challenge of enabling models to understand heterogeneous geospatial modalities. As the number and diversity of input modalities increase, their capacity for joint cross-modal reasoning becomes increasingly constrained. Future research should therefore explore unified tokenization and encoding mechanisms capable of natively representing multiple types of geospatial data, thereby internalizing multimodal fusion as a fundamental model capability rather than treating it as an external preprocessing step.

High-quality data and standardization. High-quality annotated data for GIS tasks remain insufficient across model training, fine-tuning, and evaluation. Geospatial domain-specific content constitutes only a small fraction of general pretraining corpora, limiting models' understanding of coordinate representations, spatial topological relations, and geoscientific terminology. Meanwhile, instruction data containing explicit reasoning steps and tool-use trajectories, which are essential for fine-tuning and reinforcement learning, often require substantial expert involvement and are therefore costly to construct and difficult to scale. At the evaluation level, comprehensive benchmarks spanning multiple tasks and geographic regions remain scarce, making direct comparison across studies challenging. Furthermore, the intrinsic spatial heterogeneity of geospatial data leads to dense data coverage in developed regions but persistent data gaps in remote and underrepresented areas. Such geographic imbalances are readily inherited by downstream models, constraining their reliability in data-sparse regions. Future research should establish standardized annotation protocols for multi-regional and multi-scale spatiotemporal data, explore LLM-assisted data synthesis and automated annotation, and promote open and shared data and evaluation platforms to provide a common foundation for model training, benchmarking, and cross-team comparison in GIS.

Exploration of domain-specific small models for GIS. General-purpose LLMs offer clear advantages in open-ended question answering and cross-task generalization, but their large parameter counts and deployment costs are not always optimal for well-defined GIS tasks such as remote sensing classification, traffic prediction, and address parsing. Domain-specific small models trained for targeted tasks may achieve higher accuracy with substantially lower computational overhead. Recent studies have developed specialized models with relatively small parameter scales for tasks such as geoscience question answering, remote sensing understanding, and geospatial code generation. However, most of these models are designed for individual tasks, exhibit limited cross-task transferability, and remain susceptible to degradation of general capabilities during domain knowledge injection. Future research should therefore explore GIS-oriented small-model architectures that combine domain specialization with task adaptivity. Strategies such as mixture-of-experts and dynamic routing could enable multi-task coverage within constrained parameter budgets, while knowledge distillation could transfer the reasoning capabilities of large models into lightweight models deployable on edge devices, thereby lowering the barrier to practical GIS intelligent analysis.

Integration with world models and real-world environments. LLMs acquire knowledge primarily from statistical co-occurrence patterns in training corpora and therefore lack an intrinsic understanding of the governing laws of the physical world. As a result, in tasks involving geographic process simulation and spatiotemporal forecasting, they may produce outputs that violate physical conservation laws or spatial topological constraints. Future research should explore incorporating domain knowledge, ranging from physical laws such as mass conservation and fluid-dynamics equations to empirical geographic regularities such as Tobler's First Law of Geography, into model loss functions or reasoning architectures. Beyond such knowledge constraints, LLMs could be further coupled with geospatial world models that characterize the states and dynamics of geographic environments, which may take the form of Earth system models, spatiotemporal foundation models, or simulation environments for agent interaction. Such integration would enable models to better perceive, understand, and support decision-making in real spatiotemporal environments, thereby improving the reliability and robustness of predictive outputs.

Improving interpretability and trustworthiness. The black-box nature and hallucination risks arising from the probabilistic generation mechanism of LLMs represent major barriers to deployment in high-stakes applications such as urban planning and disaster

response. Models may generate spatial-relation inferences that appear plausible but are factually incorrect, while repeated reasoning over identical inputs may yield inconsistent outputs, directly undermining decision-makers' trust. Existing approaches such as RAG and chain-of-thought prompting can improve output traceability to some extent, but current models still have limited ability to quantify predictive uncertainty or explain why the same variable relationships exhibit heterogeneous patterns across geographic regions. Future LLM4GIS research should therefore incorporate causal inference mechanisms and further strengthen spatial interpretability by providing transparent reasoning traces, uncertainty quantification, and spatially aware explanatory visualizations. These capabilities are essential for ensuring that decision-making processes are transparent, rigorous, and trustworthy, thereby enabling more reliable support for scientific discovery and societal governance.

8. Conclusions

Positioned at a critical juncture in the transition of GIS analysis from a data-driven to an AI-driven paradigm, this study provides a comprehensive review of the evolution and technical landscape of LLM4GIS. Through a quantitative analysis and in-depth synthesis of 673 core publications released between 2022 and July 2026, we trace the field's development from sporadic early exploration to rapid expansion, establish a four-dimensional taxonomy encompassing technical approaches, datasets, applications, and evaluation, and systematically examine the progress of LLMs across five core areas: Geospatial Knowledge Q&A, Geospatial Knowledge Extraction, Spatiotemporal Reasoning, Geospatial Code Generation, and Geospatial Analysis Modeling. The findings indicate that the technical trajectory of LLM4GIS has evolved from an early paradigm dominated by prompt engineering into a comprehensive framework integrating architectural design, multi-stage model training, and multi-level application strategies. In this process, LLMs are shifting from auxiliary tools for GIS analysis toward becoming a core infrastructure for next-generation GIS.

Looking ahead, the continued development of LLM4GIS will require coordinated advances across the data, model, and application levels. At the data level, unified multimodal integration mechanisms and high-quality data standards are needed to provide models with a comprehensive foundation for spatiotemporal understanding. At the model level, it is necessary to move beyond the capability limits imposed by statistical corpus-driven learning by developing domain-specific models and incorporating physical laws, thereby achieving both domain expertise and physical consistency. At the application level, interpretable, quantifiable, and verifiable trust mechanisms must be established so that LLM4GIS outputs can reliably support scientific decision-making in high-stakes scenarios. Coordinated progress across these three levels will facilitate the transformation of GIS analysis from a complex expert-oriented toolset into broadly accessible spatiotemporal intelligence infrastructure.

The significance of LLM4GIS extends beyond using LLMs to transform GIS analytical methods; it also points toward a bidirectional vision of GIS4LLM. The rigorous spatiotemporal topological logic, structured knowledge systems, and accumulated understanding of the physical world within GIS can, in turn, enhance the spatial cognition and reasoning capabilities of large language models. The bidirectional integration of LLM4GIS and GIS4LLM has the potential to foster an open, collaborative, and intelligent ecosystem for geographic information science, strengthening human understanding of complex Earth systems and supporting more informed decisions for sustainable development.

Author Contributions: Conceptualization, Huayi Wu, Shuyang Hou, Rui Li, Zhipeng Gui and Xuefeng Guan; methodology, Huayi Wu, Lutong Xie, Shuyang Hou and Yaxian Qing; software, Shuyang Hou, Shaowen Wu, Guanyu Chen, Jianyuan Liang; validation, Lutong Xie, Haoyue Jiao,

Ziqi Liu and Jianyuan Liang; formal analysis, Huayi Wu, Lutong Xie, Donghai Liu, Yaxian Qing, Rui Li and Zhipeng Gui; investigation, Huayi Wu, Zhangxiao Shen and Shaowen Wu; data curation, Zhangxiao Shen, Haoyue Jiao, Guanyu Chen and Donghai Liu; writing—original draft preparation, Lutong Xie, Zhangxiao Shen and Haoyue Jiao; writing—review and editing, Lutong Xie and Shuyang Hou; visualization, Lutong Xie, Haoyue Jiao, Shaowen Wu and Ziqi Liu; supervision, Huayi Wu and Xuefeng Guan; project administration, Huayi Wu and Xuefeng Guan; funding acquisition, Shuyang Hou. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China [grant number 425B2013], awarded to Shuyang Hou.

Data Availability Statement: The complete list of the studies included and relevant information is publicly available at https://github.com/LuojiaGeoCode/Awesome-LLM4GIS (accessed on 10 October 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Full Names of Publication Venues in Figure 2

Table A1. Full names of the publication venues represented by abbreviations in Figure 2.

Abbreviation/Label in Figure 2	Full Venue Name
arXiv	arXiv preprint repository
ACM SIGSPATIAL	ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems
IJGIS	International Journal of Geographical Information Science
ISPRS IJGI	ISPRS International Journal of Geo-Information
ACL	Annual Meeting of the Association for Computational Linguistics
IJDE	International Journal of Digital Earth
TGIS	Transactions in GIS
NeurIPS	Conference on Neural Information Processing Systems
JAG	International Journal of Applied Earth Observation and Geoinformation
EMNLP	Conference on Empirical Methods in Natural Language Processing
CEUS	Computers, Environment and Urban Systems
ISPRS JP&RS	ISPRS Journal of Photogrammetry and Remote Sensing
KDD	ACM SIGKDD Conference on Knowledge Discovery and Data Mining
Cities	Cities
ICCV	IEEE/CVF International Conference on Computer Vision
ICLR	International Conference on Learning Representations
AAAI	AAAI Conference on Artificial Intelligence
BED	Big Earth Data
CVPR	IEEE/CVF Conference on Computer Vision and Pattern Recognition
Remote Sens.	Remote Sensing

Appendix B. Bibliographic Mapping of the Studies Shown in Figures 6 and 10

Table A2. Bibliographic mapping of the studies shown in Figure 6.

Technical Approach Category	Studies
Tokenization	JiuZhou [64]; PreparedLLM [65]; K2 [66]; LandGPT [68]; TourismMinds [69]; SafeTraffic Copilot [70]; LGSID [71]; SPOK [73]; QT-Mob [72]; EarthMind [76]; GeoLLAVA-8K [77]; Traj-MLLM [78]; UrbanGPT [164]; MobGLM [217].
Model Architecture	RSGPT [79]; EarthGPT [80]; LHRS-Bot [81]; DriveGPT4 [82]; TourismMinds [69]; SkyMoE [83]; GeoProspect [84]; GeoChat [187]; SkyEyeGPT [277]; TransGPT [278]; GeoReasoner [279]; GeoLM [280]; REO-VLM [281]; GAEA [282]; UrbanMLLM [214].

Table A2. *Cont.*

Technical Approach Category	Studies
Pretraining	ClimateGPT [63]; GeoCode-GPT [85]; K 2 [66]; GeoGalactica [87]; OceanGPT [88]; PlanGPT [145]; CityGPT [97]; LHRS-Bot-Nova [91]; EarthMind [76]; BB-GeoGPT [123]; EarthDial [283]; GeoFactory [137]; OceanAgent [284]; RoadMind [285]; EarthGPT [80].
Supervised Fine-tuning	GeoCode-GPT [85]; GeoTool-GPT [93]; GRE Suite [94]; RadarQA [95]; GloRA [96]; DeepTravel [108]; PIGEON [99]; LightGPT [100]; USTBench [101].
Reinforcement Learning	CityGPT [97]; CAMS [111]; LGSID [71]; GRE Suite [94]; RadarQA [95]; Urban-R1 [114]; Geo-R1 [191]; CityRiSE [116]; RSThinker [104]; Drive-R1 [105]; VMLocPredictor [117]; AddrLLM [103]; Traffic-R1 [122]; DeepTravel [108]; STAgent [107].
Prompt Engineering	CoP [130]; FE-HS [131]; GPT4GEO [62]; LLM-Mob [61]; LLM-Move [193]; FloodDepth-GPT [286]; GeoBikeLLM [287]; T2TrajLLM [288]; Monkuu [228]; GDQA [289]; GeoGPT [8]; TrajEvo [290]; UrbanCLIP [196]; GeoQAMap [163].
Retrieval-Augmented Generation	Geo-FuB [133]; GEE-Ops [134]; GeoFactory [137]; GeoGraphRAG [138]; SpatialGPT [140]; Spatial-RAG [142]; PIGEON [99]; LLMTrajQuery [143]; UGuideRAG [291]; AskNearby [146].
LLM-based Agents	Earth-Agent [148]; STAgent [107]; PEACE [149]; PlanGPT [145]; MapGPT [150]; TrafficGPT [151]; GeoAgent [106]; IndustryScopeGPT [153]; GeoJSON Agents [154]; GeoColab [155]; CartoAgent [156]; smileGeo [157]; GeoEvolve [158]; CitySim [159].

Note: This table provides the bibliographic mapping for all studies displayed in Figure 6. Studies associated with more than one technical approach category are listed in each relevant row.

Table A3. Bibliographic mapping of the studies shown in Figure 10.

