# GEOMAPBENCH: COGNITION-AWARE CHARACTERIZATION OF GEOSPATIAL CAPABILITIES IN VISION-LANGUAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Vision-language models (VLMs) are increasingly applied to maps, remote-sensing imagery, and other geospatial data, yet their geospatial capabilities remain difficult to evaluate systematically because existing benchmarks typically cover only subsets of cartographic understanding, spatial reasoning, geographic knowledge, and structured geospatial problem solving. We develop a comprehensive taxonomy that decomposes geospatial capability into 23 fine-grained tasks organized into four domains: Visual Perception and Interpretation, Spatial Reasoning and Geometric Processing, Geospatial Language and Semantic Reasoning, and Applied Geospatial and Earth-System Reasoning. Guided by this taxonomy, we construct GeoMapBench, a benchmark containing 2,300 examples with 100 instances per task. Each example is additionally assigned a level of the revised Bloom taxonomy, enabling analysis across both geospatial capability and cognitive complexity. We also construct GeoMapRAGCorpus, a benchmark-separated multimodal retrieval corpus containing 180,344 records from seven public geospatial source families. The corpus covers structured geographic data, text, map imagery, and geocoded geographic images. We benchmark five contemporary VLMs on all 2,300 GeoMapBench examples and observe substantial variation across tasks, with no single model consistently dominating the full task space. Finally, to demonstrate the downstream utility of the released resources, we evaluate GeoMapPlus, a task-aware retrieval- and tool-augmented inference pipeline built using GeoMapRAGCorpus. GeoMapPlus improves both task-aware and strict evaluation performance over its underlying VLM, demonstrating that the corpus can support measurable improvements in geospatial reasoning.

## 1 INTRODUCTION

Geospatial reasoning is intrinsically multimodal. Maps communicate through symbols, labels, geometry, scale, topology, and coordinate systems; remote-sensing imagery adds scene semantics, object structure, and change over time; and many practical questions require geographic facts or exact operations over coordinates and networks. A capable vision–language model (VLM) must therefore do more than recognize what appears in an image. It must connect visual evidence to spatial structure, interpret geographic language, perform or invoke geometric computation, and distinguish information visible in the input from information that must be retrieved. Existing benchmarks provide valuable views of this problem, but usually isolate only one part of the capability space. Evaluations centered on maps, Earth observation, or urban geometry (Mukhopadhyay et al., 2025; Danish et al., 2025; Bao et al., 2026) consequently answer different questions, making a single score difficult to interpret as a general measure of geospatial competence. CHOICE organizes remote-sensing tasks, while Spatial-RAG studies retrieval over structured geographic data (An et al., 2025; Yu et al., 2026). Section 2 compares GeoMapBench with these and other geospatial resources.

This paper develops a unified evaluation resource around that observation. We first organize geospatial capability into a four-domain taxonomy with 23 fine-grained, operationally defined tasks spanning visual perception and interpretation, spatial reasoning and geometric processing, geospatial language and semantics, and applied geospatial and Earth-system reasoning. We then instantiate every task with 100 examples to obtain **GeoMapBench**, a 2,300-example benchmark



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

in which each record carries an explicit answer contract, evaluation rule, source metadata, and a cognitive-process label from the revised Bloom taxonomy. The benchmark therefore supports two complementary views of performance: *what geospatial capability* an example exercises and *what kind of cognitive operation* its prompt requests. Rather than collapsing heterogeneous outputs into a single exact-match format, GeoMapBench preserves task-appropriate targets including labels, text, numerical quantities, coordinates, masks, graphs, routes, and polygons, and reports both strict and normalized task-aware scores.

Evaluation alone does not address a second practical limitation of geospatial VLMs: many questions require evidence that is not contained in the query image and should not be expected to reside reliably in model parameters. We therefore construct **GeoMapRAGCorpus**, a benchmark-separated multimodal retrieval collection with 180,344 records drawn from seven public source families. The corpus spans structured geographic entities, factual text, map imagery, coordinate-system descriptions, indicators, and geocoded photographs, while retaining record-level provenance and source-specific licensing information. Benchmark-derived exclusion rules remove identifiable overlap through shared source records, substantial textual matches, exact indicator observations, and nearby geocoded content. The resulting collection supports both textual and visual evidence retrieval while remaining independent of a particular retrieval algorithm.

We use the benchmark in two complementary experiments. Five contemporary VLMs first answer the same 2,300 examples without retrieval. Their aggregate scores are relatively close compared with pronounced task-level variation: different models lead different leaves, and several applied or structurally demanding capabilities remain weak across the suite. We then study **GeoMapPlus**, a task-aware retrieval- and tool-augmented pipeline built on GeoMapRAGCorpus. Relative to direct answering with the same final VLM, the full pipeline improves both task-aware and strict performance, especially on several exact geometric and data-backed tasks, while degrading some visually grounded leaves. The result demonstrates downstream utility without implying that augmentation is uniformly beneficial or that retrieval alone causes the observed gain.

**Our contributions.** Our main contributions can be summarized as follows:

- **Geospatial capability taxonomy.** We define 23 operational tasks across visual perception, spatial reasoning, geospatial language, and applied Earth-system reasoning. A second axis based on the revised Bloom taxonomy captures the cognitive operation requested by each prompt, separating *what* capability is tested from *how* it is exercised.
- **GeoMapBench.** We construct a 2,300-example benchmark with 100 examples per task, spanning maps, remote-sensing imagery, coordinates, networks, geographic language, and applied geographic knowledge. Task-specific answer formats and evaluators preserve the structure of each problem while supporting both strict and task-aware scoring.
- **GeoMapRAGCorpus.** We release a benchmark-separated multimodal evidence resource with 180,344 records from seven public geospatial source families. The corpus combines textual, structured, cartographic, and photographic evidence with source provenance, providing a reusable substrate for retrieval- and tool-augmented geospatial systems.
- **Evaluation of VLMs and augmentation.** We benchmark five contemporary VLMs over the full task space and analyze their performance by task, domain, and Bloom level. We further evaluate GeoMapPlus in a paired setting, showing that selective retrieval and geographic tools can improve overall performance while remaining strongly task dependent.

## 2 Related Work

GeoMapBench lies at the intersection of four research threads: cartographic understanding, Earth-observation VLMs, spatial and geographic reasoning, and retrieval- or tool-augmented language models. Each thread captures an important part of geospatial intelligence, but they are typically evaluated in isolation. We review these lines below and position GeoMapBench as a common evaluation layer that connects them without collapsing their distinct problem structures.

**Map-centric benchmarks.** Recent work has begun to evaluate VLMs directly on cartographic inputs. MAPWise focuses on choropleth maps and questions that combine visual encoding with

spatial relations (Mukhopadhyay et al., 2025); CartoMapQA studies foundational map-reading skills such as symbols, scale, and routing (Ung et al., 2025); MapVerse broadens evaluation to diverse real-world map styles (Bhat et al., 2026); and PEACE targets holistic understanding of geologic maps (Huang et al., 2025). These benchmarks establish that maps pose challenges distinct from natural-image VQA, but each emphasizes a particular map family or subset of cartographic skills. GeoMapBench complements them by placing map understanding inside a larger capability space that also includes coordinate operations, graph reasoning, geographic language, remote sensing, and applied geographic knowledge.

**Remote sensing and Earth-observation VLMs.** A second line of work evaluates or develops multimodal models for Earth observation. GEOBench-VLM covers tasks including scene understanding, localization, segmentation, and temporal analysis (Danish et al., 2025); XLRS-Bench stresses reasoning over extremely large remote-sensing imagery (Wang et al., 2025); and CHOICE organizes remote-sensing capability into 23 leaf tasks (An et al., 2025). GeoChat and EarthDial instead develop geospatially specialized VLMs for grounded remote-sensing dialogue and multisensory Earth observation (Kuckreja et al., 2024; Soni et al., 2024). This literature provides strong coverage of imagery-centered geospatial perception, but it does not by itself characterize the broader mixture of cartographic conventions, coordinate systems, networks, textual relations, and structured geographic operations required in general geospatial reasoning. GeoMapBench is designed to connect these settings rather than treat them as separate evaluation regimes.

