

MultiGlobeQA: A Multilingual and Globally Diverse Benchmark for Geospatial Reasoning

Martin Böckling*, Elizaveta Nosova, Heiko Paulheim and Andreea Iana*

Data and Web Science Group, University of Mannheim, Germany

{martin.boeckling, heiko.paulheim, andreea.iana}@uni-mannheim.de
elizaveta.nosova@students.uni-mannheim.de

Abstract

Geospatial reasoning, i.e., computing distances, containment, and other spatial relations over real-world entities, is central to navigation and logistics, yet large language models (LLMs) struggle with the required geometric and topological computation despite storing considerable geographic knowledge. Existing benchmarks localize these failures only partially: they are synthetic or small-scale, largely monolingual, and offer limited control over geographic coverage. We introduce MultiGlobeQA, a multilingual benchmark of 46,060 question-answer pairs spanning 14 spatial-function families and 15 answer formats, with execution-based ground truth over three knowledge graphs. It covers 201 countries and territories via income- and density-stratified sampling, with parallel questions in English and 16 additional high- and low-resource languages. Across parametric, reasoning, and agentic settings, LLMs collapse on tasks requiring grid indexing and shape computation, while topological relations and directions fare best. Retrieval and tool use yield considerable gains, yet performance plateaus below two thirds even when gold facts are supplied, indicating that computation, not access to knowledge, is the bottleneck. Models also underperform on low-income regions, a gap that gold facts widen rather than close.

1 Introduction

Large language models (LLMs) are increasingly deployed in navigation, logistics, and planning systems that require geospatial reasoning (inferring distances, containment, and other spatial relations over geographic entities) to answer questions about physical space, directly or through retrieval-augmented or agentic workflows (Mai et al., 2024; Xie et al., 2024; Dihan et al., 2025; Yu et al., 2026; Dorobantu and Badea, 2026). While LLMs encode

substantial geographic knowledge, such as coordinate representations (Gurnee and Tegmark, 2024) and parametric factual information (Roberts et al., 2023; Bhandari et al., 2023), geospatial reasoning entails computation over geometric and topological relationships (distances, directions, coordinate transformations), a capability with which they struggle (Dihan et al., 2025; Li et al., 2025a; Truong et al., 2026). Fig. 1 illustrates this gap. A model underestimates the distance between two towns by an order of magnitude and still fails to compute a geohash even when the relevant facts are provided (top). Across question categories (bottom), explicit reasoning yields little improvement over direct answering, whereas geographic facts paired with computation tools bring substantial but uneven gains: metric and topological questions reach 65%, direction, shape, and uncertainty 63%, and centrality, grid indexing and transformation only 50%. Geographic information alone is thus insufficient for reliable geospatial reasoning.

Yet existing benchmarks capture these failures only partially. Early spatial-reasoning benchmarks are largely synthetic, testing qualitative relations over abstract objects rather than geographic entities (Mirzaee et al., 2021; Shi et al., 2022), while those grounded in real geography are small-scale (Kefalidis et al., 2023; Li et al., 2025b; Saeedan et al., 2026) or monolingual, with limited control over coverage (Dihan et al., 2025; Li et al., 2025a; Bao et al., 2026). Many also sample directly from OpenStreetMap (OSM) and Wikidata (Vrandečić and Krötzsch, 2014), inheriting their uneven coverage (Herfort et al., 2023) and socio-economic biases (Das et al., 2025). Consequently, no benchmark localizes failures jointly across spatial operations, answer formats, regions, and languages.

Contributions. We address these gaps with MultiGlobeQA, an open, large-scale multilingual benchmark for geospatial reasoning.

* Equal contribution.