Application Category	Studies
Geospatial Knowledge Q&A	STRIDE-Cosmos-Reason1-7B [292]; GAEA [282]; MFC4POI [293]; WaterCopilot [202]; GeoIR-Compiler [294]; OGD4All [295]; CARAG [296]; Change-LISA [297]; MVLT-LoRA-CC [298]; GeoRAG [299]; WorldBench [300]; LandslideAgent [301]; DigitalTurku [302]; ComPlanAI [303]; HydroLLM [304]; PH-LLM [305]; ChatPlanner [306]; pykci [307]; GIScholarBench [308]; RSICCLLM [309].
Geospatial Knowledge Extraction	FloodKG [310]; OceanAgent [284]; RoadEventGPT [311]; LLM-GeoTextCog [312]; crisesStorylinesRAG [313]; GeoGraphRAG [138]; SkyEyeGPT [277]; MapReader [314]; OmniGeo [220]; GeoMag [315]; StreetViewAI [316]; UrbanSense [317]; SAGAI [318]; PlanGPT-VL [319]; EarthMind [76]; MapBot [320]; HierLabelNet [321]; SegEarth-R1 [322].
Spatiotemporal Reasoning	LGSID [71]; S2SServiceBench [323]; Refine-POI [324]; MFC4POI [293]; PSLM4ST [325]; GeoPilot [326]; Geo-ADAPT [120]; GeoArena [327]; LandslideAgent [301]; MAS-LAND [328]; Geo-MAG [329]; RoadEventGPT [311]; VectorLLM [330]; OpenEarthAgent [331]; ClimAgent [200]; RemoteShield [332].
Geospatial Code Generation	GeoJSON Agents [154]; GIS Chat [333]; PostGISer-7B [92]; GeoPyEval [271]; GISclaw [46]; SpatialSQL [334]; GeoColab [155]; AutoGEEval [224]; AutoGEEval++ [225]; GeoJSEval [227]; GeoSQL-Bench [229]; GeoCode-GPT-7B [85]; Chain-of-Programming [130]; Geo-FuB [133]; GEE-Ops [134]; GeoCogent [171]; GIS Copilot [60]; GeoPandas-AI [226]; ShapefileGPT [195]; 5-HiNet [272]; ChatGeoAI [184].
Geospatial Analysis Modeling	FloodKG [310]; GATSim [335]; GenWorld [336]; SimPOL [337]; Flood-LLM [338]; MobCache [339]; HydroLLM [304]; RS-Claw [340]; OpenEarth-Agent [234]; MapGPT [150]; MapColorAI [341]; GISclaw [46]; ST-Attn-LLM [342]; VectorLLM [330]; Topo4Vec [343]; ClimAgent [200]; UPBench [275]; TrajGenAgent [344]; UrbanDS [345]; GVSABench [346]; Intelli-Planner [347].

Note: This table provides the bibliographic mapping for all studies displayed in Figure 10. Studies associated with more than one application category are listed in each relevant row.

References

1. Goodchild, M.F. Geographical information science. *Int. J. Geogr. Inf. Syst.* **1992**, *6*, 31–45. [CrossRef]
2. Goodchild, M.F. Geographic information systems and science: Today and tomorrow. *Ann. GIS* **2009**, *15*, 3–9. [CrossRef]
3. Fradelos, E.C.; Papathanasiou, I.V.; Mitsi, D.; Tsaras, K.; Kleisiaris, C.F.; Kourkouta, L. Health based geographic information systems (GIS) and their applications. *Acta Inform. Medica* **2014**, *22*, 402. [CrossRef] [PubMed]
4. Wang, F. Why public health needs GIS: A methodological overview. *Ann. GIS* **2020**, *26*, 1–12. [CrossRef] [PubMed]
5. Trung Hieu, H.; Le, K.G. GIS in transportation infrastructure: A systematic review and bibliometric analysis of construction, management, and digital integration. *All. Earth* **2026**, *38*, 2637588. [CrossRef]
6. Mai, G.; Xie, Y.; Jia, X.; Lao, N.; Rao, J.; Zhu, Q.; Liu, Z.; Chiang, Y.-Y.; Jiao, J. Towards the next generation of Geospatial Artificial Intelligence. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *136*, 104368. [CrossRef]
7. Li, J.; Hong, D.; Gao, L.; Yao, J.; Zheng, K.; Zhang, B.; Chanussot, J. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102926. [CrossRef]
8. Zhang, Y.; Wei, C.; He, Z.; Yu, W. GeoGPT: An assistant for understanding and processing geospatial tasks. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *131*, 103976. [CrossRef]
9. Li, W.; Hsu, C.-Y.; Wang, S.; Kedron, P. GeoAI reproducibility and replicability: A computational and spatial perspective. *Ann. Am. Assoc. Geogr.* **2024**, *114*, 2085–2103. [CrossRef]
10. Li, Z.; Ning, H. Autonomous GIS: The next-generation AI-powered GIS. *Int. J. Digit. Earth* **2023**, *16*, 4668–4686. [CrossRef]
11. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774.
12. Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F. Qwen technical report. *arXiv* **2023**, arXiv:2309.16609.
13. Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C. Deepseek-v3 technical report. *arXiv* **2024**, arXiv:2412.19437.
14. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S. Llama 2: Open foundation and fine-tuned chat models. *arXiv* **2023**, arXiv:2307.09288.
15. Minaee, S.; Mikolov, T.; Nikzad, N.; Chenaghlu, M.; Socher, R.; Amatriain, X.; Gao, J. Large language models: A survey. *arXiv* **2024**, arXiv:2402.06196.
16. Wei, J.; Bosma, M.; Zhao, V.Y.; Guu, K.; Yu, A.W.; Lester, B.; Du, N.; Dai, A.M.; Le, Q.V. Finetuned language models are zero-shot learners. *arXiv* **2021**, arXiv:2109.01652.
17. Acharya, D.B.; Kuppan, K.; Divya, B. Agentic AI: Autonomous intelligence for complex goals—A comprehensive survey. *IEEE Access* **2025**, *13*, 18912–18936. [CrossRef]
18. Zhang, K.; Zuo, Y.; He, B.; Sun, Y.; Liu, R.; Jiang, C.; Fan, Y.; Tian, K.; Jia, G.; Li, P. A survey of reinforcement learning for large reasoning models. *arXiv* **2025**, arXiv:2509.08827.
19. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Dong, Z.; Hou, Y.; Zhang, B.; Min, Y.; Zhang, J.; Liu, P. A survey of large language models. *Front. Comput. Sci.* **2026**, *20*, 2012627. [CrossRef]
20. Sahoo, S.S.; Plasek, J.M.; Xu, H.; Uzuner, Ö.; Cohen, T.; Yetisgen, M.; Liu, H.; Meystre, S.; Wang, Y. Large language models for biomedicine: Foundations, opportunities, challenges, and best practices. *J. Am. Med. Inform. Assoc.* **2024**, *31*, 2114–2124. [CrossRef] [PubMed]
21. Abdel-Rehim, A.; Zenil, H.; Orhobor, O.; Fisher, M.; Collins, R.J.; Bourne, E.; Fearnley, G.W.; Tate, E.; Smith, H.X.; Soldatova, L.N. Scientific hypothesis generation by large language models: Laboratory validation in breast cancer treatment. *J. R. Soc. Interface* **2025**, *22*, 20240674. [CrossRef] [PubMed]
22. Lei, G.; Docherty, R.; Cooper, S.J. Materials science in the era of large language models: A perspective. *Digit. Discov.* **2024**, *3*, 1257–1272. [CrossRef] [PubMed]
23. Li, C.; Deng, W.; Lu, M.; Yuan, B. AtmosSci-Bench: Evaluating the recent advance of large language model for atmospheric science. *Adv. Neural Inf. Process. Syst.* **2025**, *38*.
24. Hu, M.; Ma, C.; Li, W.; Xu, W.; Wu, J.; Hu, J.; Li, T.; Zhuang, G.; Liu, J.; Lu, Y. A survey of scientific large language models: From data foundations to agent frontiers. *arXiv* **2025**, arXiv:2508.21148.
25. Li, X.; Guo, Y. Paradigm shifts from data-intensive science to robot scientists. *Sci. Bull.* **2025**, *70*, 14–18. [CrossRef] [PubMed]
26. An, C.; Wang, J.; Zhou, C. The evolution and current landscape of AI in geographical research: A large-scale systematic review. *Geogr. Sustain.* **2026**, *7*, 100392. [CrossRef]
27. Chen, M.; Qian, Z.; Boers, N.; Creutzig, F.; Camps-Valls, G.; Hubacek, K.; Claramunt, C.; Wilson, J.P.; Nativi, S.; Jakeman, A.J. Collaboration between artificial intelligence and Earth science communities for mutual benefit. *Nat. Geosci.* **2024**, *17*, 949–952. [CrossRef]
28. Zhao, T.; Wang, S.; Ouyang, C.; Chen, M.; Liu, C.; Zhang, J.; Yu, L.; Wang, F.; Xie, Y.; Li, J. Artificial intelligence for geoscience: Progress, challenges, and perspectives. *Innovation* **2024**, *5*, 100691. [CrossRef] [PubMed]

29. Li, Z.; Ning, H.; Gao, S.; Janowicz, K.; Li, W.; Arundel, S.T.; Yang, C.; Bhaduri, B.; Wang, S.; Zhu, A.-X. GIScience in the era of Artificial Intelligence: A research agenda towards Autonomous GIS. *Ann. GIS* **2025**, *31*, 501–536. [CrossRef]
30. Qadah, T.M.; Hammad, M.; Felemban, E. The landscape of LLM-powered geoanalysis. *IEEE Access* **2025**, *13*, 188335–188354. [CrossRef]
31. Hemdan, M.; Hussein, Y.; Mokbel, M.F. Large Language Models for Spatial Analysis Tasks. In Proceedings of the 1st ACM SIGSPATIAL International Workshop on Urban Mobility Foundation Models, Minneapolis, MN, USA, 3 November 2025; pp. 15–18.
32. Hussein, Y.; Hemdan, M.; Mokbel, M.F. Large language models for spatial analysis queries. *Proc. VLDB Endow.* **2025**, *18*, 5451–5454. [CrossRef]
33. Xue, X.; Zhu, X.X. Regression in Earth Observation: Are vision–language models up to the challenge? *IEEE Geosci. Remote Sens. Mag.* **2025**, *14*, 121–150. [CrossRef]
34. Talemi, N.A.; Boone, J.; Afghah, F. Agentic ai in remote sensing: Foundations, taxonomy, and emerging systems. *arXiv* **2026**, arXiv:2601.01891.
35. Zheng, Y.; Xu, F.; Lin, Y.; Santi, P.; Ratti, C.; Wang, Q.R.; Li, Y. Urban planning in the era of large language models. *Nat. Comput. Sci.* **2025**, *5*, 727–736. [CrossRef] [PubMed]
36. Fu, X.; Li, C.; Quan, S.J.; Yigitcanlar, T.; Wasserman, D. Large language models in urban planning. *Nat. Cities* **2025**, *2*, 585–592. [CrossRef]
37. Janowicz, K.; Mai, G.; Huang, W.; Zhu, R.; Lao, N.; Cai, L. GeoFM: How will geo-foundation models reshape spatial data science and GeoAI? *Int. J. Geogr. Inf. Sci.* **2025**, *39*, 1849–1865. [CrossRef]
38. Lu, F.; Cheng, S.; Wang, P. GeoAI enabled urban computing: Status and challenges. *Ann. GIS* **2025**, *31*, 537–555. [CrossRef]
39. Zou, X.; Yan, Y.; Hao, X.; Hu, Y.; Wen, H.; Liu, E.; Zhang, J.; Li, Y.; Li, T.; Zheng, Y. Deep learning for cross-domain data fusion in urban computing: Taxonomy, advances, and outlook. *Inf. Fusion* **2025**, *113*, 102606. [CrossRef]
40. Li, X. A review of prominent paradigms for llm-based agents: Tool use, planning (including rag), and feedback learning. In Proceedings of the 31st International Conference on Computational Linguistics, Abu Dhabi, United Arab Emirates, 27–28 January 2025; pp. 9760–9779.
41. Xu, W.; Huang, C.; Gao, S.; Shang, S. LLM-Based Agents for Tool Learning: A Survey. *Data Sci. Eng.* **2025**, *10*, 533–563. [CrossRef]
42. Ning, H.; Li, Z.; Akinboyewa, T.; Lessani, M.N. An autonomous GIS agent framework for geospatial data retrieval. *Int. J. Digit. Earth* **2025**, *18*, 2458688. [CrossRef]
43. Hou, X.; Zhao, Y.; Wang, S.; Wang, H. Model context protocol (mcp): Landscape, security threats, and future research directions. *ACM Trans. Softw. Eng. Methodol.* **2025**, *35*, 1–37. [CrossRef]
44. Wang, S.; Hu, T.; Xiao, H.; Li, Y.; Zhang, C.; Ning, H.; Zhu, R.; Li, Z.; Ye, X. GPT, large language models (LLMs) and generative artificial intelligence (GAI) models in geospatial science: A systematic review. *Int. J. Digit. Earth* **2024**, *17*, 2353122. [CrossRef]
45. Van de Weghe, N.; De Sloover, L.; Cohn, A.G.; Huang, H.; Scheider, S.; Sieber, R.E.; Timpf, S.; Claramunt, C. Opportunities and challenges of integrating geographic information science and large language models. *J. Spat. Inf. Sci.* **2025**, *30*, 93–116. [CrossRef]
46. Han, J.; Lee, J.; Shim, Y.; Kim, J.; Lee, J.-J. GISclaw: A Comprehensive Open-Source LLM Agent System for Realistic Multi-Step Geospatial Analysis. *arXiv* **2026**, arXiv:2603.26845.
47. Wu, H.Y.; Shen, Z.X.; Hou, S.Y.; Liang, J.Y.; Zhao, A.Q.; Jiao, H.Y.; Guan, X.F. Large language model-driven GIS analysis: Methods, applications, and prospects. *Acta Geod. Cartogr. Sin.* **2025**, *54*, 621–635. [CrossRef]
48. Oliveira, A.; Fachada, N.; Matos-Carvalho, J.P. Data science for geographic information systems. In Proceedings of the 2024 8th International Young Engineers Forum on Electrical and Computer Engineering (YEF-ECE), Caparica/Lisbon, Portugal, 5 July 2024; pp. 1–7.
49. McLeod, K.S. Our sense of Snow: The myth of John Snow in medical geography. *Soc. Sci. Med.* **2000**, *50*, 923–935. [CrossRef] [PubMed]
50. Brody, H.; Rip, M.R.; Vinten-Johansen, P.; Paneth, N.; Rachman, S. Map-making and myth-making in Broad Street: The London cholera epidemic, 1854. *Lancet* **2000**, *356*, 64–68. [CrossRef] [PubMed]
51. Cressie, N. *Statistics for Spatial Data*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
52. Anselin, L. Local indicators of spatial association—LISA. *Geogr. Anal.* **1995**, *27*, 93–115. [CrossRef]
53. Getis, A.; Ord, J.K. The analysis of spatial association by use of distance statistics. *Geogr. Anal.* **1992**, *24*, 189–206. [CrossRef]
54. Foresman, T.W. *The History of Geographic Information Systems: Perspectives from the Pioneers*; Prentice Hall: Upper Saddle River, NJ, USA, 1998.
55. Malczewski, J. GIS-based multicriteria decision analysis: A survey of the literature. *Int. J. Geogr. Inf. Sci.* **2006**, *20*, 703–726. [CrossRef]
56. Carver, S.J. Integrating multi-criteria evaluation with geographical information systems. *Int. J. Geogr. Inf. Syst.* **1991**, *5*, 321–339. [CrossRef]