**Spatial and geospatial reasoning.** Spatial reasoning has also been studied in text and structured geographic QA. SPARTQA and StepGame probe qualitative and multi-hop spatial relations in language (Mirzaee et al., 2021; Shi et al., 2022); GeoQuestions1089 evaluates natural-language questions against GeoSPARQL-style geographic knowledge (Kefalidis et al., 2023); and MapQA couples open-domain geospatial questions with entity geometries extracted from map data (Li et al., 2025). More recent benchmarks such as CityCube and UrbanGeoEval examine cross-view urban reasoning and separate geographic memory from explicit city-scale geometric inference (Xu et al., 2026; Bao et al., 2026). These resources isolate important aspects of spatial competence, yet they typically operate in one modality or one reasoning substrate. Our taxonomy unifies visual, linguistic, numerical, and structured outputs so that these capabilities can be compared under one evaluation framework.

**Retrieval and tool augmentation.** Retrieval-augmented generation grounds language models in external evidence (Lewis et al., 2020), while ReAct and Toolformer show how language models can interleave reasoning with actions or specialized tools (Yao et al., 2023; Schick et al., 2023). In the geospatial domain, Spatial-RAG combines semantic retrieval with explicit spatial filtering over structured geographic data (Yu et al., 2026). These methods motivate augmenting models when the answer depends on up-to-date facts or exact computation, but they do not provide a common multimodal evidence resource paired with a broad geospatial benchmark. GeoMapRAGCorpus is intended to fill that role: it is released independently from GeoMapPlus and separated from GeoMapBench, allowing future work to compare retrieval, spatial filtering, and tool-use strategies while holding both the evaluation target and evidence pool fixed.

## 3 GeoMapBench

GeoMapBench is designed to expose where geospatial VLM capabilities differ rather than compressing them into a single notion of "map understanding." We first define the capability taxonomy and its complementary Bloom-based cognitive axis, then describe benchmark construction, evaluation, quality control, and separation from the retrieval corpus.

### 3.1 Capability Taxonomy

The taxonomy is task-centered: every leaf specifies a model-visible input, a geospatial operation, and a checkable output. This choice prevents superficially similar prompts from being treated as the same capability. Reading a map label is distinct from anchoring that label to a feature; identifying a road is distinct from constructing a routable graph; recognizing a coordinate reference system is distinct from executing a coordinate transformation. At the same time, the taxonomy keeps visual, textual, numerical, and structured tasks in one hierarchy without requiring them to

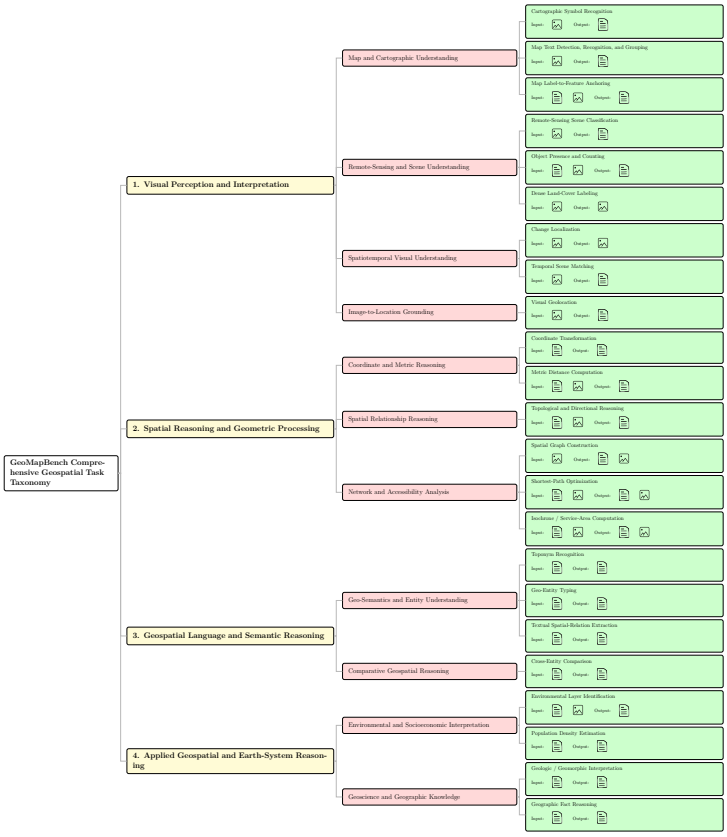



Figure 1: Compact overview of the GeoMapBench geospatial capability taxonomy. The taxonomy spans cartographic and remote-sensing understanding, spatiotemporal visual analysis, image-to-location grounding, coordinate and metric reasoning, spatial relations and network analysis, geospatial language and semantics, and applied environmental and Earth-system reasoning. It provides the organizing structure for benchmark construction, task-balanced evaluation, task-level error analysis, and task-aware retrieval or tool routing. A more detailed version including input/output modalities and task descriptions is provided in Appendix Figure 4.

share an answer format or metric. Figure 1 gives a compact overview of the hierarchy; the more detailed version with task descriptions is provided in Appendix Figure 4.

The four domains deliberately cut across conventional dataset boundaries. **Visual Perception and Interpretation** contains nine leaves covering cartographic symbols, map text and label anchoring, remote-sensing scenes and objects, dense land cover, temporal change and matching, and visual geolocation. **Spatial Reasoning and Geometric Processing** contains six leaves for coordinate transformation, metric distance, topological and directional relations, road-graph construction, shortest paths, and service areas; their outputs range from scalars and coordinates to graphs and polygons. **Geospatial Language and Semantic Reasoning** contributes four leaves for toponym recognition, entity typing, textual spatial relations, and cross-entity comparison. **Applied Geospatial and Earth-System Reasoning** contributes four more for environmental rasters, population density, geologic or geomorphic interpretation, and geographic fact reasoning. The 9/6/4/4 decomposition yields the 23 leaves summarized in Appendix A.

Task identity describes *what* is tested, whereas the revised Bloom taxonomy describes the operation requested by an example: Remember, Understand, Apply, Analyze, Evaluate, or Create (Anderson & Krathwohl, 2001; Krathwohl, 2002). Assignment follows the actual prompt and reference answer rather than the task name. Each leaf declares five or six defensible operations; a deterministic conversion preserves the source target while adapting the question and expected output. Five-level leaves allocate 20 examples per level, and six-level leaves use 17/17/17/17/16/16, yielding 384 Re-

member, 424 Understand, 424 Apply, 424 Analyze, 412 Evaluate, and 232 Create records overall. These labels are diagnostic rather than a universal difficulty order because supported levels, task mixtures, and answer structures differ. Appendix A.1 details the construction.

This two-axis design is important for interpretation. Two examples may both concern maps, yet differ substantially in what they ask a model to do: one may require recognizing a symbol, another may require checking whether a proposed route satisfies a geometric constraint, and a third may require constructing a structured output from several pieces of evidence. By storing task identity and Bloom operation separately, GeoMapBench can distinguish errors caused by missing geospatial knowledge from errors caused by format, decomposition, or synthesis demands. The representation is also extensible: new builders can attach to the same leaves or introduce carefully justified new leaves without changing the meaning of the existing ones.

### 3.2 Benchmark Construction

GeoMapBench contains 100 examples for each of the 23 leaves. We draw on established geographic and remote-sensing datasets together with public geospatial resources, selecting sources according to the semantics of each task rather than forcing a common data modality. Visual tasks use resources including Maki, the Paris and Jerusalem Maps Text Dataset, EuroSAT (Helber et al., 2019), OpenEarthMap (Xia et al., 2022), RSVQA-LR (Lobry et al., 2020), SpaceNet imagery, OpenStreetMap, and geocoded Wikimedia photographs. Geometric tasks additionally use Natural Earth and PROJ/EPSG, while language and applied tasks draw on GeoWebNews, GeoNames, SpaRTUN/ReSQ (Mirzaee & Kordjamshidi, 2022), World Bank indicators, WorldClim (Fick & Hijmans, 2017), Macrostrat, and GeoQuestions1089. Figure 5 shows broad but uneven coverage for the subset with explicit reference coordinates; examples without a meaningful reference coordinate are outside the map's scope.