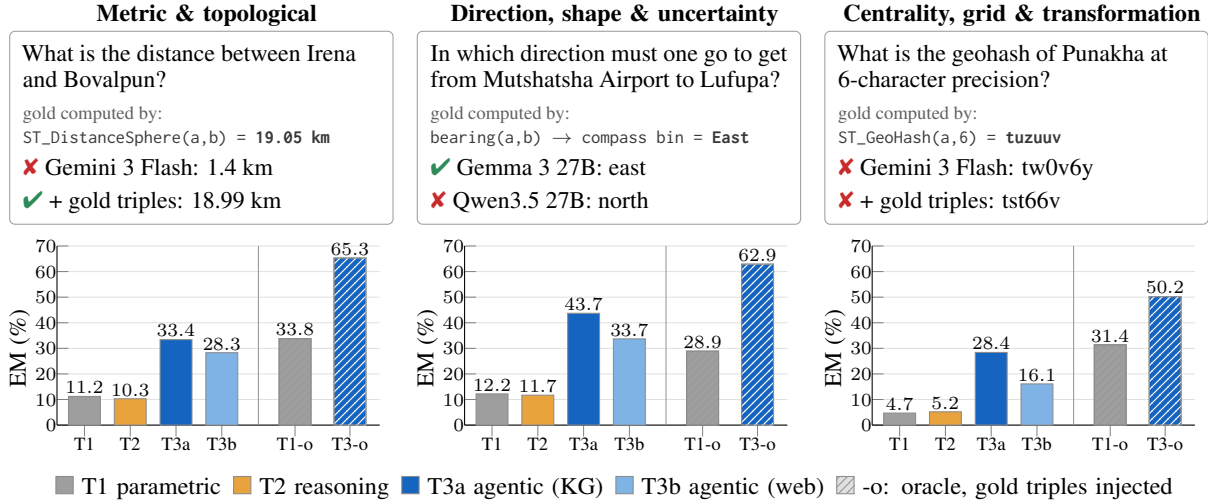


Figure 1: **LLMs systematically fail on questions requiring geospatial reasoning and computation.** *Top*: three example items (distance, direction, geohash), each with its executable gold answer and the models’ predictions (✓ correct, ✗ wrong). + *gold triples* denotes an oracle setting with perfect retrieval, where the relevant facts are injected into the prompt. *Bottom*: exact match (%) averaged over four LLMs per category, across evaluation tiers: parametric (T1), reasoning (T2), and agentic retrieval over the KG (T3a) or the web (T3b), both with code execution; hatched -o bars are the corresponding oracle conditions, with gold triples injected and retrieval disabled.

MultiGlobeQA comprises **46,060 QA pairs** instantiated from 65 templates across 14 spatial-function families, false-premise questions and a multimodal slice, with *execution-based* ground truth over three geographic knowledge graphs (KGs), so every answer is verified by construction. Compared to existing benchmarks, MultiGlobeQA is: (1) *broader* – it spans 14 spatial functions and 15 answer formats, from Boolean and set enumeration to grid-cell encodings and geometries; (2) *geographically diverse* – entities are sampled under income- and density-stratification across 201 countries and territories in four income tiers; (3) *multi-parallel* – the same questions are released in 17 high- and low-resource languages with human-verified translations. These support fine-grained analysis of failure types, geographic bias, and cross-lingual comparison.

We evaluate LLMs across parametric, reasoning, and agentic tiers, with oracle conditions injecting gold KG triples to approximate perfect retrieval. Accuracy peaks below two thirds and collapses on grid indexing and shape, which stay below 30% even with gold facts and tools, while coordinate questions reach 94%. Retrieval and tool use yield sizable gains, whereas explicit reasoning provides little benefit. Live retrieval recovers only half to three quarters of perfect retrieval, and the rest of the agentic budget goes to search that never reaches the evidence. Finally, our stratified design reveals that low-income regions lag under parametric knowl-

edge and under perfect retrieval, while performance remains largely stable across languages.

2 Related Work

Spatial and Geospatial Reasoning Benchmarks.

Textual spatial-reasoning benchmarks, such as bAbI tasks 17/19 (Weston et al., 2015), SPARTQA (Mirzaee et al., 2021), SpaRTUN (Mirzaee and Kordjamshidi, 2022), and StepGame (Shi et al., 2022), are largely synthetic, evaluating reasoning over a fixed set of spatial relations in toy worlds and leaving open how models reason about real-world geography. To ground reasoning in real geographic entities, one line studies question answering (QA) over geographic KGs (Mai et al., 2021). Extending template-based QA over linked geospatial data (Punjani et al., 2018), GeoQuestions1089 (Kefalidis et al., 2023, 2024) pairs 1,089 manually authored questions with executable GeoSPARQL queries over the union of YAGO2 (Hoffart et al., 2013) and YAGO2geo (Karalis et al., 2019), targeting primarily factual and relational retrieval.