57. Wang, S.; Huang, X.; Liu, P.; Zhang, M.; Biljecki, F.; Hu, T.; Fu, X.; Liu, L.; Liu, X.; Wang, R. Mapping the landscape and roadmap of geospatial artificial intelligence (GeoAI) in quantitative human geography: An extensive systematic review. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *128*, 103734. [CrossRef]
58. Lee, K.; Eo, M.; Jung, E.; Yoon, Y.; Rhee, W. Short-term traffic prediction with deep neural networks: A survey. *IEEE Access* **2021**, *9*, 54739–54756. [CrossRef]
59. Liu, B.; Deng, Y.; Li, M.; Yang, J.; Liu, T. Classification schemes and identification methods for urban functional zone: A Review of Recent Papers. *Appl. Sci.* **2021**, *11*, 9968. [CrossRef]
60. Akinboyewa, T.; Li, Z.; Ning, H.; Lessani, M.N. GIS copilot: Towards an autonomous GIS agent for spatial analysis. *Int. J. Digit. Earth* **2025**, *18*, 2497489. [CrossRef]
61. Wang, X.; Fang, M.; Zeng, Z.; Cheng, T. Where would i go next? large language models as human mobility predictors. *arXiv* **2023**, arXiv:2308.15197.
62. Roberts, J.; Lüddecke, T.; Das, S.; Han, K.; Albanie, S. GPT4GEO: How a language model sees the world's geography. *arXiv* **2023**, arXiv:2306.00020.
63. Thulke, D.; Gao, Y.; Pelser, P.; Brune, R.; Jalota, R.; Fok, F.; Ramos, M.; Van Wyk, I.; Nasir, A.; Goldstein, H. Climategpt: Towards ai synthesizing interdisciplinary research on climate change. *arXiv* **2024**, arXiv:2401.09646.
64. Chen, Z.; Lin, M.; Zang, M.; Wang, Z.; Li, J.; Bai, Y. JiuZhou: Open foundation language models and effective pre-training framework for geoscience. *Int. J. Digit. Earth* **2025**, *18*, 2449708. [CrossRef]
65. Chen, Z.; Lin, M.; Wang, Z.; Zang, M.; Bai, Y. PreparedLLM: Effective pre-pretraining framework for domain-specific large language models. *Big Earth Data* **2024**, *8*, 649–672. [CrossRef]
66. Deng, C.; Zhang, T.; He, Z.; Chen, Q.; Shi, Y.; Xu, Y.; Fu, L.; Zhang, W.; Wang, X.; Zhou, C.; et al. K2: A foundation language model for geoscience knowledge understanding and utilization. In Proceedings of the 17th ACM International Conference on Web Search and Data Mining, Yucatán, Mexico, 4–8 March 2024; pp. 161–170.
67. Hou, S.; Jiao, H.; Liu, Z.; Liang, J.; Xie, L.; Guan, X.; Gui, Z.; Wu, H. GEEToker: You can boost LLMs' JavaScript-based geospatial code generation for Google Earth Engine with a model-agnostic tokenizer. *Int. J. Geogr. Inf. Sci.* **2026**, 1–49. [CrossRef]
68. Zhu, G.; Tang, M.; Ma, Y.; Hu, Z.; Yu, C.; Zhang, X.; Hu, H.; Guan, Q.; Yao, Y. LandGPT: A multimodal large Language model for parcel-level land use classification with multi-source data. *Int. J. Geogr. Inf. Sci.* **2025**, *39*, 2862–2885. [CrossRef]
69. Ye, Z.; Su, X.; He, D.; Zhang, Z.; Sun, Y.; Ji, J.; Xie, X.; Wang, J.; Wei, R.; Zhang, Y. Tourismminds: A geo-augmented llm framework for semantic-aware trajectory analytics and generation. In Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, Espoo, Finland, 12–16 October 2025; Volume 9, pp. 1–32.
70. Zhao, Y.; Wang, P.; Zhao, Y.; Du, H.; Yang, H.F. SafeTraffic Copilot: Adapting large language models for trustworthy traffic safety assessments and decision interventions. *Nat. Commun.* **2025**, *16*, 8846. [CrossRef] [PubMed]
71. Jiang, H.; Wang, G.; Zhou, D.; Yu, S.; Zeng, Y.; Zeng, W.; Gai, K.; Zhou, G. Llm-aligned geographic item tokenization for local-life recommendation. In Proceedings of the AAAI Conference on Artificial Intelligence, Singapore, 20–27 January 2026; pp. 14928–14936.
72. Chen, Y.; Tao, Y.; Jiang, Y.; Liu, S.; Yu, H.; Cong, G. Enhancing large language models for mobility analytics with semantic location tokenization. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, Toronto, ON, Canada, 3–7 August 2025; pp. 262–273.
73. Huang, F.; Lv, J.-R.; Li, G.-L.; Yue, Y. SPOK: Tokenizing geographic space for enhanced spatial reasoning in GeoAI. *Int. J. Geogr. Inf. Sci.* **2025**, *39*, 2768–2808. [CrossRef]
74. Chen, W.; Wu, Y.; Chen, J.; Zhou, X.; Liang, Y. UrbanFM: Scaling urban spatio-temporal foundation models. *arXiv* **2026**, arXiv:2602.20677.
75. Meng, F.; Yuan, Y.; Ding, J.; Feng, J.; Han, C.; Li, Y. MoveFM-R: Advancing Mobility Foundation Models via Language-driven Semantic Reasoning. *arXiv* **2025**, arXiv:2509.22403.
76. Shu, Y.; Ren, B.; Xiong, Z.; Paudel, D.P.; Van Gool, L.; Demir, B.; Sebe, N.; Rota, P. Earthmind: Leveraging cross-sensor data for advanced earth observation interpretation with a unified multimodal llm. *arXiv* **2025**, arXiv:2506.01667.
77. Wang, F.; Chen, M.; Li, Y.; Wang, D.; Wang, H.; Guo, Z.; Wang, Z.; Boqi, S.; Lan, L.; Wang, Y. Geollava-8k: Scaling remote-sensing multimodal large language models to 8k resolution. *Adv. Neural Inf. Process. Syst.* **2026**, *38*, 159185–159218.
78. Liu, S.; Yao, D.; Lin, Y.; Cong, G.; Bi, J. Traj-MLLM: Can Multimodal Large Language Models Reform Trajectory Data Mining? In Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1, Jeju Island, Republic of Korea, 9–13 August 2026; pp. 903–914.
79. Hu, Y.; Yuan, J.; Wen, C.; Lu, X.; Liu, Y.; Li, X. Rsgpt: A remote sensing vision language model and benchmark. *ISPRS J. Photogramm. Remote Sens.* **2025**, *224*, 272–286. [CrossRef]
80. Zhang, W.; Cai, M.; Zhang, T.; Zhuang, Y.; Mao, X. EarthGPT: A universal multimodal large language model for multisensor image comprehension in remote sensing domain. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5917820. [CrossRef]

81. Muhtar, D.; Li, Z.; Gu, F.; Zhang, X.; Xiao, P. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; pp. 440–457.
82. Xu, Z.; Zhang, Y.; Xie, E.; Zhao, Z.; Guo, Y.; Wong, K.-Y.K.; Li, Z.; Zhao, H. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robot. Autom. Lett.* **2024**, *9*, 8186–8193. [CrossRef]
83. Liu, J.; Fu, R.; Sun, L.; Liu, H.; Yang, X.; Zhang, W.; Na, X.; Duan, Z.; Yang, B. SkyMoE: A vision-language foundation model for enhancing geospatial interpretation with mixture of experts. In Proceedings of the AAAI Conference on Artificial Intelligence, Singapore, 20–27 January 2026; pp. 7168–7178.
84. Wu, G.; Wang, H.; Zhang, K.; Zhu, S.; Cui, J.; Fan, M.; Wang, H.; Guo, H.; Zhang, G.; Zhang, D. GeoProspect: A domain-specific geological large language model with enhanced continual learning. *Neurocomputing* **2025**, *650*, 130801. [CrossRef]
85. Hou, S.; Shen, Z.; Zhao, A.; Liang, J.; Gui, Z.; Guan, X.; Li, R.; Wu, H. GeoCode-GPT: A large language model for geospatial code generation. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *138*, 104456. [CrossRef]
86. Roziere, B.; Gehring, J.; Gloeckle, F.; Sootla, S.; Gat, I.; Tan, X.E.; Adi, Y.; Liu, J.; Sauvestre, R.; Remez, T. Code llama: Open foundation models for code. *arXiv* **2023**, arXiv:2308.12950.
87. Lin, Z.; Deng, C.; Zhou, L.; Zhang, T.; Xu, Y.; Xu, Y.; He, Z.; Shi, Y.; Dai, B.; Song, Y. Geogalactica: A scientific large language model in geoscience. *arXiv* **2023**, arXiv:2401.00434.
88. Bi, Z.; Zhang, N.; Xue, Y.; Ou, Y.; Ji, D.; Zheng, G.; Chen, H. Oceangpt: A large language model for ocean science tasks. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, 11–16 August 2024; pp. 3357–3372.
89. Zhu, H.; Zhang, W.; Huang, N.; Li, B.; Niu, L.; Fan, Z.; Lun, T.; Tao, Y.; Su, J.; Gong, Z. PlanGPT: Enhancing urban planning with tailored language model and efficient retrieval. *arXiv* **2024**, arXiv:2402.19273.
90. Xu, F.; Zhang, J.; Gao, C.; Feng, J.; Li, Y. Urban generative intelligence (ugi): A foundational platform for agents in embodied city environment. *arXiv* **2023**, arXiv:2312.11813.
91. Li, Z.; Muhtar, D.; Gu, F.; He, Y.; Zhang, X.; Xiao, P.; He, G.; Zhu, X. Lhrs-bot-nova: Improved multimodal large language model for remote sensing vision-language interpretation. *ISPRS J. Photogramm. Remote Sens.* **2025**, *227*, 539–550. [CrossRef]
92. Hou, S.; Liu, Z.; Xie, L.; Chen, G.; Jiao, H.; Wu, S.; Guan, X.; Wu, H. PostGISer: The first end-to-end fine-tuned large language model for PostGIS GeoSQL query generation. *Inf. Process. Manag.* **2027**, *64*, 105042. [CrossRef]
93. Wei, C.; Zhang, Y.; Zhao, X.; Zeng, Z.; Wang, Z.; Lin, J.; Guan, Q.; Yu, W. GeoTool-GPT: A trainable method for facilitating Large Language Models to master GIS tools. *Int. J. Geogr. Inf. Sci.* **2025**, *39*, 707–731. [CrossRef]
94. Wang, C.; Ye, X.; Pan, X.; Pan, Z.; Wang, H.; Song, Y. Gre suite: Geo-localization inference via fine-tuned vision-language models and enhanced reasoning chains. *Adv. Neural Inf. Process. Syst.* **2026**, *38*, 92379–92409.
95. He, X.; You, Z.; Gong, J.; Liu, C.; Yue, X.; Zhuang, P.; Zhang, W.; Bai, L. Radarqa: Multi-modal quality analysis of weather radar forecasts. *Adv. Neural Inf. Process. Syst.* **2026**, *38*, 131737–131767.
96. Zhang, Y.; Li, J.; Zhang, W.; Wang, Z.; Zeng, Z.; Luo, W.; Guan, Q.; Lin, J.; Yu, W. GLoRA: A novel parameter-efficient fine-tuning framework for GIS large language models. *Int. J. Geogr. Inf. Sci.* **2026**, *40*, 2247–2275. [CrossRef]
97. Feng, J.; Liu, T.; Du, Y.; Guo, S.; Lin, Y.; Li, Y. Citygpt: Empowering urban spatial cognition of large language models. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining v.2, Toronto, ON, Canada, 3–7 August 2025; pp. 591–602.
98. Ji, P.; Xue, X.; Wang, S.; Yan, W. Entropy-Constrained Strategy Optimization in Urban Floods: A Multi-Agent Framework with LLM and Knowledge Graph Integration. *arXiv* **2025**, arXiv:2508.14654.
99. Lan, X.; Feng, J.; Sun, Y.; Gao, C.; Lei, J.; Shi, X.; Luo, H.; Li, Y. Open-set living need prediction with large language models. In Proceedings of the Findings of the Association for Computational Linguistics: ACL, 2025, Vienna, Austria, 27 July–1 August 2025; pp. 5454–5472.
100. Lai, S.; Xu, Z.; Zhang, W.; Liu, H.; Xiong, H. Llmlight: Large language models as traffic signal control agents. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1, Toronto, ON, Canada, 3–7 August 2025; pp. 2335–2346.
101. Lai, S.; Ning, Y.; Yuan, Z.; Chen, Z.; Liu, H. USTBench: Benchmarking and dissecting spatiotemporal reasoning capabilities of LLMs as urban agents. In Proceedings of the The Fourteenth International Conference on Learning Representations, Rio de Janeiro, Brazil, 23–27 April 2026.
102. Guo, D.; Yang, D.; Zhang, H.; Song, J.; Wang, P.; Zhu, Q.; Xu, R.; Zhang, R.; Ma, S.; Bi, X. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv* **2025**, arXiv:2501.12948.
103. Yang, Q.; Hong, Z.; Cao, D.; Wang, H.; Xie, Z.; He, T.; Liu, Y.; Yang, Y.; Zhang, D. Addrllm: Address rewriting via large language model on nationwide logistics data. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1, Toronto, ON, Canada, 3–7 August 2025; pp. 2756–2767.

