HIERARCHIES OVER PIXELS: A BENCHMARK FOR COGNITIVE GEOSPATIAL REASONING FOR AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Beyond perception, reasoning is crucial in remote sensing, enabling advanced interpretation, inference, and decision-making. Recent advances in large language models (LLMs) have given rise to tool-augmented agents that enhance reasoning by leveraging external tools for complex analytical tasks. However, existing research on these agents in remote sensing largely focuses on perception-oriented tasks, with cognitive geospatial reasoning remaining underexplored. In this work, we systematically evaluate the geospatial reasoning capabilities of LLM-powered tool-augmented agents. To this end, we introduce **GeoHOP**, a benchmark for hierarchical geospatial reasoning. GeoHOP comprises 417 scenario-driven, hierarchy-aware tasks—such as hazard vulnerability assessment, urban heat island analysis, and forest fragmentation dynamics—spanning optical, Synthetic Aperture Radar (SAR), and infrared (IR) imagery. GeoHOP advances evaluation beyond monitoring-based recognition to cognitive-level geospatial analysis. Building upon GeoHOP, we propose **GeoPlanner**, an agent powered by LLMs that organizes 5 toolkits into functional hierarchies and executes fault-tolerant reasoning pipelines. GeoPlanner enables structured abstraction, robust recovery from tool failures, and stable long-horizon planning. Extensive experiments across diverse geospatial reasoning tasks demonstrate that GeoPlanner excels in hierarchical reasoning, cross-modal transfer, and error handling.

1 INTRODUCTION

Large language models (LLMs), empowered by generative pretraining and instruction tuning, have substantially advanced zero-shot task completion across diverse applications (Yang et al., 2024; Zhou et al., 2024). Building on this progress, LLM-driven agents can decompose goals into subtasks and orchestrate external tools, enabling robust multi-step workflows (Zhao et al., 2024; Li, 2025). However, as tasks demand increasingly granular understanding—particularly in remote sensing (RS) scenarios—these general-domain agents encounter substantial limitations. Their performance degradation stems from RS-specific challenges, including heterogeneity across modalities (optical, Synthetic Aperture Radar (SAR), infrared) and extensive variation in object size, scale, and orientation across diverse landscapes worldwide.

To address these challenges, researchers have begun adapting LLMs to RS through tool-augmented agents. Examples include Remote Sensing ChatGPT(Guo et al., 2024), RS-Agent(Xu et al., 2024), GeoLLM-Engine(Singh et al., 2024), Change-Agent(Liu et al., 2024a), Tree-GPT(Du et al., 2023). While effective for perception-oriented tasks like classification, localization, counting, and visual question answering, these approaches remain confined to visual perception. They fall short of addressing the cognitively demanding reasoning needed in realistic geospatial applications.

Beyond perception, reasoning is crucial for informed decision-making and advanced scene interpretation in Earth observation. For instance, geospatial reasoning enables identifying buildings in proximity to water bodies for flood-risk screening (Oubennaceur et al., 2019), quantifying the proportion of cropland adjacent to water sources for irrigation assessment (Fu et al., 2022), or detecting barren patches that are fully enclosed by forest to identify internal clearings requiring ecological stabilization (Hansen et al., 2009). These scenarios underscore the need for reasoning capabilities that extend beyond perception, requiring models to reason over complex spatial relationships and execute multi-step analytical workflows.

Despite recent efforts, systematic evaluation of geospatial reasoning in LLM-driven agents remains limited. For instance, ThinkGeo(Shabbir et al., 2025) introduces a benchmark for geospatial reasoning, but its scope is confined to optical imagery, leaving multi-modal reasoning underexplored. To bridge this gap, we introduce GeoHOP, a benchmark explicitly designed for cognitive-level geospatial reasoning across multiple modalities. GeoHOP consists of 417 scenario-driven, hierarchy-aware tasks spanning multiple modalities—including optical, SAR, and infrared imagery—and encompassing complex planning scenarios such as hazard risk estimation, urban heat island analysis, and ecosystem fragmentation detection. This benchmark enables rigorous evaluation across perception, geospatial reasoning, and advanced decision-making.

Building on **GeoHOP**, we further propose **GeoPlanner**, an LLM-driven agent that organizes analytic tools into functional hierarchies and executes fault-tolerant reasoning pipelines. GeoPlanner facilitates structured abstraction, dynamic workflow tracking, and robust recovery from intermediate failures, enabling cognitive reasoning in complex geospatial tasks.

Our main contributions are summarized as follows:

- We present **GeoHOP**, the first benchmark for geospatial reasoning, comprising 417 tasks across optical, SAR, and infrared modalities.
- We propose **GeoPlanner**, an agentic framework that enables cognitive reasoning with hierarchical planning, robust error handling, and support for multi-modal RS scenarios.
- We establish new baselines by comprehensively evaluating state-of-the-art LLMs on **GeoHOP**, revealing both their strengths and limitations in cognitively demanding geospatial reasoning.

2 Related Work

The extension of LLMs to RS has attracted growing interest, giving rise to a range of approaches encompassing interactive assistants, domain-specific frameworks, modular toolchains, and foundation-model paradigms. Early works primarily target specialized tasks: for example, TreeGPT (Du et al., 2023) addresses forestry applications via individual tree segmentation and ecological parameter extraction, while Change-Agent (Liu et al., 2024a) supports change detection and captioning, enabling interactive interpretation of changed regions.