Sampling is balanced using task-relevant factors such as scene class, temporal-pair type, coordinate mode, and geographic region. Reference answers are inherited from authoritative annotations when available or derived from the underlying geographic structure when the task requires an exact operation, including segmentation masks, road networks, geodesic distance, shortest paths, and country–year indicators. Each example is associated with its task identity, source provenance, reference answer, evaluation rule, and Bloom label, allowing the same benchmark to support capability-level, source-level, and cognitive-level analyses.

The answer contract follows task semantics rather than forcing one generic response type. References may be labels, free-form text, numerical values with units, coordinates, spatial relations, masks, graphs, routes, or polygons. *Strict* scoring measures exact correctness under the task's reference criterion, whereas *task-aware* scoring maps prediction quality to $[0, 1]$ using a leaf-appropriate measure such as entity $F_1$, mask overlap, graph agreement, route validity, or normalized numerical error. We first average within each leaf and then macro-average across all 23 leaves, so each capability contributes equally to the headline score. Invalid or unscorable responses remain in the denominator rather than being silently discarded.

### 3.3 Quality and Separation

Quality control is applied at both the benchmark and task levels. We verify the intended cardinality of every leaf and use task-specific checks for target ontologies, coordinate conventions, distance units, reference years, and image–mask consistency. Because every leaf contains 100 examples, the four domains contain 900, 600, 400, and 400 instances, and task-macro averages coincide with instance-weighted means. Source, grouping, task, and Bloom annotations further support controlled analyses within a source family or across matched task variants, making it possible to inspect whether apparent gains are concentrated in particular data regimes rather than treating the headline score as sufficient evidence.

Benchmark separation addresses a different concern: direct overlap between evaluation examples and the external evidence pool. GeoMapRAGCorpus excludes identifiable matches with GeoMapBench through shared source records, substantial textual overlap, exact indicator observations, and nearby geocoded items when location itself could reveal the answer. These safeguards reduce direct benchmark–corpus contamination but cannot rule out semantic paraphrases,

shared public facts, or information seen during model pretraining. We therefore describe the resource as *benchmark-separated*, not contamination-free, and retain provenance so that stronger overlap audits can be performed in future work.

## 4 GEOMAPRAGCORPUS

Many geospatial questions depend on evidence that is neither visible in the input nor appropriate to memorize in model parameters. GeoMapRAGCorpus provides a shared external evidence layer for such cases. We describe its source composition and modality coverage, then discuss benchmark separation and the design choice to keep the corpus independent of any particular retriever or augmentation pipeline.

**Composition.** GeoMapRAGCorpus contains 180,344 records drawn from seven public geospatial source families (Table 1). The sources were selected to cover complementary evidence rather than maximize raw scale: GeoNames and Wikidata contribute geographic entities and attributes; OpenStreetMap contributes map features and network structure; World Bank provides country-level indicators; EPSG/PROJ provides coordinate-reference information; Wikipedia supplies descriptive geographic text; and Wikimedia Commons contributes geocoded imagery. The resulting collection spans factual text, structured geographic records, map-related data, and visual evidence, reflecting the heterogeneous information that real geospatial questions may require.

| Source family | Primary evidence type | Records |
|---|---|---|
| GeoNames | Geographic entities and coordinates | 50,000 |
| Wikidata | Structured geographic facts | 48,275 |
| OpenStreetMap | Map features and networks | 36,692 |
| World Bank | Country–year indicators | 31,451 |
| EPSG/PROJ | Coordinate-reference information | 7,013 |
| Wikipedia | Descriptive geographic text | 6,251 |
| Wikimedia Commons | Geocoded imagery | 662 |
| Total | | 180,344 |

Table 1: Composition of GeoMapRAGCorpus. The source families provide complementary textual, structured, cartographic, and visual evidence.

**Evidence design.** The corpus is intentionally organized around evidence units rather than benchmark-style questions. A retrieved record should supply information that can support an answer—for example, a place attribute, an indicator value, a coordinate-system description, a map feature, or a geographically grounded image—without encoding the benchmark target itself. This distinction matters because many GeoMapBench leaves differ in what kind of external information is useful: factual and semantic tasks benefit from descriptive or structured evidence, whereas coordinate and network tasks may require retrieved geographic context together with exact computation. A multimodal evidence pool therefore provides a more appropriate substrate than a text-only knowledge base for studying augmented geospatial systems.

**Benchmark separation.** We construct the corpus with explicit safeguards against direct benchmark contamination. Records that can be matched to benchmark source items are removed, together with substantial text matches, exact benchmark-specific indicator observations, and nearby geocoded items in settings where location itself could reveal the target. These checks address identifiable overlap between the released benchmark and corpus, but they do not imply semantic independence from all public Web knowledge or from information a pretrained VLM may already have seen. We therefore describe GeoMapRAGCorpus as *benchmark-separated*, rather than making a stronger leakage-free claim.

**Retriever independence and provenance.** GeoMapRAGCorpus is released as a resource rather than as a fixed retrieval system. The same records can support semantic, spatial, multimodal, or hybrid retrieval, and can be combined with exact geographic tools when a task calls for com-

putation rather than factual lookup. This separation is important for controlled future comparisons: improvements in retrieval or tool use can be studied without changing the benchmark or replacing the underlying evidence collection. Each source family retains provenance and source-specific licensing information, which also makes retrieved evidence auditable and exposes coverage biases that would be hidden in an opaque retrieval backend.

Taken together, these design choices make the corpus useful for more than the single GeoMapPlus experiment in Section 5.2. It provides a fixed, inspectable evidence layer for studying when retrieval helps geospatial VLMs, when structured or visual evidence is preferable, and when augmentation should abstain because the original benchmark input already contains the information needed to answer.

# 5 EXPERIMENTS

Our experiments address two complementary questions: how well current VLMs cover the geospatial capability space without external assistance, and how much selective retrieval and geographic tools can change that picture. We therefore begin with a five-model evaluation on GeoMapBench and then compare direct answering with the GeoMapPlus augmented pipeline under the same benchmark protocol.

## 5.1 VLM EVALUATION

We evaluate GPT-5.4 mini (OpenAI, 2026), Gemini 3.5 Flash-Lite (Google DeepMind, 2026), Qwen3.8-27B (Qwen Team, 2026), Claude Sonnet 5 (Anthropic, 2026), and Llama 4 Scout (Meta AI, 2025) on all 2,300 examples under the same benchmark inputs and evaluator. This experiment measures the models' native geospatial capabilities: no corpus retrieval or external geographic tools are available. We compute task-aware and strict scores over the 100 examples in each leaf and then macro-average the 23 leaf scores; because all leaves are equally sized, these macro averages coincide with instance-weighted means. Model-specific generation settings are held fixed within each run and are reported with the released evaluation configuration.

Table 2 shows relatively close aggregate scores but clearly different capability profiles. Claude Sonnet 5 obtains the highest task-aware and strict macro scores, 47.90% and 37.83%, while Gemini 3.5 Flash-Lite reaches 46.50% and 37.09%. Gemini has the strongest visual-domain average in the table, whereas Claude is strongest on spatial, language, and applied tasks. The modest aggregate gaps are much smaller than many leaf-level differences, which motivates treating the benchmark as a capability profile rather than interpreting a single score as a complete measure of geospatial competence.

| Model | Task | Strict | Visual | Spatial | Language | Applied |
|---|---|---|---|---|---|---|
| GPT-5.4 mini | 42.11 | 34.22 | 45.9 | 53.1 | 34.0 | 25.2 |
| Gemini 3.5 Flash-Lite | <u>46.50</u> | <u>37.09</u> | **51.3** | <u>53.6</u> | <u>38.9</u> | <u>32.6</u> |
| Qwen3.8-27B | 42.04 | 34.30 | 45.6 | 52.8 | 34.9 | 25.1 |
| Claude Sonnet 5 | **47.90** | **37.83** | <u>48.1</u> | **59.9** | **44.5** | **32.9** |
| Llama 4 Scout | 37.35 | 30.17 | 43.7 | 41.9 | 28.2 | 25.5 |

Table 2: Complete benchmark results (%, higher is better). Task and Strict are macro-averaged over 23 equally sized leaves; domain columns are task-aware averages over the leaves in each domain. The best score in each column is **bold** and the second best is <u>underlined</u>.