104. Liu, J.; Sun, L.; Fu, R.; Yang, B. Towards faithful reasoning in remote sensing: A perceptually-grounded geospatial chain-of-thought for vision-language models. In Proceedings of the International Conference on Learning Representations, Rio de Janeiro, Brazil, 23–27 April 2026; pp. 6535–6563.
105. Li, Y.; Tian, M.; Zhu, D.; Zhu, J.; Lin, Z.; Xiong, Z.; Zhao, X. Drive-r1: Bridging reasoning and planning in vlms for autonomous driving with reinforcement learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Singapore, 20–27 January 2026; pp. 6708–6716.
106. Huang, C.; Chen, S.; Li, Z.; Qu, J.; Xiao, Y.; Liu, J.; Chen, Z. Geoagent: To empower llms using geospatial tools for address standardization. In Proceedings of the Findings of the Association for Computational Linguistics: ACL, 2024, Bangkok, Thailand, 11–16 August 2024; pp. 6048–6063.
107. Team, A.A.A.; Hu, Y.; Zhang, X.; Ouyang, S.; Yi, H.; Xu, L.; Lang, Q.; Tan, L.; Cheng, X.; Ye, T. AMAP Agentic Planning Technical Report. *arXiv* **2025**, arXiv:2512.24957.
108. Ning, Y.; Liu, R.; Wang, J.; Chen, K.; Li, W.; Fang, J.; Zheng, K.; Tan, N.; Liu, H. Deeptravel: An end-to-end agentic reinforcement learning framework for autonomous travel planning agents. *arXiv* **2025**, arXiv:2509.21842.
109. Kaelbling, L.P.; Littman, M.L.; Moore, A.W. Reinforcement learning: A survey. *J. Artif. Intell. Res.* **1996**, *4*, 237–285. [CrossRef]
110. Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C.D.; Ermon, S.; Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 53728–53741. [CrossRef]
111. Du, Y.; Feng, J.; Yuan, J.; Li, Y. Cams: A citygpt-powered agentic framework for urban human mobility simulation. *arXiv* **2025**, arXiv:2506.13599.
112. Mroueh, Y. Reinforcement Learning with Verifiable Rewards: GRPO's Effective Loss, Dynamics, and Success Amplification. *arXiv* **2025**, arXiv:2503.06639.
113. Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv* **2024**, arXiv:2402.03300.
114. Wang, Q.; Zou, X.; Jiang, Y.; Wen, H.; Wei, J.; Wen, Q.; Liang, Y. Urban-R1: Reinforced MLLMs Mitigate Geospatial Biases for Urban General Intelligence. *arXiv* **2025**, arXiv:2510.16555.
115. Xu, C.; Yu, F.; Bianco, M.J.; Kovarskiy, J.; Tang, R.; Zhang, Q.; Xu, Z.; LeVine, W.; Dubbs, B.; Liao, H. Geo-r1: Unlocking vlm geospatial reasoning with cross-view reinforcement learning. *arXiv* **2025**, arXiv:2510.00072.
116. Liu, T.; Pang, H.; Zhang, X.; Feng, J.; Li, Y.; Hui, P. Cityrise: Reasoning urban socio-economic status in vision-language models via reinforcement learning. *arXiv* **2025**, arXiv:2510.22282.
117. Zhang, R.; Zhang, Y.; Zhu, T.; Sun, L.; Lv, W. Eyes Will Shut: A Vision-Based Next GPS Location Prediction Model by Reinforcement Learning from Visual Map Feed Back. *arXiv* **2025**, arXiv:2507.18661.
118. Wu, B.; Fang, M.; Chen, L.; Xu, K.; Cheng, T.; Wang, J. Vision-language reasoning for geolocalization: A reinforcement learning approach. In Proceedings of the AAAI Conference on Artificial Intelligence, Singapore, 20–27 January 2026; pp. 33890–33898.
119. Wang, D.; Liu, S.; Jiang, W.; Wang, F.; Liu, Y.; Qin, X.; Luo, Z.; Zhou, C.; Guo, H.; Zhang, J. Geozero: Incentivizing reasoning from scratch on geospatial scenes. *arXiv* **2025**, arXiv:2511.22645.
120. Yu, B.; Yang, F.; Liu, Y.; Wang, C.; Luo, X.; Li, T.; Ke, R.; Zhou, X.; Liu, C. Locatability-Guided Adaptive Reasoning for Image Geo-Localization with Vision-Language Models. *arXiv* **2026**, arXiv:2603.13628.
121. Zhang, G.; Geng, H.; Yu, X.; Yin, Z.; Zhang, Z.; Tan, Z.; Zhou, H.; Li, Z.; Xue, X.; Li, Y. The landscape of agentic reinforcement learning for llms: A survey. *arXiv* **2025**, arXiv:2509.02547.
122. Zou, X.; Yang, Y.; Chen, Z.; Hao, X.; Chen, Y.; Huang, C.; Liang, Y. Traffic-r1: Reinforced llms bring human-like reasoning to traffic signal control systems. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), San Diego, CA, USA, 2–7 July 2026; pp. 21823–21838.
123. Zhang, Y.; Wang, Z.; He, Z.; Li, J.; Mai, G.; Lin, J.; Wei, C.; Yu, W. BB-GeoGPT: A framework for learning a large language model for geographic information science. *Inf. Process. Manag.* **2024**, *61*, 103808. [CrossRef]
124. Feng, J.; Wang, S.; Liu, T.; Xi, Y.; Li, Y. Urbanllava: A multi-modal large language model for urban intelligence. In Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–25 October 2025; pp. 6209–6219.
125. Chu, T.; Chen, Y.; Zhu, R.; Zeng, F. Estimating urban flooding depth by integrating multimodal image-text data: A segment-level direct preference optimization-based multimodal large language model. *ISPRS J. Photogramm. Remote Sens.* **2025**, *230*, 895–917. [CrossRef]
126. Marvin, G.; Hellen, N.; Jjingo, D.; Nakatumba-Nabende, J. Prompt engineering in large language models. In Proceedings of the International Conference on Data Intelligence and Cognitive Informatics, Tirunelveli, India, 27–28 June 2023; pp. 387–402.
127. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; Wang, H. Retrieval-augmented generation for large language models: A survey. *arXiv* **2023**, arXiv:2312.10997.
128. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837. [CrossRef]

129. Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Ma, J.; Li, R.; Xia, H.; Xu, J.; Wu, Z.; Chang, B. A survey on in-context learning. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA, 12–16 November 2024; pp. 1107–1128.
130. Hou, S.; Jiao, H.; Shen, Z.; Liang, J.; Zhao, A.; Zhang, X.; Wang, J.; Wu, H. Chain-of-programming (CoP): Empowering large language models for geospatial code generation task. *Int. J. Digit. Earth* **2025**, *18*, 2509812. [CrossRef]
131. Wang, S.; Li, R.; Wu, H.; Li, J.; Shen, Y. Fine-grained flood disaster information extraction incorporating multiple semantic features. *Int. J. Digit. Earth* **2025**, *18*, 2448221. [CrossRef]
132. Tang, Z.; Yue, S.; Mu, F.; Shu, Y.; Sun, Z.; Wen, Y. MapMate: A framework bridging natural language interaction and map design through large language models. *Int. J. Appl. Earth Obs. Geoinf.* **2026**, *146*, 105073. [CrossRef]
133. Hou, S.; Zhao, A.; Liang, J.; Shen, Z.; Wu, H. Geo-FuB: A method for constructing an Operator-Function knowledge base for geospatial code generation with large language models. *Knowl.-Based Syst.* **2025**, *319*, 113624. [CrossRef]
134. Hou, S.; Liang, J.; Zhao, A.; Wu, H. GEE-OPs: An operator knowledge base for geospatial code generation on the Google Earth Engine platform powered by large language models. *Geo-Spat. Inf. Sci.* **2026**, *29*, 429–450. [CrossRef]
135. Khosravi Kazazi, A.; Li, Z.; Lessani, M.N.; Cervone, G. From questions to queries: An ai-powered multi-agent framework for spatial text-to-sql. *Int. J. Digit. Earth* **2026**, *19*, 2687193. [CrossRef]
136. Peng, Z.; Kuai, X.; Ke, S.; Dong, X.; Guo, R. Enhancing geodatabases operability: Advanced human-computer interaction through RAG and Multi-Agent Systems. *Big Earth Data* **2025**, *9*, 217–242. [CrossRef]
137. Chen, Z.; Wang, X.; Zhang, X.; Lin, M.; Liao, Y.; Li, J.; Bai, Y. GeoFactory: An LLM performance enhancement framework for geoscience factual and inferential tasks. *Big Earth Data* **2025**, *9*, 714–746. [CrossRef]
138. Liang, J.; Hou, S.; Jiao, H.; Qing, Y.; Zhao, A.; Shen, Z.; Xiang, L.; Wu, H. GeoGraphRAG: A graph-based retrieval-augmented generation approach for empowering large language models in automated geospatial modeling. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *142*, 104712. [CrossRef]
139. Liang, J.; Hou, S.; Zhao, A.; Xu, Q.; Xiang, L.; Li, R.; Wu, H. Design and application of a semantic-driven geospatial modeling knowledge graph based on large language models. *Geo-Spat. Inf. Sci.* **2025**, *28*, 2927–2946. [CrossRef]
140. Jiang, Z.; Wang, X. SpatialGPT: Zero-Shot Vision-and-Language Navigation via Spatial CoT over Structured Spatial Memory. In Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems, Minneapolis, MN, USA, 3–6 November 2025; pp. 423–435.
141. Zhang, J.; Zhu, J.; Guo, Z.; Wu, J.; Guo, Y.; Lai, J.; Li, W. More intelligent knowledge graph: A large language model-driven method for knowledge representation in geospatial digital twins. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *139*, 104527. [CrossRef]
142. Yu, D.; Bao, R.; Ning, R.; Peng, J.; Mai, G.; Zhao, L. Spatial-rag: Spatial retrieval augmented generation for real-world geospatial reasoning questions. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2026, San Diego, CA, USA, 2–7 July 2026; pp. 11094–11112.
143. Yu, S.; Liu, J.; Gao, C.; Niu, W.; Guo, X.; Liu, H.; Xu, M. LLMTrajQuery: An LLM-based generative approach to semantic trajectory queries. *Int. J. Geogr. Inf. Sci.* **2026**, *40*, 1651–1679. [CrossRef]
144. Wu, C.-C.; Huang, S.-B.; Song, Y.-L.; Lin, P.-H.; Lin, M.C.-L.; Lin, Y.-T. Democratizing Multi-Granularity Spatio-Temporal Intelligence with Multi-Agent Systems. In Proceedings of the 1st ACM SIGSPATIAL International Workshop on Generative and Agentic AI for Multi-Modality Space-Time Intelligence, Minneapolis, MN, USA, 3 November 2025; pp. 22–26.
145. Zhu, H.; Chen, G.; Zhang, W. PlanGPT: Enhancing urban planning with a tailored agent framework. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), Vienna, Austria, 28–30 July 2025; pp. 764–783.
146. Niu, L.; Deng, Z.; Li, B.; Huang, N.; Liu, R.; Zhang, W. AskNearby: An LLM-Based Application for Neighborhood Information Retrieval and Personalized Cognitive-Map Recommendations. In Proceedings of the 8th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, Minneapolis, MN, USA, 3–6 November 2025; pp. 22–32.
147. Rahman, M. Physically interpretable AlphaEarth foundation model embeddings enable LLM-based land surface intelligence. *Remote Sens. Appl. Soc. Environ.* **2026**, *42*, 102045. [CrossRef]
148. Feng, P.; Lv, Z.; Ye, J.; Wang, X.; Huo, X.; Yu, J.; Xu, W.; Zhang, W.; Bai, L.; He, C. Earth-agent: Unlocking the full landscape of earth observation with agents. In Proceedings of the International Conference on Learning Representations, Rio de Janeiro, Brazil, 23–27 April 2026; pp. 55700–55764.
149. Huang, Y.; Gao, T.; Xu, H.; Zhao, Q.; Song, Y.; Gui, Z.; Lv, T.; Chen, H.; Cui, L.; Li, S. Peace: Empowering geologic map holistic understanding with mllms. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 3899–3908.
150. Zhang, Y.; He, Z.; Li, J.; Lin, J.; Guan, Q.; Yu, W. MapGPT: An autonomous framework for mapping by integrating large language model and cartographic tools. *Cartogr. Geogr. Inf. Sci.* **2024**, *51*, 717–743. [CrossRef]
151. Zhang, S.; Fu, D.; Liang, W.; Zhang, Z.; Yu, B.; Cai, P.; Yao, B. Trafficgpt: Viewing, processing and interacting with traffic foundation models. *Transp. Policy* **2024**, *150*, 95–105. [CrossRef]