Recent efforts aim to create more general-purpose RS agents. Remote Sensing ChatGPT (Guo et al., 2024) integrates ChatGPT (Brown et al., 2020) with pretrained RS networks to handle a range of perception-oriented tasks. RS-Agent (Xu et al., 2024) expands the task spectrum via scalable tool integration, handling workflows that require specialized domain expertise. In parallel, GeoLLM-Engine (Singh et al., 2024) leverages fully operational APIs with dynamic map and web interfaces to execute geospatial tasks, while UnivEARTH (Kao et al., 2025) curates domain-grounded QA tasks from NASA Earth Observatory articles to evaluate the ability of LLMs to generate executable Earth Engine code.

Despite these advances, existing frameworks often lack planning transparency and fine-grained step-level reasoning. To date, ThinkGeo (Shabbir et al., 2025) is the only work that introduces step-wise evaluation protocols for perception, planning, and geospatial reasoning. However, its coverage is restricted to optical imagery, with no support for additional modalities such as SAR and limited handling of specialized geospatial operations.

3 GeoHOP Dataset

We propose **GeoHOP**, a benchmark designed to assess the geospatial reasoning capabilities of tool-augmented agents powered by LLMs. GeoHOP integrates diverse imagery from optical, SAR, and infrared modalities with expert-curated knowledge and tool-augmented query pipelines, yielding 417 high-quality instances. Each instance couples real-world imagery with structured, multi-step reasoning challenges that require both low-level perception (e.g., segmentation, detection) and high-level decision-making (e.g., urban planning, disaster assessment). Unlike prior remote sensing benchmarks that focus narrowly on perception (e.g., classification or detection), GeoHOP emphasizes *end-to-end reasoning*, from perception through spatial analysis to actionable conclusions,

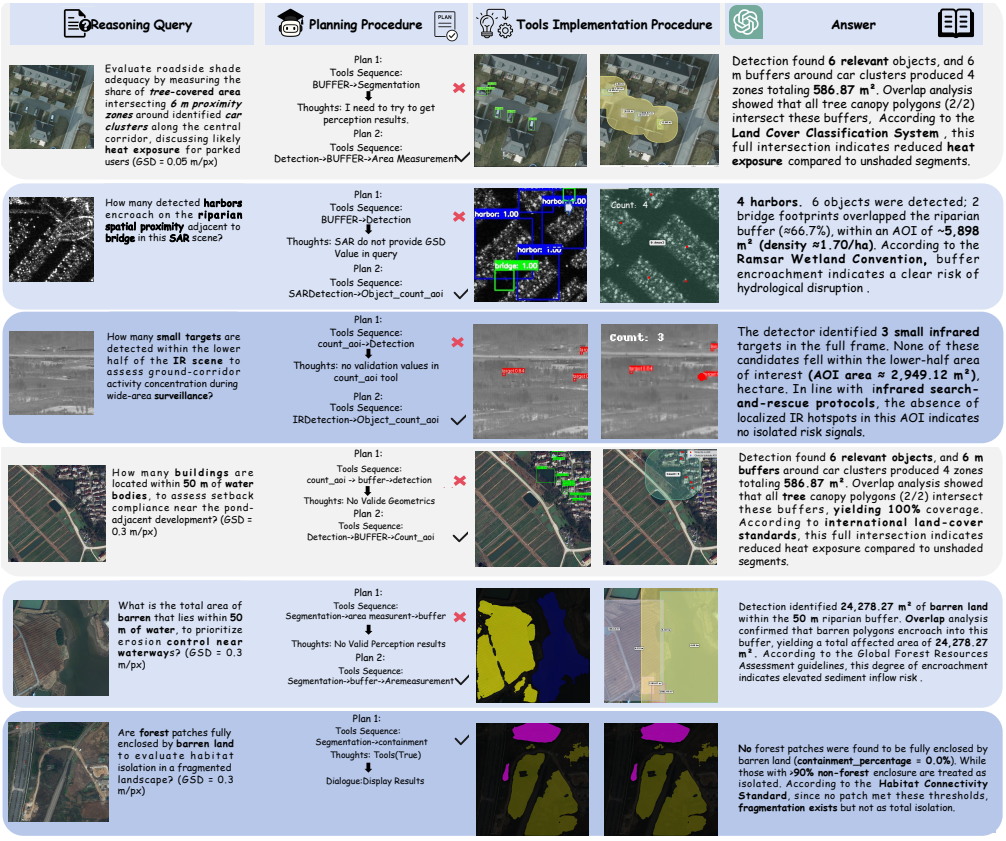


Figure 1: Representative samples from the GeoHOP benchmark.

thereby providing a rigorous testbed for agents. Figure 1 illustrates three representative samples from the GeoHOP benchmark.

3.1 Geospatial Reasoning Scenarios

A central challenge in RS is bridging the gap between low-level perceptual outputs and the high-level intelligence required for real-world decision-making. To address this, we adopt a structured taxonomy of geospatial reasoning tasks, systematically organized into seven primary domains (see Table 1).

3.2 Data Construction Pipeline

We construct GeoHOP with a two-stage pipeline (see Figure 2) that combines diverse source imagery, knowledge-augmented scenario generation, and multi-pass expert adjudication to produce 417 validated instances. The source datasets used in this pipeline are summarized in Table 2.

Stage 1: Knowledge-augmented scenario generation. We obtain a stratified sample of source imagery (by modality and scene type) and apply explicit hardness controls to select cases that require multi-tool reasoning and compositional spatial relations (e.g., proximity, containment, topology). Candidate queries and tool-chains are generated with ChatGPT-5 via in-context learning: prompts are modality-specific and seeded with expert-authored exemplars.