The leaf matrix in Appendix G makes this heterogeneity explicit. Claude has the highest task-aware score on nine of 23 leaves, Gemini on six, GPT on three, Qwen on three, and Llama on two. GPT leads coordinate transformation (0.620) and dense land-cover labeling (0.481); Qwen leads map label-to-feature anchoring (0.627) and temporal scene matching (0.600); Gemini leads remote-sensing scene classification (0.710); and Claude leads shortest paths (0.442) and visual geolocation (0.700). By contrast, geologic or geomorphic interpretation remains at or below 0.127 for every model, and geographic fact reasoning reaches at most 0.160. The decomposition therefore reveals both task-dependent reversals and shared weak regions hidden by a single macro score.

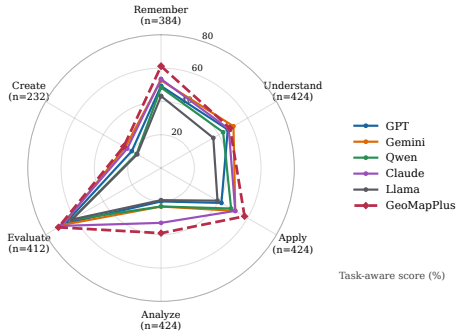



Figure 2: Task-aware scores by revised Bloom level for five direct VLMs and GeoMapPlus (Claude Sonnet 5 with retrieval and tools). GeoMapPlus is shown as a dashed line. Labels give the number of questions at each level; these descriptive means reflect different task and answer-format mixtures.

At the domain level, performance is higher on visual and spatial tasks than on applied geospatial questions for most models, although domain averages conceal large differences among their constituent leaves. The strongest visual and spatial averages reach 51.3% and 59.9%, respectively, whereas the best applied-task average is 32.9%. Many applied leaves require geographic knowledge, numeric interpretation, or evidence beyond what is directly visible. This contrast illustrates why domain and leaf scores are needed alongside the aggregate benchmark result.

Bloom profiles offer another diagnostic view. Claude's task-aware means range from 0.234 on Create and 0.329 on Analyze to 0.694 on Evaluate; Gemini's corresponding values are 0.249, 0.231, and 0.682. We do not interpret this ordering as intrinsic difficulty: each level contains a different mixture of leaves and output structures, and a judgment at Evaluate can be easier to score than constructing a graph at Create. Figure 2 shows these descriptive profiles for the five direct VLMs and GeoMapPlus; controlled within-leaf comparisons are needed to isolate cognitive demand.

### 5.2 GeoMapPlus

GeoMapPlus asks whether the released evidence resource can support a stronger geospatial system rather than only a static benchmark. Figure 3 summarizes the pipeline. A task router decides whether an example should use corpus retrieval, geographic tools, both, or neither. Retrieved evidence is filtered for relevance before being supplied alongside the original input, and the system can abstain from retrieval when additional context is unlikely to help. Geographic tools are reserved for tasks that benefit from exact operations such as coordinate transformation, distance computation, or graph reasoning. The final prediction is scored by the same benchmark evaluator used for direct answering.

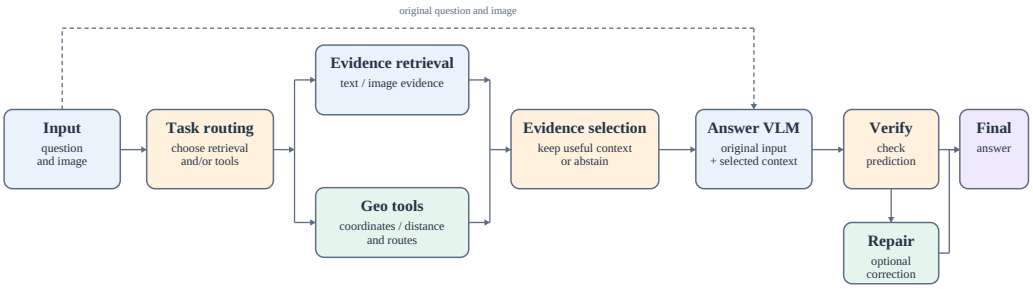



Figure 3: GeoMapPlus overview. Task routing determines whether to use corpus evidence and geographic tools; evidence selection can retain useful context or abstain before the VLM answers. Verification can accept the prediction or invoke optional repair, while the original question and image remain available to the answering model.

The comparison is paired at the example level. Direct Claude Sonnet 5 and GeoMapPlus receive the same 2,300 benchmark questions and images, use the same final answering model, and are evaluated by the same scoring procedure. The augmented condition additionally has access to GeoMapRAGCorpus, selective evidence retrieval, geographic tools, and answer verification. The resulting comparison therefore measures the *full augmented pipeline*; it should not be interpreted as isolating the causal contribution of the corpus, the retriever, or the tools individually. Component-wise ablations remain an important direction for future analysis.

Direct answering reaches 47.90% task-aware and 37.83% strict accuracy, versus 52.24% and 45.26% for GeoMapPlus: gains of 4.34 and 7.43 percentage points, with reported delta intervals of [3.09, 5.67] and [5.96, 8.91]. The paired comparison contains 395 wins, 207 losses, and 1,698 ties. The large number of ties reflects the selective design of the pipeline: retrieval is not forced on every example, and the augmented system can remain close to direct answering when the original visual input already contains the relevant evidence.

Selective intervention is important because irrelevant evidence can distract the final model. GeoMapPlus therefore retrieves only when external context or exact tools are plausibly useful and can abstain when the original input should dominate. The paired evaluation measures this behavior over the complete benchmark rather than only on a favorable subset.

Task-level effects are sharply nonuniform. Coordinate transformation rises from 0.541 to 1.000, metric distance from 0.642 to 1.000, population density from 0.694 to 0.891, and shortest paths from 0.442 to 0.609, consistent with benefits from exact tools, structured records, or checkable evidence. Ten leaves decline, including object presence and counting (0.340 to 0.235), topological and directional reasoning (0.567 to 0.460), and service areas (0.746 to 0.682); Appendix H contains all paired scores. Thus the aggregate gain should not obscure new failure modes introduced by context and intermediate stages. The case study supports the corpus as an experimental resource while motivating stronger routing, abstention, and component-level ablations.

## 6 Discussion and Conclusion

The experiments expose a consistent theme: geospatial capability is heterogeneous, and aggregate accuracy alone obscures where current VLMs succeed or fail. GeoMapBench shows that geospatial competence is not a single skill: models that perform strongly on visual recognition can remain weak on exact geometry, network reasoning, geospatial language, or applied geographic knowledge, and model rankings frequently reverse across leaves. The taxonomy therefore provides a more informative capability profile than a single benchmark score, while the Bloom annotations offer a complementary view of the cognitive operation requested by each example.

The results also show the value of preserving task-specific outputs and evaluators. Because GeoMapBench includes categorical, textual, numerical, geometric, and structured targets, errors can be localized to specific capabilities rather than collapsed into one generic response format. GeoMapRAGCorpus provides a shared multimodal evidence layer, and GeoMapPlus shows that selective retrieval and exact geographic tools can improve performance, especially on computation- and data-dependent tasks, while task-level variation suggests that augmentation should remain selective.

The Bloom annotations should be interpreted diagnostically rather than as a calibrated scale of intrinsic difficulty. GeoMapRAGCorpus is benchmark-separated rather than semantically independent of all public knowledge or model pretraining, and GeoMapPlus evaluates the full augmented pipeline rather than isolating each component. Public geographic data also raises considerations around coverage, redistribution terms, and potentially sensitive geolocated content, which require continued review as the resources are extended.

Taken together, the taxonomy, GeoMapBench, GeoMapRAGCorpus, and augmentation study provide a common foundation for evaluating geospatial VLMs across visual, linguistic, geometric, and knowledge-intensive settings. Future work can extend geographic and linguistic coverage, study retrieval and tool components more independently, and test whether new models improve consistently across the full capability space rather than only on aggregate scores.

### AI USE STATEMENT

Generative AI assisted with drafting and editing the manuscript. The authors are responsible for checking the scientific claims, citations, data, and final text.

### ETHICS STATEMENT

The benchmark uses publicly sourced geographic data; its provenance and source licenses accompany the records. We discuss geolocation, redistribution, and geographic coverage risks in Section 6. Before release, the authors will check source-specific permissions and any sensitive coordinates or photographs.