152. Chen, Y.; Wang, W.; Lobry, S.; Kurtz, C. An llm agent for automatic geospatial data analysis. *arXiv* **2024**, arXiv:2410.18792.
153. Wang, S.; Liang, C.; Gao, Y.; Liu, Y.; Li, J.; Wang, H. Decoding urban industrial complexity: Enhancing knowledge-driven insights via industryscopegpt. In Proceedings of the 32nd ACM International Conference on Multimedia, Victoria, Australia, 28 October–1 November 2024; pp. 4757–4765.
154. Luo, Q.; Lin, Q.; Xu, L.; Wu, S.; Mao, R.; Wang, C.; Feng, H.; Huang, B.; Du, Z. Geojson agents: A multi-agent llm architecture for geospatial analysis—Function calling vs. code generation. *Big Earth Data* **2026**, *10*, 1731–1784. [CrossRef]
155. Wu, H.; Jiao, H.; Hou, S.; Liang, J.; Shen, Z.; Zhao, A.; Qing, Y.; Jin, F.; Guan, X.; Gui, Z. GeoColab: An LLM-based multi-agent collaborative framework for geospatial code generation. *Int. J. Digit. Earth* **2025**, *18*, 2569405. [CrossRef]
156. Wang, C.; Kang, Y.; Gong, Z.; Zhao, P.; Feng, Y.; Zhang, W.; Li, G. CartoAgent: A multimodal large language model-powered multi-agent cartographic framework for map style transfer and evaluation. *Int. J. Geogr. Inf. Sci.* **2025**, *39*, 1904–1937. [CrossRef]
157. Han, X.; Zhu, C.; Zhu, H.; Zhao, X. Swarm intelligence in geo-localization: A multi-agent large vision-language model collaborative framework. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, Toronto, ON, Canada, 3–7 August 2025; pp. 814–825.
158. Luo, P.; Lou, X.; Zheng, Y.; Zheng, Z.; Ermon, S. Geoevolve: Automating geospatial model discovery via multi-agent large language models. *arXiv* **2025**, arXiv:2509.21593.
159. Bougie, N.; Watanabe, N. Citysim: Modeling urban behaviors and city dynamics with large-scale llm-driven agent simulation. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, Suzhou, China, 4–9 November 2025; pp. 215–229.
160. Zhou, Z.; Lin, Y.; Jin, D.; Li, Y. Large language model for participatory urban planning. *arXiv* **2024**, arXiv:2402.17161.
161. Ni, H.; Wang, Y.; Liu, H. Planning, living and judging: A multi-agent llm-based framework for cyclical urban planning. *arXiv* **2024**, arXiv:2412.20505.
162. Xuan, W.; Wang, J.; Qi, H.; Chen, Z.; Zheng, Z.; Zhong, Y.; Xia, J.; Yokoya, N. DynamicVL: Benchmarking multimodal large language models for dynamic city understanding. In Proceedings of the Advances in Neural Information Processing Systems, San Diego, California, USA, 2–7 December 2025 and Mexico City, Mexico, 30 November–5 December 2025; Volume 38.
163. Feng, Y.; Ding, L.; Xiao, G. Geoqamap-geographic question answering with maps leveraging LLM and open knowledge base (short paper). In Proceedings of the 12th International Conference on Geographic Information Science (GIScience 2023), Leeds, UK, 12–15 September 2023.
164. Li, Z.; Xia, L.; Tang, J.; Xu, Y.; Shi, L.; Xia, L.; Yin, D.; Huang, C. Urbangpt: Spatio-temporal large language models. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Barcelona, Spain, 25–29 August 2024; pp. 5351–5362.
165. Rashik, M.; Fekete, J.-D.; Mahyar, N. CLAImate: AI-enabled climate change communication through personalized and localized narrative visualizations. In Proceedings of the 2025 IEEE Visualization and Visual Analytics (VIS), Vienna, Austria, 1–7 November 2025; pp. 36–40.
166. Aghzal, M.; Plaku, E.; Yao, Z. Can large language models be good path planners? a benchmark and investigation on spatial-temporal reasoning. *arXiv* **2023**, arXiv:2310.03249.
167. Gardelakos, E.-O.; Kyriakopoulos, V.; Pantazi, D.-A.; Kapopoulos, O.-M.; Tsourma, M.; Koubarakis, M. Can Large Reasoning Models Reason about Spatial Relations? In Proceedings of the 8th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, Minneapolis, MN, USA, 3 November 2025; pp. 81–91.
168. Jin, Y.; Ma, J. Large language model as parking planning agent in the context of mixed period of autonomous vehicles and human-driven vehicles. *Sustain. Cities Soc.* **2024**, *117*, 105940. [CrossRef]
169. Singh, S.; Fore, M.; Stamoulis, D. Geollm-engine: A realistic environment for building geospatial copilots. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 17–18 June 2024; pp. 585–594.
170. Yuan, Z.; Xiong, Z.; Mou, L.; Zhu, X.X. Chatearthnet: A global-scale image-text dataset empowering vision-language geo-foundation models. *Earth Syst. Sci. Data.* **2025**, *17*, 1245–1263. [CrossRef]
171. Hou, S.; Jiao, H.; Liang, J.; Shen, Z.; Zhao, A.; Wu, H. GeoCogent: An LLM-based agent for geospatial code generation. *Int. J. Geogr. Inf. Sci.* **2026**, *40*, 1073–1106. [CrossRef]
172. Mooney, P.; Cui, W.; Guan, B.; Juhász, L. 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, Hamburg, Germany, 13 November 2023; pp. 85–94.
173. Paz-Argaman, T.; Bauman, T.; Mondshine, I.; Omer, I.; Dalyot, S.; Tsarfaty, R. HeGeL: A novel dataset for geo-location from Hebrew text. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, Toronto, ON, Canada, 9–14 July 2023; pp. 7311–7321.

174. Paz-Argaman, T.; Palowitch, J.; Kulkarni, S.; Baldridge, J.; Tsarfaty, R. Where do we go from here? multi-scale allocentric relational inferencefrom natural spatial descriptions. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), St. Julian's, Malta, 17–22 March 2024; pp. 1026–1040.
175. Lin, H.; Hong, D.; Ge, S.; Luo, C.; Jiang, K.; Jin, H.; Wen, C. RS-MoE: A vision–language model with mixture of experts for remote sensing image captioning and visual question answering. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5614918. [CrossRef]
176. Jiang, Y.; Chao, Q.; Chen, Y.; Li, X.; Liu, S.; Cong, G. Urbanllm: Autonomous urban activity planning and management with large language models. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP, 2024, Miami, FL, USA, 12–16 November 2024; pp. 1810–1825.
177. Unlu, E. Chatmap: Large language model interaction with cartographic data. *arXiv* **2023**, arXiv:2310.01429.
178. Chen, Z.; Wang, X.; Liao, Y.; Lin, M.; Bai, Y. Climatechat: Designing data and methods for instruction tuning llms to answer climate change queries. *arXiv* **2025**, arXiv:2506.13796.
179. Goecks, V.G.; Waytowich, N.R. Disasterresponsegpt: Large language models for accelerated plan of action development in disaster response scenarios. *arXiv* **2023**, arXiv:2306.17271.
180. Xia, Y.; Qu, A.; Zheng, Y.; Tang, Y.; Zhuang, D.; Liang, Y.; Wang, S.; Wu, C.; Sun, L.; Zimmermann, R. Reimagining urban science: Scaling causal inference with large language models. *arXiv* **2025**, arXiv:2504.12345.
181. Gupta, D.V.; Ishaqui, A.S.A.; Kadiyala, D.K. Geode: A zero-shot geospatial question-answering agent with explicit reasoning and precise spatio-temporal retrieval. *arXiv* **2024**, arXiv:2407.11014.
182. Hou, S.; Shen, Z.; Liang, J.; Jiao, H.; Zhao, A.; Qing, Y.; Peng, D.; Gui, Z.; Guan, X.; Xiang, L. Can large language models generate geospatial code? *Geo-Spat. Inf. Sci.* **2026**, *29*, 753–787. [CrossRef]
183. Zhang, Z.; Laparra, E.; Bethard, S. Improving toponym resolution by predicting attributes to constrain geographical ontology entries. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), Mexico City, Mexico, 16–21 June 2024; pp. 35–44.
184. Mansourian, A.; Oucheikh, R. ChatGeoAI: Enabling geospatial analysis for public through natural language, with large language models. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 348. [CrossRef]
185. Li, X.; Li, J.; Zhang, K.; Fang, Y.; Lin, L.; Wang, H.; Wu, H.; Fan, Z. Decoding the Delta: Unifying Remote Sensing Change Detection and Understanding with Multimodal Large Language Models. *arXiv* **2026**, arXiv:2604.14044.
186. Deng, P.; Zhou, W.; Wu, H. Deltavlm: Interactive remote sensing image change analysis via instruction-guided difference perception. *Remote Sens.* **2026**, *18*, 541. [CrossRef]
187. Kuckreja, K.; Danish, M.S.; Naseer, M.; Das, A.; Khan, S.; Khan, F.S. Geochat: Grounded large vision-language model for remote sensing. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 27831–27840.
188. Ni, J.; Wang, S.; He, Q.; Jin, M.; Jin, W. Streasoner: Empowering llms for spatio-temporal reasoning in time series via spatial-aware reinforcement learning. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), San Diego, CA, USA, 2–7 July 2026; pp. 15369–15416.
189. Li, L.; Zhou, Y.; Liang, Y.; Tsung, F.; Wei, J. Recognition through reasoning: Reinforcing image geo-localization with large vision-language models. *Adv. Neural Inf. Process. Syst.* **2026**, *38*, 62132–62159.
190. Jia, F.; Dai, L.; Deng, W.; Zhang, F.; Hu, C.; Jiang, D.; Liu, Y. Spotagent: Grounding visual geo-localization in large vision-language models through agentic reasoning. In Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, Jeju Island, South Korea, 9–13 August 2026; pp. 2062–2073.
191. Zhang, Z.; Guan, Z.; Zhao, T.; Shen, H.; Cai, Y.; Su, Z.; Shang, Y.; Liu, Z.; Yin, J.; Li, X. Geo-R1: Improving few-shot geospatial referring expression understanding with reinforcement fine-tuning. *ISPRS J. Photogramm. Remote Sens.* **2026**, *237*, 113–129. [CrossRef]
192. Truong, T.H.; Lau, J.H.; Qi, J. GPSBench: Do Large Language Models Understand GPS Coordinates? *arXiv* **2026**, arXiv:2602.16105.
193. Feng, S.; Lyu, H.; Li, F.; Sun, Z.; Chen, C. Where to move next: Zero-shot generalization of llms for next poi recommendation. In Proceedings of the 2024 IEEE Conference on Artificial Intelligence (cai), Singapore, 25–27 June 2024; pp. 1530–1535.
194. Wang, J.; Xuan, W.; Qi, H.; Dai, P.; Liu, K.; Chen, H.; Zheng, Z.; Xia, J.; Ermon, S.; Yokoya, N. Can LLM Agents Respond to Disasters? Benchmarking Heterogeneous Geospatial Reasoning in Emergency Operations. *arXiv* **2026**, arXiv:2605.11633.
195. Lin, Q.; Hu, R.; Li, H.; Wu, S.; Li, Y.; Fang, K.; Feng, H.; Du, Z.; Xu, L. ShapefileGPT: A multi-agent large language model framework for automated shapefile processing. *Int. J. Digit. Earth* **2025**, *18*, 2577884. [CrossRef]
196. Huang, W.; Wang, J.; Cong, G. Zero-shot urban function inference with street view images through prompting a pretrained vision-language model. *Int. J. Geogr. Inf. Sci.* **2024**, *38*, 1414–1442. [CrossRef]
197. Huang, Z.; Peng, P.; Lu, F.; Zhang, H. An LLM-Based Method for Quality Information Extraction From Web Text for Crowed-Sensing Spatiotemporal Data. *Trans. GIS* **2025**, *29*, e13294. [CrossRef]
198. Wang, S.; Feng, J.; Liu, T.; Pei, D.; Li, Y. Mitigating Geospatial Knowledge Hallucination in Large Language Models: Benchmarking and Dynamic Factuality Aligning. In Proceedings of the EMNLP (Findings), Suzhou, China, 4–9 November 2025; pp. 870–888.