To ground generation in domain knowledge, we inject a compact knowledge corpus into prompts. The corpus covers four guidance categories: (i) urban greening and heat-mitigation frameworks (Twohig-Bennett & Jones, 2018; Bowler et al., 2010; Aram et al., 2019; Norton et al., 2015; Rigolon, 2016); (ii) international land-cover standards (Di Gregorio & Jansen, 1998; Mosca et al., 2020); (iii)

Domain	Tasks
Urban Planning	Urban morphology analysis; impervious surface quantification; green space accessibility; canopy coverage analysis
Disaster Assessment	Ecosystem connectivity (forest fragmentation); vegetative buffer delineation; land stability risk mapping
Environmental Monitoring	Land-use composition (agricultural intensity); agro-forestry interface mapping; ecological mosaic detection
Transportation Analysis	Infrastructure–ecosystem conflict detection; corridor impact assessment
Aviation Monitoring	Restricted airspace monitoring; multi-target intrusion detection; security violation flagging
Maritime Monitoring	Maritime anomaly detection; small-target tracking; search-and-rescue prioritization
Industrial Sites	Critical infrastructure localization; proximity-based risk assessment; diameter/area measurement; spatial relation mapping

Table 1: Domain–Task Matrix of GeoHOP, showing 7 primary domains and their associated reasoning tasks.

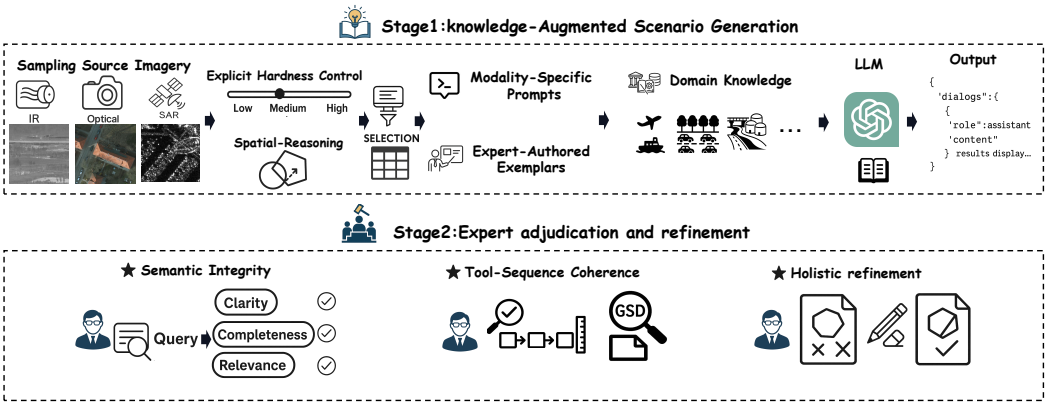


Figure 2: Pipeline for constructing the GeoHOP benchmark.

Name	Annotation Type	Resolution	Modality
LoveDA (Wang et al., 2021)	Masks	0.3 m/px	Optical
ISPRS Potsdam (Song & Kim, 2020)	Masks	0.05 m/px	Optical
OGSOD (Li et al., 2025)	Bounding boxes & Scene label	1–3 m/px	SAR
DMIST (Chen et al., 2024)	Masks & Bounding boxes	-	IR

Table 2: Datasets used as image sources in the GeoHOP benchmark.

aviation and maritime search-and-rescue doctrine for IR small-target tasks (Kim et al., 2020); and (iv) industrial safety rules for separation and proximity of hazardous assets (Ricci et al., 2021; Kukfisz et al., 2022). These anchors force the ChatGPT-5 to produce quantitative thresholds (distances, areas, class compositions) consistent with established frameworks. We also enforce operator legality checks (including ground distance sampling (GSD)-aware units) to prevent invalid tool sequences.

Stage 2: Expert adjudication and refinement. All generated candidates undergo a hierarchical, three-pass review by a panel of eight remote-sensing experts. Reviewers apply an auditable rubric that scores (1) *semantic integrity* — whether the query is meaningful and relevant; (2) *tool-sequence coherence* — whether the proposed tool-chain and parameters are logically consistent and GSD-aware; and (3) *holistic refinement* — correction of ambiguous phrasing, geometry errors, or minor inconsistencies. Every edit is recorded in a curation interface to preserve an audit trail.

Item	All	Optical	SAR	IR
Total queries	417	252	119	42
Total tool calls	1255	811	358	86
2/3/4-tools use	67/275/74	22/157/74	2/118/0	43/0/0
Maximum question length	283	279	193	283
Minimum question length	97	101	97	130
Average question length	138	136	129	176

Table 3: Statistics of GeoHOP across modalities (Optical, SAR, and IR).