### REPRODUCIBILITY STATEMENT

We will release the benchmark, evaluation code, task definitions, source provenance, model prompts and configurations, and the GeoMapRAGCorpus resources needed to reproduce the reported experiments. Fixed benchmark splits and task-specific scoring rules will be provided so that future models can be evaluated under the same protocol. For information obtained from evolving public data sources, the release will also record retrieval dates and source versions whenever available.

## REFERENCES

Xiao An, Jiaxing Sun, Zihan Gui, and Wei He. Choice: Benchmarking the remote sensing capabilities of large vision-language models, 2025. URL https://arxiv.org/abs/2411.18145.

Lorin W. Anderson and David R. Krathwohl (eds.). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman, New York, 2001. ISBN 9780321084057.

Anthropic. Claude Sonnet 5 System Card, June 2026. URL https://www.anthropic.com/claude-sonnet-5-system-card.

Mutian Bao, Qiuyi Qi, Tian Liang, Jinjian Zhang, Wei Zhou, Ming Kong, Linjian Mo, and Qiang Zhu. UrbanGeoEval: A city-scale benchmark for evaluating large language models in geospatial reasoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40183–40223, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1867. URL https://aclanthology.org/2026.acl-long.1867/.

Sharat Bhat, Harshita Khandelwal, Tushar Kataria, and Vivek Gupta. Mapverse: A benchmark for geospatial question answering on diverse real-world maps, 2026. URL https://arxiv.org/abs/2602.10518.

Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. Geobench-vlm: Benchmarking vision-language models for geospatial tasks, 2025. URL https://arxiv.org/abs/2411.19325.

Stephen E. Fick and Robert J. Hijmans. WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12):4302–4315, 2017. doi: 10.1002/joc.5086.

Google DeepMind. Gemini 3.5 Flash-Lite model card, July 2026. URL https://deepmind.google/models/model-cards/gemini-3-5-flash-lite/.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification, 2019. URL https://arxiv.org/abs/1709.00029.

Yangyu Huang, Tianyi Gao, Haoran Xu, Qihao Zhao, Yang Song, Zhipeng Gui, Tengchao Lv, Hao Chen, Lei Cui, Scarlett Li, and Furu Wei. Peace: Empowering geologic map holistic understanding with mllms, 2025. URL https://arxiv.org/abs/2501.06184.

Sergios-Anestis Kefalidis, Dharmen Punjani, Eleni Tsalapati, Konstantinos Plas, Mariangela Pollali, Michail Mitsios, Myrto Tsokanaridou, Manolis Koubarakis, and Pierre Maret. Benchmarking geospatial question answering engines using the dataset GeoQuestions1089. In *The Semantic Web – ISWC 2023*, pp. 266–284. Springer, 2023. doi: 10.1007/978-3-031-47243-5_15.

David R. Krathwohl. A revision of bloom's taxonomy: An overview. *Theory Into Practice*, 41(4): 212–218, 2002. doi: 10.1207/s15430421tip4104_2.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27831–27840, June 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Zekun Li, Malcolm Grossman, Eric (Ehsan) Qasemi, Mihir Kulkarni, Muhao Chen, and Yao-Yi Chiang. MapQA: Open-domain geospatial question answering on map data, 2025. URL https://arxiv.org/abs/2503.07871.

Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566, 2020. doi: 10.1109/TGRS.2020.2988782.

Meta AI. Llama 4 Model Card, April 2025. URL https://github.com/meta-llama/llama-models/blob/main/models/llama4/MODEL_CARD.md.

Roshanak Mirzaee and Parisa Kordjamshidi. Transfer learning with synthetic corpora for spatial role labeling and reasoning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6148–6165, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.413. URL https://aclanthology.org/2022.emnlp-main.413/.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. SPARTQA: A textual question answering benchmark for spatial reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4582–4598. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.364. URL https://aclanthology.org/2021.naacl-main.364/.

Srija Mukhopadhyay, Abhishek Rajgaria, Prerana Khatiwada, Manish Shrivastava, Dan Roth, and Vivek Gupta. MAPWise: Evaluating vision-language models for advanced map queries. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9348–9378, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.473. URL https://aclanthology.org/2025.naacl-long.473/.

OpenAI. Introducing GPT-5.4 mini and nano, March 2026. URL https://openai.com/index/introducing-gpt-5-4-mini-and-nano/.

Qwen Team. Qwen3.8-27B. Hugging Face model card, August 2026. URL https://huggingface.co/Qwen/Qwen3.8-27B.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 11321–11329, 2022. doi: 10.1609/aaai.v36i10.21383. URL https://ojs.aaai.org/index.php/AAAI/article/view/21383.

Sagar Soni, Akshay Dudhane, Hiyam Debary, Mustansar Fiaz, Muhammad Akhtar Munir, Muhammad Sohail Danish, Paolo Fraccaro, Campbell D. Watson, Levente J. Klein, Fahad Shahbaz Khan, and Salman Khan. Earthdial: Turning multi-sensory earth observations to interactive dialogues, 2024. URL https://arxiv.org/abs/2412.15190.

Huy Quang Ung, Guillaume Habault, Yasutaka Nishimura, Hao Niu, Roberto Legaspi, Tomoki Oya, Ryoichi Kojima, Masato Taya, Chihiro Ono, Atsunori Minamikawa, and Yan Liu. Cartomapqa: A fundamental benchmark dataset evaluating vision-language models on cartographic map understanding, 2025. URL https://arxiv.org/abs/2512.03558.

Fengxiang Wang, Hongzhen Wang, Zonghao Guo, Di Wang, Yulin Wang, Mingshuo Chen, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, Zhiyuan Liu, and Maosong Sun. Xlrs-bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14325–14336, 2025. doi: 10.1109/CVPR52734.2025.01336.

Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for global high-resolution land cover mapping, 2022. URL https://arxiv.org/abs/2210.10732.

Haotian Xu, Yue Hu, Zhengqiu Zhu, Chen Gao, Ziyou Wang, Junreng Rao, Wenhao Lu, Weishi Li, Quanjun Yin, and Yong Li. CityCube: Benchmarking cross-view spatial reasoning on vision-language models in urban environments. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9190–9215, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.416. URL https://aclanthology.org/2026.acl-long.416/.

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*, 2023. URL https://arxiv.org/abs/2210.03629.

Dazhou Yu, Riyang Bao, Ruiyu Ning, Jinghong Peng, Gengchen Mai, and Liang Zhao. Spatial-RAG: Spatial retrieval augmented generation for real-world geospatial reasoning questions. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 11094–11112, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.539. URL https://aclanthology.org/2026.findings-acl.539/.

# A BENCHMARK COMPOSITION

This appendix summarizes the domain composition of GeoMapBench and the assignment of revised-Bloom variants across benchmark examples.