199. Guo, H.; Su, X.; Wu, C.; Du, B.; Zhang, L.; Li, D. Remote sensing chatgpt: Solving remote sensing tasks with chatgpt and visual models. In Proceedings of the IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium, Athens, Greece, 7–12 July 2024; pp. 11474–11478.
200. Wang, H.; Han, J.; Fan, W.; Liu, H. ClimAgent: LLM as Agents for Autonomous Open-ended Climate Science Analysis. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2026, San Diego, CA, USA, 2–7 July 2026; pp. 21219–21241.
201. Pai, P.-L.; Lu, H.-H.; Su, W.-R.; Liao, H.-M.; Chan, T.-C.; Lin, Y.; Chiang, Y.-Y. AI-Supported Smart Search for Multi-Period Historical Maps Using the mapKurator Place Name Index. *J. Map Geogr. Libr.* **2026**, *22*, 23–45. [CrossRef]
202. Vickneswaran, K.; Andarcia, M.G.; Retief, H.; Dickens, C.; Silva, P. WaterCopilot: An AI-driven virtual assistant for water management. *arXiv* **2026**, arXiv:2601.08559.
203. Zeng, Q.; Zhao, R.; Mao, J.; Li, H.; Xu, F.; Li, Y. CrimeMind: Simulating urban crime with multi-modal LLM agents. *arXiv* **2025**, arXiv:2506.05981.
204. Chu, M.; Chen, Y.; Gui, H.; Yu, S.; Wang, Y.; Jia, J. TraveLLaMA: Facilitating Multi-modal Large Language Models to Understand Urban Scenes and Provide Travel Assistance. *arXiv* **2025**, arXiv:2504.16505.
205. Kefalidis, S.-A.; Punjani, D.; Tsalapati, E.; Plas, K.; Pollali, M.-A.; Maret, P.; Koubarakis, M. The question answering system GeoQA2 and a new benchmark for its evaluation. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *134*, 104203. [CrossRef]
206. Yao, L.; Ren, F.; Du, K.; Du, Q. From knowledge graph construction to retrieval-augmented generation: A framework for comprehensive earthquake emergency support. *Geo-Spat. Inf. Sci.* **2026**, *29*, 509–529. [CrossRef]
207. Zhu, J.; Dang, P.; Cao, Y.; Lai, J.; Guo, Y.; Wang, P.; Li, W. A flood knowledge-constrained large language model interactable with GIS: Enhancing public risk perception of floods. *Int. J. Geogr. Inf. Sci.* **2024**, *38*, 603–625. [CrossRef]
208. Zhao, A.; Gui, Z.; Liang, J.; Liu, Y.; Peng, D.; Wei, W.; Hou, S.; Wu, H. Extraction of geoprocessing modeling knowledge from crowdsourced Google Earth Engine scripts by coordinating large and small language models. *Int. J. Geogr. Inf. Sci.* **2026**, *40*, 808–841. [CrossRef]
209. Liang, Y.; Liu, Y.; Wang, X.; Zhao, Z. Exploring large language models for human mobility prediction under public events. *Comput. Environ. Urban Syst.* **2024**, *112*, 102153. [CrossRef]
210. Ning, Y.; Liu, H. Urbankgent: A unified large language model agent framework for urban knowledge graph construction. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 123127–123154. [CrossRef]
211. Cheng, F.; Wang, J.; Wang, S.; Huang, Z.; Li, X. Geoguess: Multimodal reasoning based on hierarchy of visual information in street view. *arXiv* **2025**, arXiv:2506.16633.
212. Wang, J.; Xuan, W.; Qi, H.; Chen, Z.; Chen, H.; Zheng, Z.; Xia, J.; Zhong, Y.; Yokoya, N. CityVLM: Towards sustainable urban development via multi-view coordinated vision–language model. *ISPRS J. Photogramm. Remote Sens.* **2026**, *232*, 62–74. [CrossRef]
213. Zhang, X.; Wu, Y.; Zhang, B.; Lin, C.; Shen, C.; Backes, M.; Zhang, Y. GEO-Detective: Unveiling Location Privacy Risks in Images with LLM Agents. *arXiv* **2025**, arXiv:2511.22441.
214. Zhang, X.; Ouyang, T.; Shang, Y.; Liao, Q.; Li, Y. UrbanMLLM: Joint learning of cross-view imagery for urban understanding. In Proceedings of the 43rd International Conference on Machine Learning (ICML 2026), Seoul, Republic of Korea, 6–11 July 2026. [PubMed]
215. Cheng, A.-C.; Yin, H.; Fu, Y.; Guo, Q.; Yang, R.; Kautz, J.; Wang, X.; Liu, S. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 135062–135093. [CrossRef]
216. Wang, J.; Jiang, R.; Yang, C.; Wu, Z.; Onizuka, M.; Shibasaki, R.; Koshizuka, N.; Xiao, C. Large language models as urban residents: An llm agent framework for personal mobility generation. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 124547–124574. [CrossRef]
217. Zhang, K.; Pang, Y.; Zhang, Y.; Sekimoto, Y. Mobglm: A large language model for synthetic human mobility generation. In Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems, Atlanta, GA, USA, 29 October–1 November 2024; pp. 629–632.
218. Liu, L.; Yu, S.; Wang, R.; Ma, Z.; Shen, Y. How can large language models understand spatial-temporal data? *arXiv* **2024**, arXiv:2401.14192.
219. Du, Y.; Feng, J.; Zhao, J.; Li, Y. Trajagent: An llm-agent framework for trajectory modeling via large-and-small model collaboration. *Adv. Neural Inf. Process. Syst.* **2026**, *38*, 21595–21625.
220. Yuan, L.; Mo, F.; Huang, K.; Wang, W.; Zhai, W.; Zhu, X.; Li, Y.; Xu, J.; Nie, J.-Y. Omnigeo: Towards a multimodal large language models for geospatial artificial intelligence. *arXiv* **2025**, arXiv:2503.16326.
221. Manvi, R.; Khanna, S.; Mai, G.; Burke, M.; Lobell, D.; Ermon, S. Geollm: Extracting geospatial knowledge from large language models. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024; pp. 38791–38807.
222. Zhao, J.; Feng, J.; Li, Y. AgentTravel: Knowledge-Augmented LLM Agent Framework for Urban Travel Planning. In Proceedings of the NORA@ NeurIPS, Mexico City, Mexico, 1 December 2025; pp. 55–67.

223. Guo, X.; Peng, M.; Hao, X.; Zou, X.; Wang, Q.; Ruan, S.; Liang, Y. AgentSense: LLMs empower generalizable and explainable web-based participatory urban sensing. In Proceedings of the ACM Web Conference 2026, Dubai, United Arab Emirates, 29 June–3 July 2026; pp. 5439–5450.
224. Wu, H.; Shen, Z.; Hou, S.; Liang, J.; Jiao, H.; Qing, Y.; Zhang, X.; Li, X.; Gui, Z.; Guan, X. AutoGEEval: A multimodal and automated evaluation framework for geospatial code generation on GEE with large language models. *ISPRS Int. J. Geo-Inf.* **2025**, *14*, 256. [CrossRef]
225. Wu, H.; Shen, Z.; Hou, S.; Jiao, H.; Liu, Z.; Xie, L.; Liang, J.; Qing, Y.; Zhang, X.; Gui, Z. AutoGEEval++: A multi-level and multi-geospatial-modality automated evaluation framework for large language models in geospatial code generation on Google Earth Engine. *Big Earth Data* **2025**, *9*, 747–796. [CrossRef]
226. Merten, G.; Dejaegere, G.; Sakr, M. GeoPandas-AI: A Smart Class Bringing LLM as Stateful AI Code Assistant. In Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems, Minneapolis, MN, USA, 3–6 November 2025; pp. 538–548.
227. Chen, G.; Jiao, H.; Hou, S.; Liu, Z.; Xie, L.; Wu, S.; Wu, H.; Guan, X.; Gui, Z. GeoJSEval: An automated evaluation framework for large language models on JavaScript-based geospatial computation and visualization code generation. *ISPRS Int. J. Geo-Inf.* **2025**, *14*, 382. [CrossRef]
228. Yu, C.; Yao, Y.; Zhang, X.; Zhu, G.; Guo, Y.; Shao, X.; Shibasaki, M.; Hu, Z.; Dai, L.; Guan, Q. Monkuu: A LLM-powered natural language interface for geospatial databases with dynamic schema mapping. *Int. J. Geogr. Inf. Sci.* **2026**, *40*, 588–609. [CrossRef]
229. Hou, S.; Jiao, H.; Liu, Z.; Xie, L.; Chen, G.; Wu, S.; Guan, X.; Wu, H. GeoSQL-Eval: First evaluation of LLMs on PostGIS-based NL2GeoSQL queries. *Expert Syst. With Appl.* **2026**, 132122.
230. Feng, Y.; Zhang, P.; Xiao, G.; Ding, L.; Meng, L. Towards a barrier-free GeoQA portal: Natural language interaction with geospatial data using multi-agent LLMs and semantic search. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *144*, 104825. [CrossRef]
231. Wan, Z.; Hu, W.; Zhang, C.J.; Song, Y.; Li, S.; Xiao, R.; Wei, X.-Y.; Wong, R.C.-W. Osmt: Bridging openstreetmap queries and natural language with open-source tag-aware language models. In Proceedings of the 2026 IEEE 42nd International Conference on Data Engineering (ICDE), Montreal, QC, Canada, 4–8 May 2026; pp. 517–530.
232. Lee, C.; Paramanayakam, V.; Karatzas, A.; Jian, Y.; Fore, M.; Liao, H.; Yu, F.; Li, R.; Anagnostopoulos, I.; Stamoulis, D. Multi-agent geospatial copilots for remote sensing workflows. In Proceedings of the IGARSS 2025-2025 IEEE International Geoscience and Remote Sensing Symposium, Brisbane, Australia, 3–8 August 2025; pp. 1084–1089.
233. Yang, Q.; Qiu, S.; Li, B.; Shan, X.; Feng, J.; Zhou, S.; Zhou, X.; Xing, T.; Guo, J.; Dong, X. HydroAgent: Formalizing Forecaster Expertise into Skill-Orchestrated Flood Forecasting Workflows. *arXiv* **2026**, arXiv:2607.23983.
234. Zhao, S.; Liu, F.; Zhang, X.; Chen, H.; Gu, X.; Jiang, Z.; Ling, F.; Fei, B.; Zhang, W.; Wang, J. Openearth-agent: From tool calling to tool creation for open-environment earth observation. *arXiv* **2026**, arXiv:2603.22148.
235. Feng, X.; Cao, Y. Enhancing the accessibility of regionalization techniques through large language models: A case study in conversational agent guidance. *Int. J. Geogr. Inf. Sci.* **2025**, *39*, 2045–2068. [CrossRef]
236. Gao, K.; Lu, D.; Li, L.; Chen, N.; He, H.; Du, J.; Xu, L.; Li, J. Instructor–Worker large language model system for policy recommendation: A case study on air quality analysis of the January 2025 Los Angeles wildfires. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *143*, 104774. [CrossRef]
237. Xu, H.; Yu, X.-Y. From PowerPoint UI sketches to web-based applications: Pattern-driven code generation for GIS dashboard development using knowledge-augmented LLMs, context-aware visual prompting, and the React framework. *arXiv* **2025**, arXiv:2502.08756.
238. Manivannan, V.V.; Jafari, Y.; Eranky, S.; Ho, S.; Yu, R.; Watson-Parris, D.; Ma, Y.; Bergen, L.; Berg-Kirkpatrick, T. Climaqa: An automated evaluation framework for climate question answering models. In Proceedings of the International Conference on Learning Representations, Singapore, 24–28 April 2025; pp. 25357–25382.
239. Li, X.; Ding, J.; Elhoseiny, M. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 3229–3242. [CrossRef]
240. Nunes, R.O.; Spritzer, A.S.; Balreira, D.G.; Freitas, C.M.; Carbonera, J.L. An evaluation of large language models for geological named entity recognition. In Proceedings of the 2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI), Herndon, VA, USA, 28–30 October 2024; pp. 494–501.
241. Srivastava, V.; Lei, F.; Mukhopadhyay, S.; Gupta, V.; Maciejewski, R. Mapiq: Evaluating multimodal large language models for map question answering. *arXiv* **2025**, arXiv:2507.11625.
242. Chen, J.; Zhou, P.; Hua, Y.; Chong, D.; Cao, M.; Li, Y.; Chen, W.; Zhu, B.; Liang, J.; Yuan, Z. ClimateIQA: A new dataset and benchmark to advance vision-language models in meteorology anomalies analysis. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, Toronto, ON, Canada, 3–7 August 2025; pp. 5322–5333.
243. Danish, M.; Munir, M.A.; Shah, S.R.A.; Kuckreja, K.; Khan, F.S.; Fraccaro, P.; Lacoste, A.; Khan, S. Geobench-vlm: Benchmarking vision-language models for geospatial tasks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–25 October 2025; pp. 7132–7142.