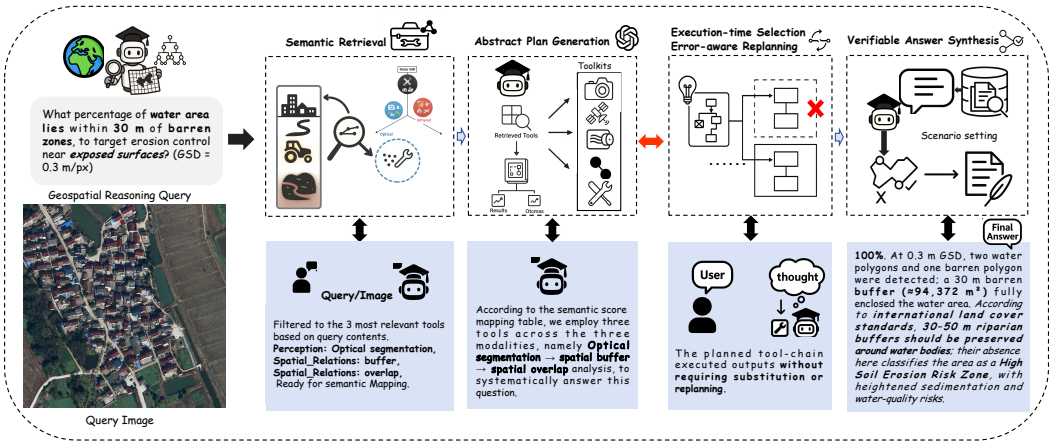


Figure 3: End-to-end workflow of the GeoPlanner Agent.

3.3 Benchmark Statistics

GeoHOP contains 417 queries across optical, SAR, and IR modalities. In total, 12 tools are invoked 1,255 times, with most queries requiring multi-tool composition. Query lengths range from short factual questions to complex compositional ones, demonstrating both diversity and reasoning depth. Detailed statistics are reported in Table 3.

Each reasoning task in GeoHOP is instantiated as a distinct scene context type, characterized by three components: (i) a Spatial Trigger (quantitative thresholds or topological relations), (ii) an Expert Interpretation (domain-grounded semantics), and (iii) a Verifiable Tool-Chain. This hierarchical design bridges perception and cognition, offering structured priors and explicit multi-step pathways while ensuring interpretability and verifiability. By covering both canonical and high-complexity reasoning scenarios, GeoHOP provides rigorous testbeds for evaluating diverse agentic capabilities, including fine-grained spatial understanding, multi-step tool composition, quantitative analysis, and context-aware risk assessment.

4 The GeoPlanner Agent

We propose **GeoPlanner**, an agent tailored for cognitively demanding geospatial reasoning in RS. GeoPlanner's LLM controller orchestrates an end-to-end workflow (see Figure 3): (i) semantic retrieval of a task-relevant toolset from a hierarchical, typed tool library; (ii) abstract, modality-aware plan generation over the retrieved tools; (iii) execution-time operator selection and parameterization with error-aware adaptation (within-toolkit substitution and prefix-preserving replanning); and (iv) verifiable answer synthesis strictly from structured tool outputs.

GeoPlanner extends the toolkit paradigm (Liu et al., 2024b) and agent framework (Fallahpour et al., 2025) to geospatial analysis by replacing flat tool sets with a multi-level, domain-specific hierarchy that incorporates typed I/O and geospatial constraints, thereby enabling reliable, modality-aware reasoning across optical, SAR, and IR data. GeoPlanner is explicitly designed to meet four RS-specific requirements: (1) *modality awareness* enforced in both retrieval and composition; (2) *hierarchical*

abstraction with typed operators plus spatial unit checks and legality constraints; (3) *fault tolerance* via within-toolkit substitution and prefix-preserving replanning with structured error context; and (4) *verifiable grounding*, validating spatial outputs (e.g., geometry validity), computing all quantitative results with tools, and having the LLM synthesize interpretable, context-aware analyses by integrating validated tool I/O with pre-encoded scenario knowledge.

4.1 Semantic Retrieval over a Hierarchical, Typed Tool Library

We organize all tools into a multi-level hierarchy $\mathcal{H}$ with **five** top-level toolkits: (1) *Perception* (Optical segmentation, Optical detection, Optical classification), (2) *Spatial-Relations* (buffer, overlap, containment), (3) *Spatial-Statistics* (distance calculation, area measurement, object counting), (4) *SAR Tools* (SAR detection, SAR classification), and (5) *IR Tools* (IR detection). Each operator advertises (i) typed I/O (e.g., vector geometries, rasters, masks, bounding boxes), (ii) spatial unit requirements, and (iii) legality constraints, and is annotated with modality tags.

Given a natural-language query Q, a semantic retriever $\mathcal{R}$ aligns Q with textual tool descriptors in $\mathcal{H}$ using sentence-transformer embeddings (Reimers & Gurevych, 2019) to induce a candidate set:

$$\mathcal{C}_{\text{candidate}} = \mathcal{R}(Q, \mathcal{H}).$$

This step reduces the planning search space, enforces modality consistency, and prevents invalid compositions at the outset.

Integration of external expert models. GeoPlanner grounds high-level plans in verifiable computations by integrating state-of-the-art open-source models into the hierarchy: **RemoteSAM** (Yao et al., 2025) (*Perception*), a unified optical segmentation/detection/classification model trained on 297 categories with fine-grained attributes; **SARATR-X** (Li et al., 2025; 2024) (*SAR Tools*), a foundation model for SAR detection/classification pretrained on 180K samples; and **DMIST/LASNet** (Chen et al., 2024) (*IR Tools*) for infrared detection. Together with Spatial-Relations and Spatial-Statistics operators, these expert tools form GeoPlanner's operational backbone, ensuring reproducibility and scalability.

4.2 Abstract Plan Generation over the Retrieved Toolset

Conditioned on $\mathrm{desc}(\mathcal{C}_{\text{candidate}})$, the planner ρ_θ (few-shot prompted with task-decomposition instructions and a constrained action schema) produces a multi-step abstract plan

$$\mathcal{P} \sim \rho_\theta(Q, \mathrm{desc}(\mathcal{C}_{\text{candidate}}))\,.$$

Each step binds to an operator *family* in $\mathcal{C}_{\text{candidate}}$, declares expected I/O types and units, and defers parameterization (e.g., buffer distances, class labels, thresholds) to execution. This yields a modality-aware blueprint that preserves semantic validity while retaining flexibility for data-driven parameter selection.