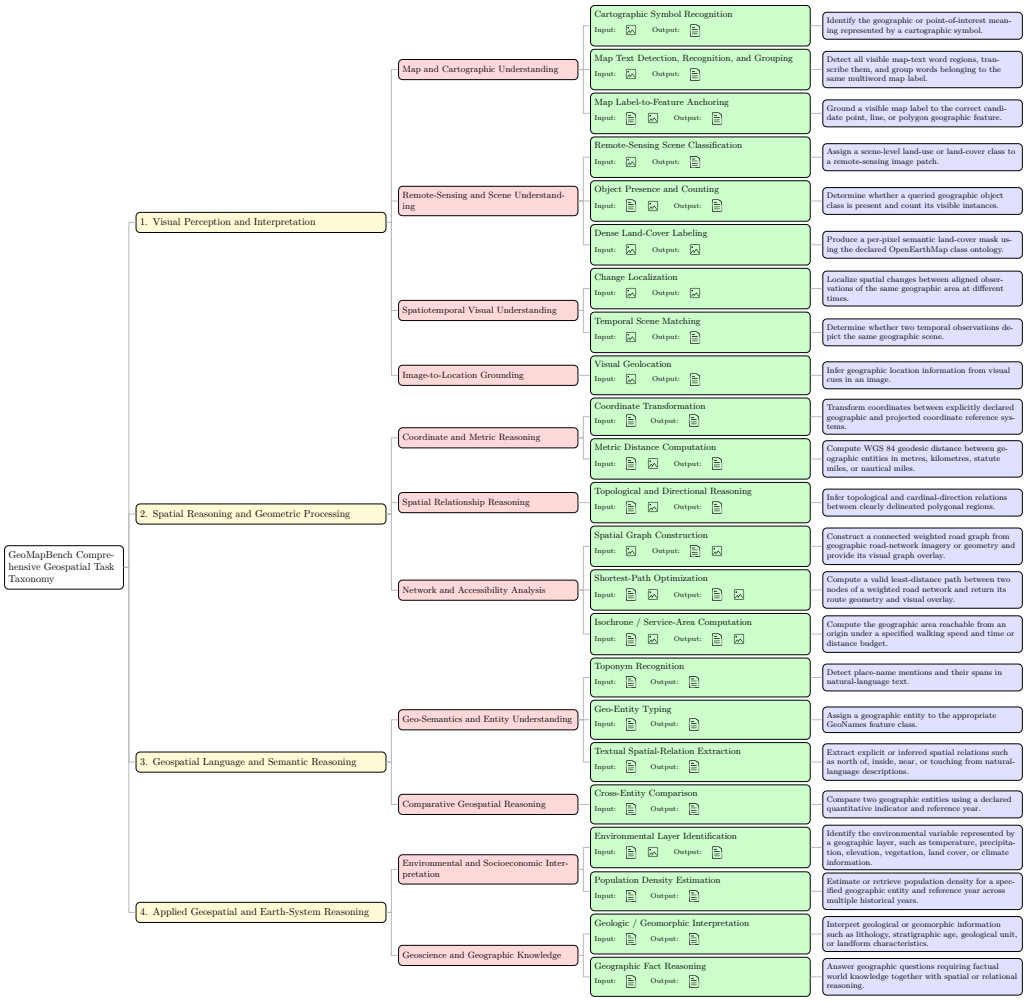



Figure 4: Detailed GeoMapBench geospatial capability taxonomy, including the four domains, 23 fine-grained tasks, input/output modalities, and task descriptions. The compact version used in the main paper is shown in Figure 1.

| Domain | Tasks | Examples |
|---|---|---|
| Visual Perception and Interpretation | 9 | 900 |
| Spatial Reasoning and Geometric Processing | 6 | 600 |
| Geospatial Language and Semantic Reasoning | 4 | 400 |
| Applied Geospatial and Earth-System Reasoning | 4 | 400 |
| Total | 23 | 2,300 |

Table 3: GeoMapBench composition by taxonomy domain. Each task contains 100 examples.

Each task has 100 examples. The full hierarchy and input/output signatures are visible in Figure 4. Source provenance and grouping annotations are retained to support auditing and controlled subset analyses. The Bloom balance is defined within each task over its supported levels: 20 examples per level when five levels are supported, and 17/17/17/17/16/16 when six are supported.

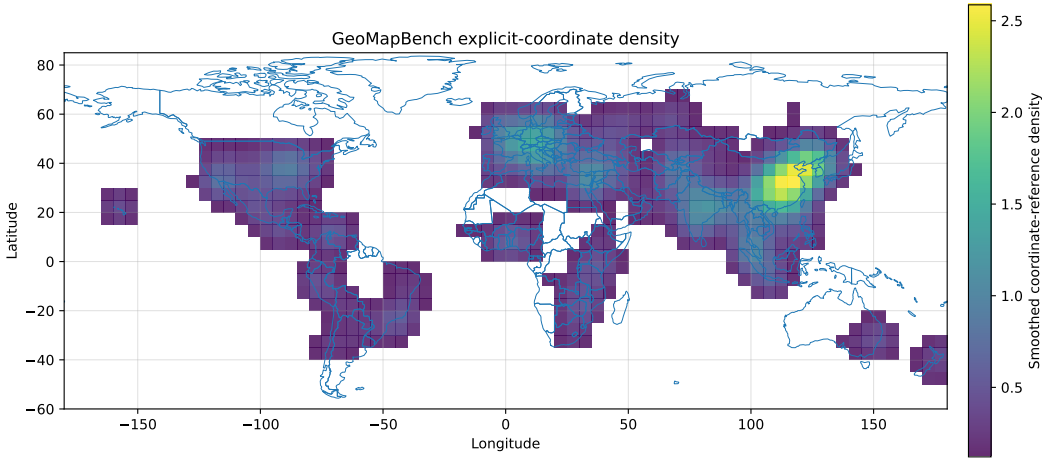



Figure 5: Smoothed geographic density of GeoMapBench examples with explicit reference coordinates. The map illustrates regional diversity and uneven concentration for this coordinate-bearing subset; examples without explicit coordinates are excluded. Color indicates smoothed density rather than a count of examples per cell.

### A.1 Bloom Variants

The revised Bloom taxonomy distinguishes a *knowledge dimension* from a *cognitive-process dimension* (Anderson & Krathwohl, 2001; Krathwohl, 2002). GeoMapBench operationalizes only the latter: Remember, Understand, Apply, Analyze, Evaluate, and Create. These labels characterize the operation requested by a prompt, not an observation of a model's internal reasoning or a universally ordered difficulty scale. Table 4 illustrates the six variants defined for the polygon-relation leaf; the descriptions paraphrase its implemented prompts.

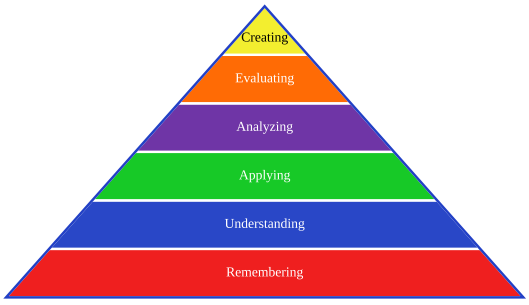



Figure 6: The six cognitive-process levels of the revised Bloom taxonomy used in GeoMapBench.

| Level | Requested operation | Example for polygons A and B | Total |
|---|---|---|---|
| Remember | Recognize | Name the A-to-B relation. | 384 |
| Understand | Interpret | Classify the relation as topological or directional. | 424 |
| Apply | Execute | Determine the relation between the given polygons. | 424 |
| Analyze | Decompose | Give the A-to-B relation and its B-to-A inverse. | 424 |
| Evaluate | Check | Judge whether a proposed relation is correct. | 412 |
| Create | Construct | Produce a two-node graph with both directed relations. | 232 |
| | | All GeoMapBench examples | 2,300 |

Table 4: Bloom operations illustrated with variants of the topological and directional reasoning task. Totals count assigned levels over the entire benchmark, not just this task.

For each leaf, we determine in advance which Bloom levels admit a meaningful and objectively scorable variant. Each of its 100 source examples is assigned exactly one primary level, and the prompt and reference answer are adapted to request the corresponding operation while preserving the underlying geospatial content. Levels without a defensible formulation are omitted for that leaf: for example, change localization has no Remember variant, while the relation task illustrated above supports all six levels. Consequently, the global level totals are unequal and each Bloom level contains a different mixture of tasks and response types. A correct/incorrect judgment at Evaluate may therefore be easier for a VLM than generating a full graph at Create; comparisons across levels should be read as diagnostic descriptions, not causal estimates of intrinsic cognitive difficulty.

## B Task Sources and Evaluation Contracts

Table 5 makes the operational scope of the 23 leaves explicit. The output and diagnostic columns describe representative task instances; Bloom variants can request a different operation or response structure within the same leaf. All task-aware instance scores lie in $[0, 1]$, and each leaf contributes equally to the macro-average. Source names and contracts are taken from the frozen benchmark source mapping and evaluation implementation.