244. Dihan, M.L.; Hassan, M.T.; Parvez, M.T.; Hasan, M.H.; Alam, M.A.; Cheema, M.A.; Ali, M.E.; Parvez, M.R. Mapeval: A map-based evaluation of geo-spatial reasoning in foundation models. *arXiv* **2024**, arXiv:2501.00316.
245. Jia, P.; Zhang, Y.; Zhao, X.; Li, Y. Geoarena: An open platform for benchmarking large vision-language models on worldwide image geolocalization. *arXiv* **2025**, arXiv:2509.04334v2.
246. O'Sullivan, K.; Schneider, N.R.; Samet, H. Metric reasoning in large language models. In Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems, Atlanta, GA, USA, 29 October–1 November 2024; pp. 501–504.
247. Zhang, Q.; Gao, S.; Wei, C.; Zhao, Y.; Nie, Y.; Chen, Z.; Chen, S.; Su, Y.; Sun, H. GeoAnalystBench: A GeoAI benchmark for assessing large language models for spatial analysis workflow and code generation. *Trans. GIS* **2025**, *29*, e70135. [CrossRef]
248. Gramacki, P.; Martins, B.; Szymański, P. Evaluation of code LLMs on geospatial code generation. In Proceedings of the 7th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, Atlanta, GA, USA, 29 October 2024; pp. 54–62.
249. Xu, L.; Zhao, S.; Lin, Q.; Chen, L.; Luo, Q.; Wu, S.; Ye, X.; Feng, H.; Du, Z. Evaluating large language models on geospatial tasks: A multiple geospatial task benchmarking study. *Int. J. Digit. Earth* **2025**, *18*, 2480268. [CrossRef]
250. Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; Steinhardt, J. Measuring massive multitask language understanding. *arXiv* **2020**, arXiv:2009.03300.
251. Huang, Y.; Bai, Y.; Zhu, Z.; Zhang, J.; Zhang, J.; Su, T.; Liu, J.; Lv, C.; Zhang, Y.; Fu, Y. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 62991–63010. [CrossRef]
252. Zhang, X.; Li, C.; Zong, Y.; Ying, Z.; He, L.; Qiu, X. Evaluating the performance of large language models on gaokao benchmark. *arXiv* **2023**, arXiv:2305.12474.
253. JuhÃsz, L.; Mooney, P.; Hochmair, H.H.; Guan, B. ChatGPT as a mapping assistant: A novel method to enrich maps with generative AI and content derived from street-level photographs. *arXiv* **2023**, arXiv:2306.03204.
254. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
255. Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 25–26 July 2004; pp. 74–81.
256. Xia, Y.; Huang, Y.; Qiu, Q.; Zhang, X.; Miao, L.; Chen, Y. A question and answering service of typhoon disasters based on the T5 large language model. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 165. [CrossRef]
257. Xu, W.; Zhao, X.; Zhou, Y.; Yue, X.; Fei, B.; Ling, F.; Zhang, W.; Bai, L. Earthse: A benchmark evaluating earth scientific exploration capability for large language models. In Proceedings of the International Conference on Learning Representations, Rio de Janeiro, Brazil, 23–27 April 2026; pp. 130942–130966.
258. Church, K.W. Word2Vec. *Nat. Lang. Eng.* **2017**, *23*, 155–162. [CrossRef]
259. Hu, X.; Kersten, J.; Klan, F.; Farzana, S.M. Toponym resolution leveraging lightweight and open-source large language models and geo-knowledge. *Int. J. Geogr. Inf. Sci.* **2026**, *40*, 670–697. [CrossRef]
260. Ji, Y.; Gao, S.; Nie, Y.; Majić, I.; Janowicz, K. Foundation models for geospatial reasoning: Assessing the capabilities of large language models in understanding geometries and topological spatial relations. *Int. J. Geogr. Inf. Sci.* **2025**, *39*, 1866–1903. [CrossRef]
261. Singh, S.; Fore, M.; Stamoulis, D. Evaluating tool-augmented agents in remote sensing platforms. *arXiv* **2024**, arXiv:2405.00709.
262. Li, Z.; Xu, R.; Hu, J.; Peng, Z.; Lu, X.; Guo, C.; Yang, B. Ocean significant wave height estimation with spatio-temporally aware large language models. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, Boise, ID, USA, 21–25 October 2024; pp. 3892–3896.
263. Beneduce, C.; Lepri, B.; Luca, M. Large language models are zero-shot next location predictors. *IEEE Access* **2025**, *13*, 77456–77467. [CrossRef]
264. Shao, C.; Xu, F.; Fan, B.; Ding, J.; Yuan, Y.; Wang, M.; Li, Y. Chain-of-planned-behaviour workflow elicits few-shot mobility generation in llms. *arXiv* **2024**, arXiv:2402.09836.
265. Balsebre, P.; Huang, W.; Cong, G. LAMP: A language model on the map. *arXiv* **2024**, arXiv:2403.09059.
266. Feng, J.; Zhang, J.; Liu, T.; Zhang, X.; Ouyang, T.; Yan, J.; Du, Y.; Guo, S.; Li, Y. Citybench: Evaluating the capabilities of large language models for urban tasks. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, Toronto, ON, Canada, 3–7 August 2025; pp. 5413–5424.
267. Wasi, A.T.; Faisal, W.; Rahman, A.; Anik, M.A.; Shahriar, M.; Topu, M.M.; Meem, S.T.; Priti, R.N.; Mitu, S.A.; Hoque, M.I. SpatiaLab: Can Vision-Language Models Perform Spatial Reasoning in the Wild? *arXiv* **2026**, arXiv:2602.03916.
268. Ren, S.; Guo, D.; Lu, S.; Zhou, L.; Liu, S.; Tang, D.; Sundaresan, N.; Zhou, M.; Blanco, A.; Ma, S. Codebleu: A method for automatic evaluation of code synthesis. *arXiv* **2020**, arXiv:2009.10297.
269. Popović, M. ChrF: Character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisbon, Portugal, 17–18 September 2015; pp. 392–395.

270. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv* **2019**, arXiv:1904.09675.
271. Wu, S.; Xie, L.; Hou, S.; Chen, G.; Jiao, H.; Liu, Z.; Guan, X.; Wu, H. GeoPyEval: An Automated Evaluation Framework for Python Code Generation Capabilities of Large Language Models in Geospatial Domains. *Trans. GIS* **2026**, *30*, e70324. [CrossRef]
272. Wu, H.; Zhao, A.; Liang, J.; Hou, S. Five-layer hierarchical network (5-HiNet) of geospatial information service for AIGC of geographic analysis model. *Acta Geod. Cartogr. Sin.* **2024**, *53*, 2053–2063. [CrossRef]
273. Krechetova, V.; Kochedykov, D. GeoBenchX: Benchmarking LLMs in agent solving multistep geospatial tasks. In Proceedings of the 1st ACM SIGSPATIAL International Workshop on Generative and Agentic AI for Multi-Modality Space-Time Intelligence, Minneapolis, MN, USA, 3 November 2025; pp. 27–35.
274. Wang, M.; Li, B.; Wang, Z.; Liu, S.; Liao, C.; Cai, Z. An Intelligent Map Cartography Framework Integrating Knowledge Graph and Large Language Model. *Geomat. Inf. Sci. Wuhan Univ.* **2026**, *51*, 840–851. [CrossRef]
275. Deng, Y.; Zhu, H.; Wang, W.; Su, J.; Chen, M.; Zhang, W. Can AI Reason Like an Urban Planner? Benchmarking Large Language Models Against Professional Judgment. *arXiv* **2026**, arXiv:2606.11678.
276. Rein, D.; Hou, B.L.; Stickland, A.C.; Petty, J.; Pang, R.Y.; Dirani, J.; Michael, J.; Bowman, S.R. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv* **2023**, arXiv:2311.12022.
277. Zhan, Y.; Xiong, Z.; Yuan, Y. Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *ISPRS J. Photogramm. Remote Sens.* **2025**, *221*, 64–77. [CrossRef]
278. Wang, P.; Wei, X.; Hu, F.; Han, W. Transgpt: Multi-modal generative pre-trained transformer for transportation. In Proceedings of the 2024 International Conference on Computational Linguistics and Natural Language Processing (CLNLP), Yinchuan, China, 19–21 July 2024; pp. 96–100.
279. Yan, Y.; Lee, J. Georeasoner: Reasoning on geospatially grounded context for natural language understanding. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, Boise, ID, USA, 21–25 October 2024; pp. 4163–4167.
280. Li, Z.; Zhou, W.; Chiang, Y.-Y.; Chen, M. Geolm: Empowering language models for geospatially grounded language understanding. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 5227–5240.
281. Xue, X.; Wei, G.; Chen, H.; Zhang, H.; Lin, F.; Shen, C.; Zhu, X.X. Reo-vlm: Transforming vlm to meet regression challenges in earth observation. *arXiv* **2024**, arXiv:2412.16583.
282. Campos, R.; Vayani, A.; Kulkarni, P.P.; Gupta, R.; Zafar, A.; Dutta, A.; Shah, M. Gaea: A geolocation aware conversational assistant. In Proceedings of the 2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Tucson, AZ, USA, 6–10 March 2026; pp. 5236–5246.
283. Soni, S.; Dudhane, A.; Debary, H.; Fiaz, M.; Munir, M.A.; Danish, M.S.; Fraccaro, P.; Watson, C.D.; Klein, L.J.; Khan, F.S. Earthdial: Turning multi-sensory earth observations to interactive dialogues. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 10–17 June 2025; pp. 14303–14313.
284. Xu, Y.; Liu, Y.; Shang, J.; Lin, J.; Ma, D. OceanAgent: A small-scale multi-modal assistant for ocean exploration. *Expert Syst. With Appl.* **2025**, *298*, 129640. [CrossRef]
285. Zguir, A.E.F.; Ofli, F.; Imran, M. Roadmind: Towards a geospatial ai expert for disaster response. *arXiv* **2025**, arXiv:2509.19354.
286. Akinboyewa, T.; Ning, H.; Lessani, M.N.; Li, Z. Automated floodwater depth estimation using large multimodal model for rapid flood mapping. *Comput. Urban Sci.* **2024**, *4*, 12. [CrossRef]
287. Alshamsi, S.; Haggag, M.; Al Rayes, L.; Alhamid, E.; Alsmirat, M.; Al Jawarneh, I.M. GeoBikeLLM: LLM-Powered Bike Route Planning with Air Quality and Vegetation-Aware Geospatial Intelligence. In Proceedings of the 2025 Sixth International Conference on Intelligent Data Science Technologies and Applications (IDSTA), Varna, Bulgaria, 1–4 September 2025; pp. 1–8.
288. Liu, L.; Pei, T.; Fang, Z.; Yan, X.; Zheng, C.; Wang, X.; Song, C.; Luan, W.; Chen, J. Extracting individual trajectories from text by fusing large language models with diverse knowledge. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *141*, 104654. [CrossRef]
289. Li, H.; Yue, P.; Wu, H.; Teng, B.; Zhao, Y.; Liu, C. A question-answering framework for geospatial data retrieval enhanced by a knowledge graph and large language models. *Int. J. Digit. Earth* **2025**, *18*, 2510566. [CrossRef]
290. Zhao, Z.; Hua, C.; Berto, F.; Lee, K.; Ma, Z.; Li, J.; Park, J. Trajevo: Trajectory prediction heuristics design via llm-driven evolution. In Proceedings of the AAAI Conference on Artificial Intelligence, Singapore, 20–27 January 2026; pp. 18072–18080.
291. Tang, J.; Kong, I.; Wang, Z. UGuideRAG: Intent-Enhanced Retrieval-Augmented Generation with User-Generated Content for Personalized Urban Tourism. In Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems, Minneapolis, MN, USA, 3–6 November 2025; pp. 90–102.
292. Ishihara, K.; Sasaki, K.; Takahashi, T.; Shiono, D.; Yamaguchi, Y. Stride-qa: Visual question answering dataset for spatiotemporal reasoning in urban driving scenes. In Proceedings of the AAAI Conference on Artificial Intelligence, Singapore, 20–27 January 2026; pp. 5257–5266.