4.3 Execution-Time Selection & Error-Aware Replanning

During execution, the controller instantiates each abstract step by selecting a concrete operator t, binding arguments, and validating explicit success signals (e.g., non-empty geometries/masks and numeric stability). Failures are treated as informative signals and handled via two strategies:

(1) Within-toolkit substitution. GeoPlanner first selects a functionally similar alternative t' inside the same toolkit to preserve the abstract workflow $\mathcal{P}$ with minimal disruption (e.g., switching from *overlap* to *containment* within Spatial-Relations).

(2) Prefix-preserving replanning (error-aware). If local substitution fails, the validated prefix of $\mathcal{P}$ is retained, while a structured error context E (failed operator, arguments, logs, and validated intermediates) is injected into retrieval. A refreshed set $\mathcal{C}(Q, E)$ supports regeneration of the failing suffix. A step succeeds if any candidate succeeds ($\bigcup_{t\in\mathcal{C}(Q,E)} \mathrm{succ}_t$); total failure occurs only when all fail ($\bigcap_{t\in\mathcal{C}(Q,E)} \mathrm{fail}_t$). Retries are capped to bound cost, yielding a budget-conscious controller that preserves progress under adverse conditions.

4.4 VERIFIABLE ANSWER SYNTHESIS

GeoPlanner synthesizes the final answer strictly from validated tool outputs: (i) all numeric values (counts, areas, distances, proportions) are computed by operators with explicit spatial units; (ii) aggregations include provenance (modality, operator names) and spatial unit conversions when necessary; and (iii) the LLM only provides natural-language contextualization. This separation reduces hallucination risk and improves transparency and reproducibility.

5 EXPERIMENT

To evaluate the reasoning and tool-use capabilities of **GeoPlanner** under real-world RS scenarios, we conduct comprehensive evaluations on the **GeoHOP** benchmark. As the controller in GeoPlanner is powered by LLMs, we evaluate a diverse set of frontier and open-source models: GPT-3.5 (Ouyang et al., 2022), GPT-4-1106-Preview (Achiam et al., 2023), GPT-4o (Hurst et al., 2024), Claude-4 (Anthropic, 2024), Gemini-2.5-flash (Team et al., 2023), DeepSeek-V3 (Bi et al., 2024), and Qwen2.5 (72B and 32B Instruct) (Team, 2024). All experiments are conducted on an NVIDIA A5000 GPU within the OpenCompass evaluation platform.

5.1 EVALUATION STRATEGY

We follow the ReAct-style (Yao et al., 2023) evaluation protocol, which includes both step-by-step and end-to-end modes. These protocols define how tool-augmented reasoning is assessed, and our evaluation criteria remain fully consistent with ReAct, ensuring comparability across different agents. In our framework, GeoPlanner orchestrates the workflow of tool invocations, which complements rather than contradicts the ReAct protocol by providing structured execution while adhering to its evaluation standards.

Unlike GTA (Wang et al., 2024), which computes final answer accuracy (AnsAcc) through deterministic string matching, we argue this approach is insufficient for GeoHOP's complex reasoning scenarios. String matching is brittle to minor lexical variations and may misclassify semantically correct but differently phrased predictions. This issue is particularly problematic in GeoPlanner, where final answers synthesize information from *query semantics, tool inputs/outputs, and scenario knowledge* rather than matching a single canonical string.

Additionally, argument consistency (ArgAcc) in GeoHOP fundamentally relates to the agent's ability to correctly propagate arguments across multi-step tool chains. Pure string-level comparison cannot adequately capture semantic equivalence between valid arguments (e.g., polygon references, buffered geometries, or object lists).

To address these limitations, we adopt **LLM-as-a-judge** for both AnsAcc and ArgAcc evaluation. For each query, we construct curated evaluation prompts and employ GPT-4o-mini as an automatic judge to assess whether predicted results and arguments are semantically consistent with ground truth. This approach provides a more reliable and context-aware measure of success in GeoHOP, aligning evaluation with the hierarchical, multi-fact reasoning required by real-world geospatial tasks while maintaining compatibility with the GTA evaluation framework.

5.2 MAIN RESULTS

We conduct experiments on the **GeoHOP benchmark** to comprehensively evaluate reasoning and tool-use abilities in real RS tasks. Unlike prior efforts relying on synthetic prompts or shallow tool interactions, GeoHOP introduces *multimodal, scenario-driven tasks* that compel agents to invoke multiple **toolkits** spanning perception, spatial analysis, SAR, and IR. These tasks are grounded in satellite and aerial imagery, requiring models to exhibit *hierarchical reasoning* and *quantitative precision* during multi-step execution, enabling systematic evaluation of cognitive-level geospatial reasoning.