Table 5: Sources, typical output types, and task-aware diagnostics for GeoMap-Bench. The original input and response contract may vary across Bloom variants.

| Task | Primary source | Typical output | Principal diagnostic |
|---|---|---|---|
| Cartographic symbol recognition | Maki | Symbol class | Accuracy |
| Map text detection, recognition, grouping | Paris/Jerusalem maps | Text and groups | Text/structure agreement |
| Map label-to-feature anchoring | OpenStreetMap | Feature anchor | Field/geometry agreement |
| Dense land-cover labeling | OpenEarthMap | Class-index mask | Mask mean IoU |
| Remote-sensing scene classification | EuroSAT | Scene class | Accuracy |
| Object presence and counting | RSVQA-LR | Presence or count | Count error/accuracy |
| Change localization | SpaceNet 7 | Change mask/IDs | Mask IoU and object $F_1$ |
| Temporal scene matching | SpaceNet 7 | Match decision | Accuracy/field agreement |
| Visual geolocation | Wikimedia Commons | City/country | Location-field accuracy |
| Coordinate transformation | Natural Earth; PROJ/EPSG | Coordinates | Coordinate agreement/error |
| Metric distance computation | Natural Earth | Distance with units | Relative numeric error |
| Topological/directional reasoning | Natural Earth | Spatial relations | Relation $F_1$ |
| Spatial graph construction | SpaceNet 3 | Road graph | Node/edge $F_1$ |
| Shortest-path optimization | SpaceNet 3 | Route | Validity and relative cost |
| Isochrone/service area | OpenStreetMap | Reachable polygon | Polygon IoU |
| Toponym recognition | GeoWebNews | Place-name spans | Entity $F_1$ |
| Geo-entity typing | GeoNames | Feature class | Accuracy |
| Textual spatial relation extraction | SpaRTUN/ReSQ | Relation tuples | Relation $F_1$ |
| Cross-entity comparison | World Bank | Comparison/values | Field/numeric agreement |
| Environmental layer identification | WorldClim | Layer class | Accuracy |
| Population density estimation | World Bank | Density/ordering | Numeric/rank agreement |
| Geologic/geomorphic interpretation | Macrostrat | Structured description | Field agreement |
| Geographic fact reasoning | GeoQuestions1089 | Geographic answer | Answer agreement |

## C Scope Relative to Existing Evaluations

Table 6 compares the primary evaluation settings of the closest benchmark families discussed in Section 2. It characterizes their published emphasis rather than claiming that a benchmark cannot contain any example outside the listed setting. In particular, CHOICE and GeoMapBench both have 23 leaves but organize different capability spaces.

| Benchmark | Principal input setting | Emphasis and typical response |
|---|---|---|
| MAPWise (Mukhopadhyay et al., 2025) | Choropleth maps | Questions about visual encodings and spatial relations |
| GEOBench-VLM (Danish et al., 2025) | Earth-observation imagery | Perception, localization, segmentation, and temporal analysis |
| CHOICE (An et al., 2025) | Remote-sensing imagery | Hierarchical perception/reasoning; 23 leaf tasks and multiple-choice problems |
| GeoMapBench | Maps, Earth-observation imagery, text, and geographic data | Four domains; categorical, numerical, geometric, and structured outputs; benchmark-separated retrieval corpus |

Table 6: Primary scope of representative geospatial VLM evaluations. Rows summarize the cited work's main evaluation setting; task overlap does not imply identical prompts or scoring.

## D Frozen-Release Integrity Checks

The frozen benchmark's automated Bloom audit reports 23 leaves and 2,300 records, with 100 records per leaf, the planned five- or six-level allocation within each leaf, and no structural audit errors. The corpus build manifest records 2,722 duplicate candidate texts skipped before finalization. Its quality report validates all 180,344 final records and reports zero duplicate identifiers, zero duplicate normalized texts, zero missing media, and zero invalid images. These checks establish structural integrity and reproducibility of the frozen artifacts; they are distinct from an independent human assessment of question or evidence quality.

Corpus modalities comprise 172,299 structured records, 6,251 text records, 1,132 map images, and 662 geocoded images. The released normalized inner-product indexes contain 180,344 text vectors (384-dimensional BGE-small-en-v1.5) and 1,794 image vectors (512-dimensional CLIP ViT-B/32), matching the two image-bearing modality counts. The benchmark–corpus exclusion guard matches source identifiers, normalized text hashes, exact country–indicator–year observations, and records within 2 km of benchmark coordinates where location could reveal the answer. Release metadata includes file checksums and the benchmark's Bloom audit. These are reproducible overlap checks, not claims about semantic paraphrases or model pretraining.

## E TAXONOMY-LEVEL PERFORMANCE PROFILES

Figures 7 and 8 organize the same task-aware results by the four domains and 23 leaves of the taxonomy. Every leaf contributes equally to its domain mean. The five direct VLMs use all 2,300 question-level scores each; GeoMapPlus uses its paired scores on the same 100 questions per leaf. The dashed GeoMapPlus curve represents the complete augmented pipeline rather than the corpus alone. Axes show percentages, and the abbreviated leaf labels follow the task order of Table 5.

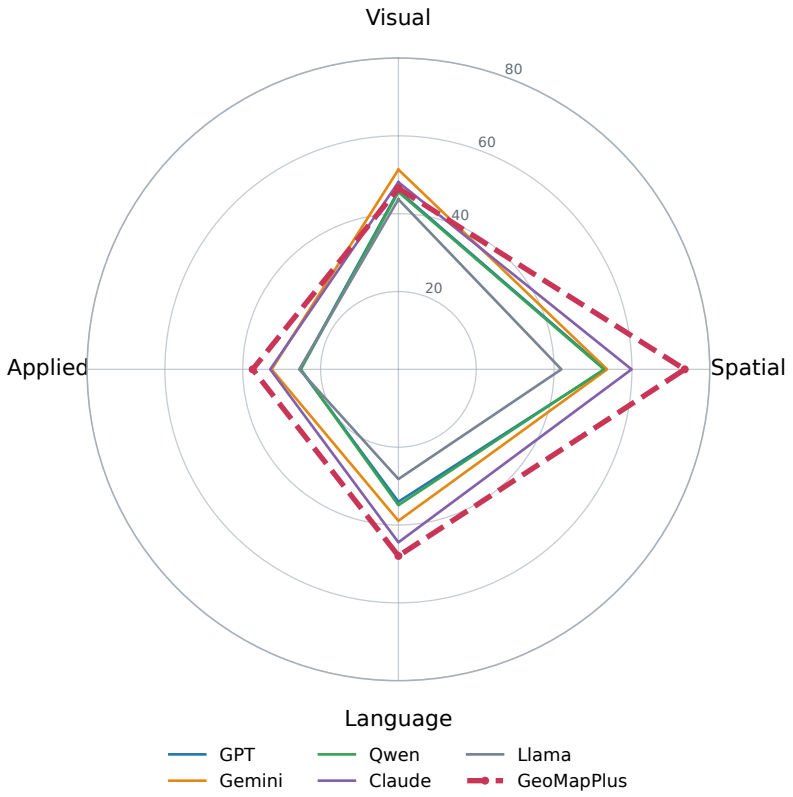



Figure 7: Task-aware score by taxonomy domain (%, four axes). Each domain is the unweighted mean of its constituent leaf scores; the model conditions use the same complete benchmark. The radial scale is 0–80%.