293. Song, Y.; Xu, W.; Liu, L.; Tiwari, P.; Tian, G.; Xie, Q.; Peng, M. MFC4POI: Multi-factor collaboration for next point-of-interest recommendation using large language models. *Inf. Process. Manag.* **2026**, *63*, 104824. [CrossRef]
294. Zhang, C.; Deng, J.; Zhang, H.; Li, L.; Sun, L.; Ma, X. GeoIR-Compiler: A Geospatial Intermediate Representation and Compilation Framework for Chinese Urban Spatial Question Answering. *ISPRS Int. J. Geo-Inf.* **2026**, *15*, 310. [CrossRef]
295. Siebenmann, M.; Sánchez-Vaquerizo, J.A.; Arisona, S.; Samp, K.; Gisler, L.; Helbing, D. OGD4All: A Framework for Accessible Interaction with Geospatial Open Government Data Based on Large Language Models. In Proceedings of the 2026 IEEE Conference on Artificial Intelligence (CAI), Granada, Spain, 8–10 May 2026; pp. 882–888.
296. Zheng, W.; Yang, M.; Ren, Y.; Wang, H.; Zeng, C.; Zhang, Y. CARAG: Context-aware retrieval-augmented generation for railway operation and maintenance question answering over spatial knowledge graph. *ISPRS Int. J. Geo-Inf.* **2026**, *15*, 78. [CrossRef]
297. Jia, X.; Chen, Z.; Zhang, S.; Xue, X. Change-LISA: Language-Guided Reasoning for Remote Sensing Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 3001415. [CrossRef]
298. León, J.L.; Nogueira, V.; Salgueiro, P.; Quaresma, P. Describing land cover changes via multi-temporal remote sensing image captioning using llm, vit, and lora. *Remote Sens.* **2026**, *18*, 166. [CrossRef]
299. Chen, T.; Miao, S.; Cai, S.; Liu, H.; Mu, Y.; Yu, K.; Deng, H.; Jin, X.; Jiang, Q.; Jiang, H. GeoRAG: A Geographic Retrieval Augmented Generation Framework Based on Urban Spatio-Temporal Knowledge Graph. In Proceedings of the International Conference on Computational Linguistics and Natural Language Processing, Suzhou, China, 18–20 July 2025; pp. 130–140.
300. Moayeri, M.; Tabassi, E.; Feizi, S. Worldbench: Quantifying geographic disparities in llm factual recall. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, Rio de Janeiro, Brazil, 3–6 June 2024; pp. 1211–1228.
301. Liu, C.; Hou, D.; Xiang, J.; Yang, C.; Cui, X.; Wang, Z.; Liu, L.; Miao, Z. LandslideAgent with Multimodal LandslideBench: A Domain-Rule-Augmented Agent for Autonomous Landslide Identification and Analysis. *arXiv* **2026**, arXiv:2606.18661.
302. Jauhiainen, J.S.; Hakanpää, S.; Honkasaari, H.-P.; Kivilompolo, N.; Kurri, M.; Lehtiranta, L.; Nurminen, M. Generative AI in Participatory Urban Planning: Synthetic Inhabitants and Experts. *Land* **2026**, *15*, 407. [CrossRef]
303. Sung, S.; Jung, C.; Park, Y. Planning with ComPlanAI: Comparative insights from a human–AI evaluation of comprehensive plans. *Cities* **2026**, *172*, 106898. [CrossRef]
304. Kizilkaya, D.; Sajja, R.; Sermet, Y.; Demir, I. Toward HydroLLM: A benchmark dataset for hydrology-specific knowledge assessment for large language models. *Environ. Data Sci.* **2025**, *4*, e31. [CrossRef]
305. Zhou, X.; Zhou, J.; Wang, C.; Xie, Q.; Ding, K.; Mao, C.; Liu, Y.; Cao, Z.; Chu, H.; Chen, X. A suite of large language models for public health infoveillance. *npj Digit. Med.* **2026**, *9*, 270. [CrossRef] [PubMed]
306. Yang, T.; Xue, C.; Chen, J. ChatPlanner: A Large Language Model Framework for Personalized Public Transit Routing. *arXiv* **2026**, arXiv:2606.15315.
307. Nguyen, H.D.A.S.; Arzoumanidis, L.; Dehbi, Y. Pykci: A Compact Urban Knowledge Graph for Semantic and Spatial Queries using LLMs. *arXiv* **2026**, arXiv:2607.01605.
308. Li, Z.; Yang, M.; Zou, L.; Ma, H.; Tian, H.; Zhou, S.; Gong, W.; Zhang, K.; Chen, B.; Zhang, M. GIScholarBench: Benchmarking LLM Overconfidence in GIS Research. *arXiv* **2026**, arXiv:2606.08036.
309. Wang, Y.; Song, Z.; Ye, S.; Yang, C.; Wang, M.; Xu, Y.; An, Z.; Xu, Y.; Yu, Z. RSICCLLM: A Multimodal Large Language Model for Remote Sensing Image Change Captioning. *arXiv* **2026**, arXiv:2606.28266.
310. Lou, L.; Ke, X.; Liu, S.; Zhao, C.; Peng, J. A graph-guided LLM prompting for supply-demand reasoning in substation flood prevention. *Knowl.-Based Syst.* **2025**, *333*, 115004. [CrossRef]
311. He, Y.; Hu, C.; Wu, H.; Huang, W.; Liu, C. RoadEventGPT: A Data-Driven Autonomous Framework for Detection and Geo-Visualization of Urban Road Event Using Natural Chinese Text Data. *Trans. GIS* **2026**, *30*, e70343. [CrossRef]
312. Wang, X.; Jiang, B.; Ren, Y.; Xu, D.; Sun, Y.; Yu, R.; Li, T. LLM-GeoTextCog: A Cognitive Enhancement Framework for Geospatial Scene Understanding in Geographic Texts Using Large Language Models. *Trans. GIS* **2026**, *30*, e70203. [CrossRef]
313. Ronco, M.; Bandelli, L.; Bertolini, L.; Consoli, S.; Delforge, D.; Spadaro, A.; Verile, M.; Corbane, C. Disaster storylines and knowledge graphs from global news with large language models and retrieval-augmented generation. *Sci. Data* **2026**, *13*, 689. [CrossRef] [PubMed]
314. Zhang, Y.; Zhang, W.; Zeng, Z.; Jiang, K.; Li, J.; Min, W.; Luo, W.; Guan, Q.; Lin, J.; Yu, W. MapReader: A framework for learning a visual language model for map analysis. *Int. J. Geogr. Inf. Sci.* **2025**, *39*, 1973–2008. [CrossRef]
315. Ma, X.; Li, J.; Pei, C.; Liu, H. Geomag: A vision-language model for pixel-level fine-grained remote sensing image parsing. In Proceedings of the 33rd ACM International Conference on Multimedia, Dublin, Ireland, 27–31 October 2025; pp. 5441–5450.
316. Froehlich, J.E.; Fiannaca, A.J.; Jaber, N.M.; Tsaran, V.; Kane, S.K. Streetviewai: Making street view accessible using context-aware multimodal ai. In Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology, Busan, Republic of Korea, 28 September–1 October 2025; pp. 1–22.
317. Yin, J.; Zhong, J.; Li, P.; Pan, R.; Zeng, P.; Cao, L.; Tai, H.; Li, J.; Zhang, M.; Lu10, S. A framework for quantitative analysis of urban streetscapes leveraging vision large language models. *arXiv* **2025**, arXiv:2506.10342.

318. Perez, J.; Fusco, G. Streetscape Analysis with Generative AI (SAGAI): Vision-language assessment and mapping of urban scenes. *Geomatica* **2025**, *77*, 100063. [CrossRef]
319. Zhu, H.; Su, J.; Chen, M.; Wang, W.; Deng, Y.; Chen, G.; Zhang, W. PlanGPT-VL: Enhancing urban planning with domain-specific vision-language models. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, Suzhou, China, 4–9 November 2025; pp. 2461–2483.
320. Weiss, M.; Rahaman, N.; Pal, C. Mapbot: A multi-modal agent for geospatial analysis. In Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems, Detroit, MI, USA, 19–23 May 2025; pp. 3059–3061.
321. Chen, Z.; Zhao, L. HierLabelNet: A Two-Stage LLMs Framework with Data Augmentation and Label Selection for Geographic Text Classification. *ISPRS Int. J. Geo-Inf.* **2025**, *14*, 268. [CrossRef]
322. Li, K.; Xin, Z.; Pang, L.; Pang, C.; Deng, Y.; Yao, J.; Xia, G.; Meng, D.; Wang, Z.; Cao, X. Segearth-r1: Geospatial pixel reasoning via large language model. *arXiv* **2025**, arXiv:2504.09644.
323. Li, C.; Deng, W.; Sun, Z.; Jin, M.; Cui, H.; Li, H.; Li, S.; Yu, M.K.; Lai, M.L.; Yang, Y. S2SServiceBench: A Multimodal Benchmark for Last-Mile S2S Climate Services. *arXiv* **2026**, arXiv:2602.14017.
324. Li, P.; Ao, S.; Xue, H.; Song, Y.; de Rijke, M.; Barthélemy, J.; Bednarz, T.; Salim, F.D. Refine-POI: Reinforcement Fine-Tuned Large Language Models for Next Point-of-Interest Recommendation. *arXiv* **2025**, arXiv:2506.21599.
325. Zheng, H.; Xu, Z.; Pan, Q.; Zhao, Z.; Kong, X. Plugging Small Models in Large Language Models for POI Recommendation in Smart Tourism. *Algorithms* **2025**, *18*, 376. [CrossRef]
326. Zhou, X.; Wei, X.; Qu, Z.; Sakai, Y.; Kamigaito, H.; Watanabe, T. A Cross-Domain Tool-Augmented Vision–Language Framework for Remote Sensing Image Understanding. *Remote Sens.* **2026**, *18*, 1613. [CrossRef]
327. Jia, P.; Zhang, Y.; Zhao, X.; Li, S. GeoArena: Evaluating Open-World Geographic Reasoning in Large Vision-Language Models. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), San Diego, CA, USA, 2–7 July 2026; pp. 20886–20901.
328. Riche, A.; Zeggada, A.; Caleca, F.; Alruqimi, M.; Confuorto, P.; Guermoui, M.; Bianchini, S.; Tofani, V.; Melgani, F. MAS-LAND: A multi-agent system for landslide detection and rapid response. *Int. J. Disaster Risk Reduct.* **2026**, *143*, 106263. [CrossRef]
329. Ma, K.; Zhan, Q.; Zhang, T.; Shi, Y.; Fang, Z.; Shah, S.Y.A.; Kablan, O.A.B.K.; Zhang, B. Geo-MAG: A Knowledge Graph (KG)-enhanced Multimodal Retrieval-Augmented Generation (RAG) Framework for Geological Map Understanding. *Trans. GIS* **2026**, *30*, e70226. [CrossRef]
330. Zhang, T.; Wei, S.; Chen, S.; Yu, W.; Luo, M.; Ji, S. VectorLLM: Human-like extraction of structured building contours via multimodal LLMs. *ISPRS J. Photogramm. Remote Sens.* **2026**, *233*, 55–68. [CrossRef]
331. Shabbir, A.; Sheikh, M.U.; Munir, M.A.; Debary, H.; Fiaz, M.; Zaheer, M.Z.; Fraccaro, P.; Khan, F.S.; Khan, M.H.; Zhu, X.X. Openearthagent: A unified framework for tool-augmented geospatial agents. *arXiv* **2026**, arXiv:2602.17665.
332. Min, R.; Yao, L.; Miao, S.; Xu, S.; Liu, Y.; Zhang, C.; Di, S.; Liu, F. RemoteShield: Enable Robust Multimodal Large Language Models for Earth Observation. *arXiv* **2026**, arXiv:2604.17243.
333. Fehér, Z.Z. Bridging Natural Language and Desktop GIS Automation with LLM-Powered GIS Plugins. *EarthArXiv* **2026**. [CrossRef]
334. Ma, Y.; Zhang, Y.; Zhang, W.; Zhao, X.; Yu, W. Text-to-SpatialSQL: A LLM Based Method for Generating Spatial SQL Queries With Geo-Knowledge Extracted From Software User Manual. *Trans. GIS* **2026**, *30*, e70312. [CrossRef]
335. Liu, Q.; Li, C.; Ma, W. GATSim: Urban mobility simulation with generative agents. *Transp. Res. Part C Emerg. Technol.* **2026**, *186*, 105576. [CrossRef]
336. Li, G.; Lan, J.; Xu, P.; Wu, Z.; Ogura, M.; Feng, T. GenWorld: Empirically Grounded Urban Simulation Infrastructure for Scalable LLM-Agent Studies. *arXiv* **2026**, arXiv:2606.27650.
337. Reia, S.M.; Ferraz de Arruda, H.; Ruan, S.; Anderson, T.; Kavak, H.; Pfoser, D. Towards universal urban patterns-of-life simulation. *npj Complex.* **2026**, *3*, 25. [CrossRef]
338. Jiang, J.; Wang, Y.; Manfredini, M. Flood-LLM: An AI-Driven Framework for Property-Level Flood Risk Assessment Using Multi-Source Urban Data. *Sustainability* **2026**, *18*, 2957. [CrossRef]
339. Yan, H.; Tan, H.; Zhang, Y.; Yang, Y. Mobility-Aware Cache Framework for Scalable LLM-Based Human Mobility Simulation. *arXiv* **2026**, arXiv:2602.16727.
340. Liu, L.; Wang, Z.; Li, Z.; Ouyang, K.; Tang, Z.; Liu, C.; Li, H.; Yu, H.; Yang, W.; Yang, C. RS-Claw: Progressive Active Tool Exploration via Hierarchical Skill Trees for Remote Sensing Agents. *arXiv* **2026**, arXiv:2605.13391.
341. Yang, N.; Wang, Y.; Wei, Z.; Wu, F. MapColorAI: Designing contextually relevant choropleth map color schemes using a large language model. *Cartogr. Geogr. Inf. Sci.* **2026**, *53*, 479–497. [CrossRef]
342. Hao, X.; Zhong, X.; Yang, J.; Sun, S.; Zhang, J. Multi-task learning-based high-speed railway inflow and outflow prediction with large language models. *Transp. A Transp. Sci.* **2025**, 1–26. [CrossRef]
343. Li, H.; Chu, C.; Biljecki, F.; Shahabi, C.; Li, W. Automated Quality Assessment of Geospatial Vector Data: A GeoAI Approach using Spatial Representation Learning. *arXiv* **2026**, arXiv:2606.28390.

344. Li, S.; Tran, T.; Zhao, L.; Shafique, K.; Xiong, L. TrajGenAgent: A Hierarchical LLM Agent for Human Mobility Trajectory Generation. *arXiv* **2026**, arXiv:2606.12657.
345. Zhou, Z.; Yu, J.; Lin, Y.; Yongquan, S.; Jin, D.; Li, Y. UrbanDS: A Graph-Guided LLM Multi-Agent System for Data-Intensive Urban Tasks. *arXiv* **2026**, arXiv:2607.26724.
346. Liu, C.; Wei, Z.; Liao, H.; Dong, W. Can AI Observe Geographical Space? GVSABench For Evaluating Geo-Visuospatial Ability of Large Multimodal Models. *Trans. GIS* **2026**, *30*, e70189. [CrossRef]
347. Yong, X.; Sun, P.; Wang, Z.; Zhou, X. Intelli-planner: Towards customized urban planning via large language model empowered reinforcement learning. In Proceedings of the ACM Web Conference 2026, Dubai, United Arab Emirates, 29 June–3 July 2026; pp. 9385–9396.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.