Table 4 reports step-by-step (Inst., Tool., Arg., Summ.) and end-to-end (Ans.) performance on GeoHOP. Figure 4 presents example answers generated by different models for a single query. **GPT-4o** attains the highest end-to-end accuracy and leads in instruction understanding, tool selection, and summary accuracy. **Claude 4** is a close second and achieves the strongest **argument consistency**,

Model	Step-by-Step Metrics				End-to-End Metrics
	Inst.	Tool.	Arg.	Summ.	Ans.
GPT-4o	**89.52**	**52.32**	**6.52**	**14.78**	**15.65**
Claude 4	71.42	41.49	7.61	14.39	15.11
GPT-4-1106 Preview	57.33	34.35	4.38	7.19	7.91
Gemini-2.5-flash	24.10	13.60	1.60	4.30	4.80
DeepSeek-V3	18.60	10.90	0.70	3.60	3.80
Qwen2.5-72B-Instruct	19.10	12.30	2.40	3.60	3.60
Qwen2.5-32B-Instruct	17.40	10.30	1.60	2.60	2.90
GPT-3.5	9.90	2.00	0.30	2.60	2.90

Table 4: Evaluation results across models in **GeoPlanner** on the **GeoHOP** benchmark. Models are sorted by end-to-end performance (Ans.: final answer accuracy). The table reports step-by-step execution metrics (Inst.: instruction understanding, Tool.: tool selection, Arg.: argument consistency, Summ.: summary accuracy) and the final answer accuracy.

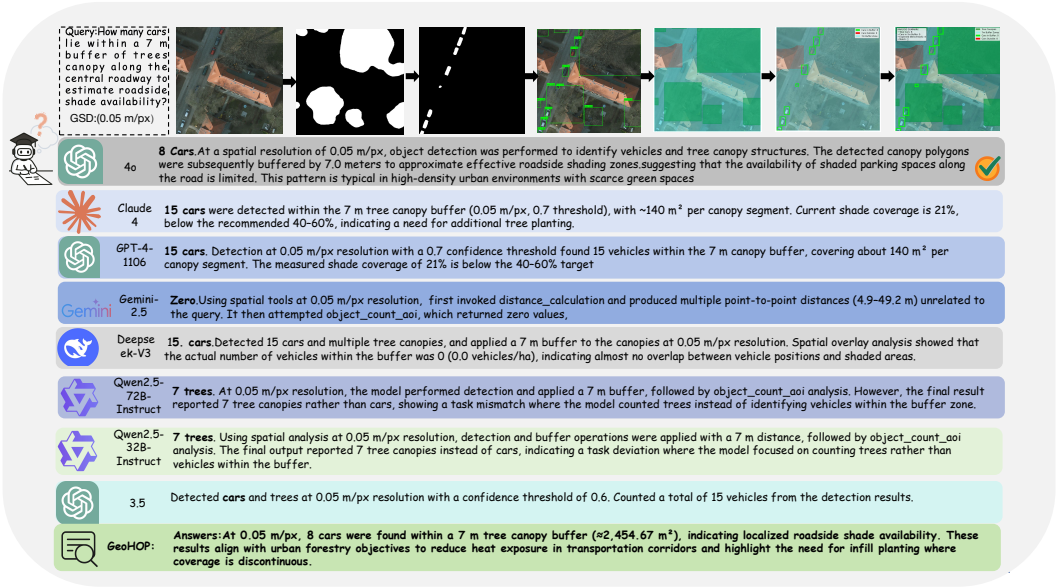


Figure 4: Comparative Evaluation of LLMs on Spatial Reasoning Tasks.

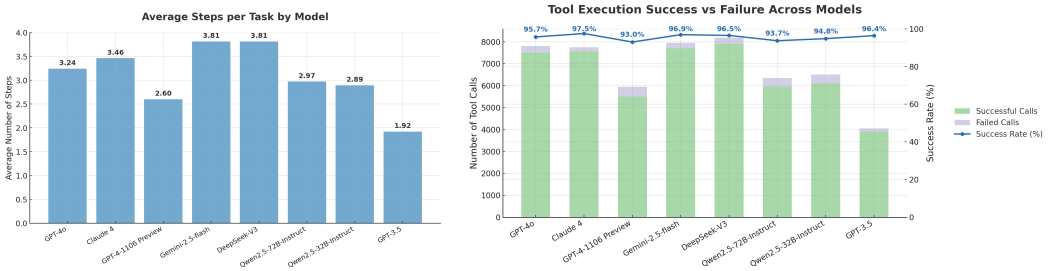
but lags GPT-4o on Inst./Tool. **GPT-4-1106 Preview** forms a mid-tier baseline. **Gemini-2.5-flash** and **DeepSeek-V3** trail further, while **Qwen2.5-72B/32B-Instruct** and **GPT-3.5** exhibit poor instruction adherence and fragile tool formatting. Across models, **argument consistency** is the dominant bottleneck, and weaker **tool selection** strongly correlates with degraded end-to-end accuracy. Even for top models, absolute scores remain modest, underscoring the difficulty of robust geospatial reasoning and faithful summary generation on GeoHOP. As shown in Figure 4, the figure presents the comparative evaluation of multiple LLMs on the GeoHOP benchmark. GPT-4o performed the best among the evaluated models, yet the results highlight that even state-of-the-art models like GPT-4o are not fully "intelligent" when compared with the ground truth. For example, they often lack higher-level reasoning such as *"These results align with urban forestry objectives to reduce heat exposure in transportation corridors"*, which is explicitly included in the ground truth.