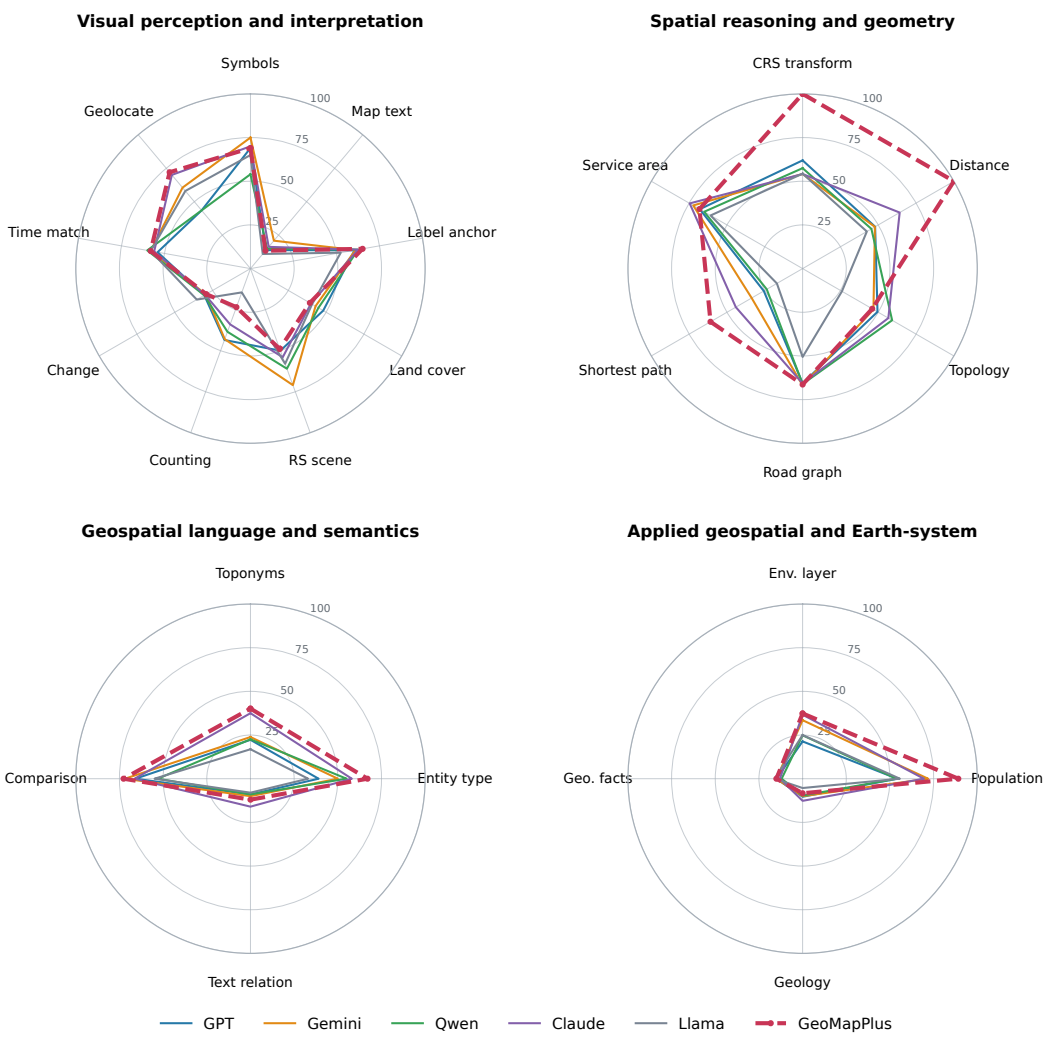



Figure 8: Task-aware score (%) on all 23 leaves, grouped by taxonomy domain. Each leaf contains 100 examples in every condition. The common radial scale is 0–100%; names are shortened only for display and follow Table 5.

## F Paired Aggregate Uncertainty

We resample the same question IDs for the two leading direct models within each of the 23 leaves (5,000 paired bootstrap replicates) and macro-average the resulting leaf scores. Claude Sonnet 5 exceeds Gemini 3.5 Flash-Lite by 1.40 percentage points in task-aware score (95% percentile interval $[-0.22, 2.97]$) and 0.74 points in strict score (interval $[-0.83, 2.35]$). Both intervals include zero, so their observed aggregate order does not establish a reliable separation under this resampling procedure. This paired calculation complements the per-model intervals in Figure 9.

## G Additional Results

This appendix provides aggregate and task-level views of the five-model evaluation beyond the summary table in the main paper.

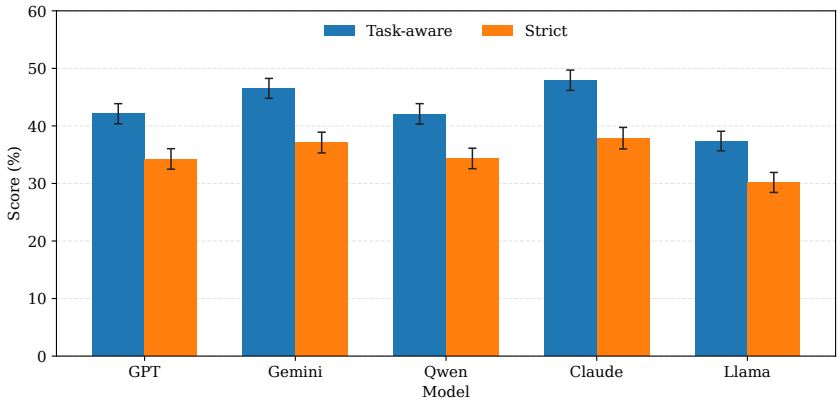



Figure 9: Aggregate task-aware score and strict accuracy for the five complete model runs. Whiskers show 95% percentile intervals from 5,000 bootstrap resamples of questions within each of the 23 equally weighted tasks. All models share the same resampled question IDs.

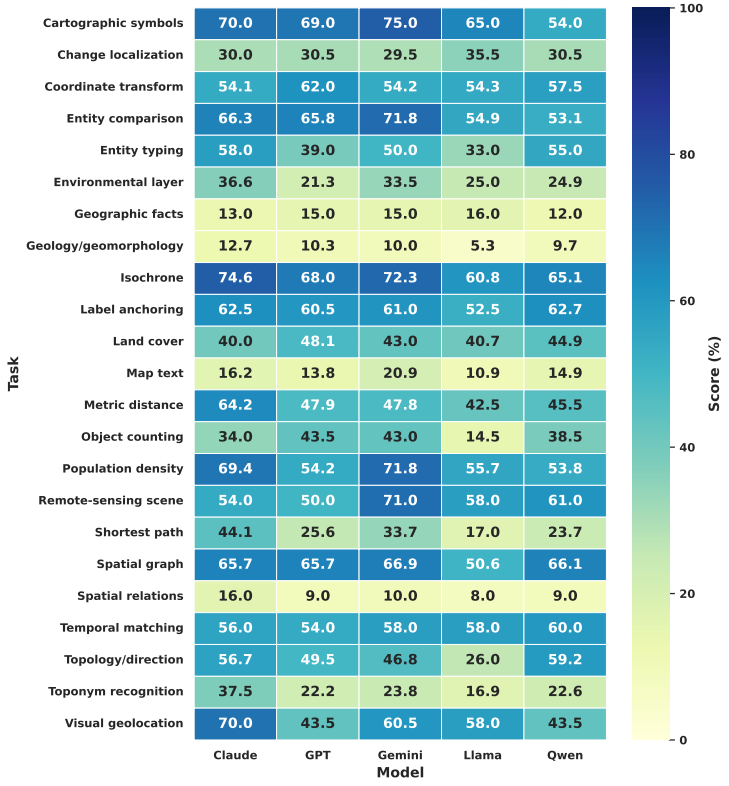



Figure 10: Task-aware score on each of the 23 benchmark leaves. All five models are evaluated on all 100 examples of each leaf.

## H GeoMapPlus by Task

This appendix reports the paired task-aware scores for direct answering and the full GeoMapPlus pipeline on each benchmark leaf.

| Task | Direct | GeoMapPlus | Δ (pp) |
|---|---|---|---|
| Cartographic symbol recognition | **0.700** | 0.690 | -1.0 |
| Change localization | **0.300** | 0.292 | -0.8 |
| Coordinate transformation | 0.541 | **1.000** | +45.9 |
| Cross entity comparison | 0.663 | **0.727** | +6.3 |
| Dense land cover labeling | **0.400** | 0.393 | -0.7 |
| Environmental layer identification | 0.366 | **0.373** | +0.7 |
| Geo entity typing | 0.580 | **0.670** | +9.0 |
| Geographic fact reasoning | 0.130 | **0.150** | +2.0 |
| Geologic geomorphic interpretation | **0.127** | 0.083 | -4.3 |
| Isochrone service area | **0.746** | 0.682 | -6.5 |
| Map label feature anchoring | 0.625 | **0.652** | +2.7 |
| Map text detection recognition grouping | **0.162** | 0.134 | -2.9 |
| Metric distance computation | 0.642 | **1.000** | +35.8 |
| Object presence counting | **0.340** | 0.235 | -10.5 |
| Population density estimation | 0.694 | **0.891** | +19.7 |
| Remote sensing scene classification | **0.540** | 0.490 | -5.0 |
| Shortest path optimization | 0.442 | **0.609** | +16.7 |
| Spatial graph construction | 0.657 | **0.664** | +0.6 |
| Temporal scene matching | 0.560 | **0.580** | +2.0 |
| Textual spatial relation extraction | **0.160** | 0.120 | -4.0 |
| Topological directional reasoning | **0.567** | 0.460 | -10.7 |
| Toponym recognition | 0.375 | **0.401** | +2.6 |
| Visual geolocation | 0.700 | **0.720** | +2.0 |

Table 7: Task-aware score on the same 100 examples per leaf for direct Claude Sonnet 5 and the full GeoMapPlus pipeline. Deltas are percentage points. The better score in each row is **bold** and the second best is underlined.