5.3 Task Analysis

Across models, we observe clear differences in task completion and execution dynamics (Figures 5 and 6). In Figure 5, API-based models—notably **GPT-4o** and **GPT-4-1106 Preview**—complete

the largest number of tasks with comparatively few failures, whereas open-source models such as **Qwen2.5-32B/72B** and **DeepSeek-V3** accumulate substantially more failures. Figure 6(a) shows that **Gemini-2.5-Flash** and **DeepSeek-Chat** tend to use longer reasoning chains (≈ 3.81 steps/task), **Claude-Sonnet-4** is also relatively long (≈ 3.46), **GPT-4o** is moderate (≈ 3.24), while **GPT-3.5** is shortest (≈ 1.92), often at the expense of task success. Importantly, tool reliability alone is not sufficient: Figure 6(b) indicates **Claude 4** attains the highest tool-call success rate ($> 97\%$) yet lags behind **GPT-4o** in overall task success, suggesting that stable local execution must be paired with effective global planning, argument selection, and summarization. Overall, **GPT-4o** offers the best balance of chain length, tool reliability, and task-level success across diverse scenarios.

Figure 5: Task-level success and failure distribution across models. The plot reports, for each evaluated model, the number of tasks completed successfully versus failed attempts. Open-source models exhibit higher failure counts compared to API-based models, while GPT-4o family shows the highest success rates.



(a) (b)

Figure 6: Comparison of (a) average reasoning steps required per task across models and (b) tool-execution success vs. failure distributions across the same models.

6 Conclusion

In this work, we presented **GeoHOP**, the first benchmark explicitly designed for cognitive-level geospatial reasoning across optical, SAR, and IR modalities. GeoHOP comprises 417 hierarchy-aware, scenario-driven tasks that extend beyond perception to demand structured, multi-step reasoning. To address these challenges, we introduced **GeoPlanner**, an LLM-powered agent that organizes analytic tools into functional hierarchies, supports modality-aware planning, and ensures robust error recovery during execution. Extensive experiments on GeoHOP demonstrate that frontier models such as the GPT-4 family currently achieve the strongest performance. Simultaneously, our benchmark reveals significant potential for improvement in argument accuracy, summary generation, and multimodal toolkit integration. These findings indicate that while progress has been made, substantial headroom remains for advancing reasoning, planning, and execution capabilities in real-world remote sensing scenarios. By providing both a high-fidelity benchmark and a fault-tolerant agentic framework, our work establishes a rigorous foundation for evaluating and advancing multimodal reasoning in RS.

USE OF LLMS

In this work, we employed large language models (LLMs) to assist with language refinement and to enhance the overall coherence of the manuscript. Specifically, LLMs were used to polish sentence-level grammar and improve the logical flow between sections. In addition, we utilized LLMs to generate a set of scalable vector graphics (SVGs), which served as schematic figures to illustrate the conceptual framework of our study. These applications were limited to stylistic editing and visualization support; all research design, data analysis, and substantive conclusions were conducted and validated independently by the authors.

ETHICS STATEMENT

This work uses only publicly available remote sensing datasets that do not contain personal or sensitive information. All experiments were conducted in compliance with relevant data usage licenses. We acknowledge potential risks of misuse of geospatial AI technologies, such as unauthorized surveillance or environmental misinterpretation, and emphasize that our methods are intended solely for scientific and societal applications, including environmental monitoring and sustainable development. No human subjects were involved in this study.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our results, we provide the complete framework code of the Spatial Reasoning Agent, including tool integration, planning modules, and evaluation scripts.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Anthropic. Claude opus 4 and claude sonnet 4 system card. `https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf`, 2024. Accessed: 2025-09-24.

Farshid Aram, Ester Higueras García, Ebrahim Solgi, and Soran Mansournia. Urban green space cooling effect in cities. *Heliyon*, 5(4), 2019.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

Diana E Bowler, Lisette Buyung-Ali, Teri M Knight, and Andrew S Pullin. Urban greening to cool towns and cities: A systematic review of the empirical evidence. *Landscape and urban planning*, 97(3):147–155, 2010.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Shengjia Chen, Luping Ji, Sicheng Zhu, Mao Ye, Haohao Ren, and Yongsheng Sang. Towards dense moving infrared small target detection: New datasets and baseline. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

Antonio Di Gregorio and Louisa JM Jansen. Land cover classification system (lccs): classification concepts and user manual. *FAO, Rome*, 1998.

Siqi Du, Shengjun Tang, Weixi Wang, Xiaoming Li, and Renzhong Guo. Tree-gpt: Modular large language model expert system for forest remote sensing image understanding and interactive analysis. *arXiv preprint arXiv:2310.04698*, 2023.

A. Fallahpour, J. Ma, A. Munim, H. Lyu, and B. Wang. Medrax: Medical reasoning agent for chest x-ray. *arXiv preprint arXiv:2502.02673*, 2025.

Jianyu Fu, Weiguang Wang, Benjamin Zaitchik, Wanshu Nie, Esther Xu Fei, Scot M Miller, and Ciaran J Harman. Critical role of irrigation efficiency for cropland expansion in western china arid agroecosystems. *Earth's Future*, 10(9):e2022EF002955, 2022.

Haonan Guo, Xin Su, Chen Wu, Bo Du, Liangpei Zhang, and Deren Li. Remote sensing chatgpt: Solving remote sensing tasks with chatgpt and visual models. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pp. 11474–11478. IEEE, 2024.

Matthew C Hansen, Stephen V Stehman, Peter V Potapov, Belinda Arunarwati, Fred Stolle, and Kyle Pittman. Quantifying changes in the rates of forest clearing in indonesia from 1990 to 2005 usingremotely sensed data sets. *Environmental Research Letters*, 4(3):034001, 2009.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Chia Hsiang Kao, Wenting Zhao, Shreelekha Revankar, Samuel Speas, Snehal Bhagat, Rajeev Datta, Cheng Perng Phoo, Utkarsh Mall, Carl Vondrick, Kavita Bala, et al. Towards llm agents for earth observation. *arXiv preprint arXiv:2504.12110*, 2025.

Inchul Kim, Chongju Chae, and Soyeong Lee. Simulation study of the iamsar standard recovery maneuvers for the improvement of serviceability. *Journal of Marine Science and Engineering*, 8 (6):445, 2020.

Bożena Kukfisz, Aneta Kuczyńska, Robert Piec, and Barbara Szykuła-Piec. Research on the safety and security distance of above-ground liquefied gas storage tanks and dispensers. *International journal of environmental research and public health*, 19(2):839, 2022.

Weijie Li, Wei Yang, Tianpeng Liu, Yuenan Hou, Yuxuan Li, Zhen Liu, Yongxiang Liu, and Li Liu. Predicting gradient is better: Exploring self-supervised learning for sar atr with a joint-embedding predictive architecture. *ISPRS Journal of Photogrammetry and Remote Sensing*, 218:326–338, 2024. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2024.09.013. URL https://www.sciencedirect.com/science/article/pii/S0924271624003514.

Weijie Li, Wei Yang, Yuenan Hou, Li Liu, Yongxiang Liu, and Xiang Li. Saratr-x: Toward building a foundation model for sar target recognition. *IEEE Transactions on Image Processing*, 34:869–884, 2025. doi: 10.1109/TIP.2025.3531988.

Xinzhe Li. A review of prominent paradigms for LLM-based agents: Tool use, planning (including RAG), and feedback learning. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 9760–9779, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.652/.

Chenyang Liu, Keyan Chen, Haotian Zhang, Zipeng Qi, Zhengxia Zou, and Zhenwei Shi. Changeagent: Towards interactive comprehensive remote sensing change interpretation and analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 2024a.

Yanming Liu, Xinyue Peng, Jiannan Cao, Yuwei Zhang, Xuhong Zhang, Sheng Cheng, Xun Wang, Jianwei Yin, and Tianyu Du. Tool-planner: Task planning with clusters across multiple tools. *arXiv preprint arXiv:2406.03807*, 2024b.

Nicola Mosca, Antonio Di Gregorio, Matieu Henry, Rashed Jalal, and Palma Blonda. Object-based similarity assessment using land cover meta-language (lcml): Concept, challenges, and implementation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:3790–3805, 2020.

Briony A Norton, Andrew M Coutts, Stephen J Livesley, Richard J Harris, Annie M Hunter, and Nicholas SG Williams. Planning for cooler cities: A framework to prioritise green infrastructure to mitigate high temperatures in urban landscapes. *Landscape and urban planning*, 134:127–138, 2015.

Khalid Oubennaceur, Karem Chokmani, Miroslav Nastev, Rachid Lhissou, and Anas El Alem. Flood risk mapping for direct damage to residential buildings in quebec, canada. *International journal of disaster risk reduction*, 33:44–54, 2019.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Federica Ricci, Giordano Emrys Scarponi, Elsa Pastor, Eulàlia Planas, and Valerio Cozzani. Safety distances for storage tanks to prevent fire damage in wildland-industrial interface. *Process Safety and Environmental Protection*, 147:693–702, 2021.

Alessandro Rigolon. A complex landscape of inequity in access to urban parks: A literature review. *Landscape and urban planning*, 153:160–169, 2016.

Akashah Shabbir, Muhammad Akhtar Munir, Akshay Dudhane, Muhammad Umer Sheikh, Muhammad Haris Khan, Paolo Fraccaro, Juan Bernabe Moreno, Fahad Shahbaz Khan, and Salman Khan. Thinkgeo: Evaluating tool-augmented agents for remote sensing tasks. *arXiv preprint arXiv:2505.23752*, 2025.

Simranjit Singh, Michael Fore, and Dimitrios Stamoulis. Geollm-engine: A realistic environment for building geospatial copilots. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 585–594, 2024.

Ahram Song and Yongil Kim. Semantic segmentation of remote-sensing imagery using heterogeneous big data: International society for photogrammetry and remote sensing potsdam and cityscape datasets. *ISPRS International Journal of Geo-Information*, 9(10):601, 2020.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Qwen Team. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2, 2024.

Caoimhe Twohig-Bennett and Andy Jones. The health benefits of the great outdoors: A systematic review and meta-analysis of greenspace exposure and health outcomes. *Environmental research*, 166:628–637, 2018.

Jize Wang, Ma Zerun, Yining Li, Songyang Zhang, Cailian Chen, Kai Chen, and Xinyi Le. Gta: a benchmark for general tool agents. *Advances in Neural Information Processing Systems*, 37: 75749–75790, 2024.

Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021.

Wenjia Xu, Zijian Yu, Boyang Mu, Zhiwei Wei, Yuanben Zhang, Guangzuo Li, and Mugen Peng. Rs-agent: Automating remote sensing tasks through intelligent agent. *arXiv preprint arXiv:2406.07089*, 2024.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32, 2024.

Liang Yao, Fan Liu, Delong Chen, Chuanyi Zhang, Yijun Wang, Ziyun Chen, Wei Xu, Shimin Di, and Yuhui Zheng. Remotesam: Towards segment anything for earth observation. *arXiv preprint arXiv:2505.18022*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19632–19642, 2024.

Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. Larger and more instructable language models become less reliable. *Nature*, 634(8032):61–68, 2024.

A Appendix

You may include other additional sections here.