A second line directly evaluates foundation models: LLMs encode geographic coordinates (Gurnee and Tegmark, 2024) and recall facts without retrieval (Roberts et al., 2023; Bhandari et al., 2023), yet they struggle to compute over them (Truong et al., 2026; Li et al., 2025a), particularly for questions requiring distance, direction, and counting (Dihan et al., 2025). Retrieval-, tool-, and agent-

augmented systems offload computations to external resources or GIS pipelines, which improves accuracy (Dihan et al., 2025; Yu et al., 2026; Krechetova and Kochedykov, 2025; Hasan et al., 2026; Zhang et al., 2025), but conflates tool assistance with the model’s reasoning ability (Bao et al., 2026). Suizu et al. (2026), closest to our setting yet single-country, automatically generate geospatial questions by composing spatial and entity constraints, likewise finding that models ground entities well but fail at precise spatial reasoning.

Execution-Based Benchmark Construction. Benchmarks increasingly obtain verifiable ground truth by executing structured queries over a knowledge source rather than matching surface forms, the standard in text-to-SQL, where correctness is measured by execution accuracy (Yu et al., 2018; Li et al., 2023). The same recipe underlies non-spatial KG and graph QA benchmarks such as CRAG (Yang et al., 2024), STaRK (Wu et al., 2024), and GRBench (Jin et al., 2024), which target retrieval and general graph reasoning rather than spatial computation, as well as geospatial QA like MapQA (Li et al., 2025b) and GS-QA (Saeedan et al., 2026), which compute answers via spatial SQL over OSM.

Limitations of Current Benchmarks. Existing benchmarks are limited along several dimensions. They are small and manually authored, cover a narrow range of spatial operations, and are city- or region-specific and nearly exclusively monolingual. They also inherit the high-income, urban skew of their OSM- and Wikidata-derived data (Herfort et al., 2023; Das et al., 2025), a bias that propagates into model behavior across regions and languages (Manvi et al., 2024; Moayeri et al., 2024; Faisal and Anastasopoulos, 2023). MultiGlobeQA addresses these gaps jointly: it derives large-scale, execution-verified, multilingual questions over a wide range of spatial operations and three independently built geographic KGs, sampled under income- and density-stratification for diverse geographic coverage, extending multilingual geographic QA (Roh et al., 2025; Hwang et al., 2025) from cross-lingual knowledge to spatial reasoning.

3 MultiGlobeQA

We construct MultiGlobeQA in four stages: we define a taxonomy of typed templates (§3.1), sample entities from three KGs (§3.2), instantiate each template with an executor that produces verified English QA pairs (§3.3), and translate the templates

Spatial Operation (#TIDs)	Spatial functions (SFIDs)
Metric & topological (23)	(A) distance, (B) containment, (C) topology
Network & path (4)	(D) network/path
Hierarchy & comparison (9)	(E) administrative hierarchy, (F) comparison
Centrality, grid & transformation (10)	(G) centrality, (H) grid, (M) transformation, (N) coord.
Direction, shape & uncertainty (14)	(J) direction, (K) shape, (L) uncertainty
Spatio-temporal events (5)	(I) spatio-temporal events

Table 1: **MultiGlobeQA’s template taxonomy.** #TIDs counts top-level templates.

into 16 target languages, re-instantiating them with multilingual entity labels (§3.5).

3.1 Template Taxonomy

We construct MultiGlobeQA around a taxonomy of 65 top-level templates (TIDs), organized into 14 spatial-function categories (SFIDs). Table 1 summarizes the SFID coverage across six clusters of common geospatial reasoning operations (Egenhofer and Franzosa, 1991; Randell et al., 1992; Cohn and Hazarika, 2001; Worboys and Duckham, 2004). Each TID expands into one or more parameterized sub-templates (129 in total). To increase linguistic diversity, we generate 1-5 LLM-drafted, human-verified paraphrases per sub-template (315 English natural-language variants in total). Templates use typed placeholders for entities, values, and parameters, and support 15 answer formats (Boolean, numeric, geometric, categorical, and temporal). Table 5 lists all templates with examples.

3.2 Stratified Entity Sampling

Data Sources. We sample entities from three complementary KGs: (i) *WorldKG* (Dsouza et al., 2021), built from tagged OSM nodes, contributing named places and amenities as points with the broadest country coverage; (ii) *KnowWhereGraph* (Janowicz et al., 2022), a cross-domain KG integrating hazard, health, and administrative hierarchies from heterogeneous, largely US-based sources; and (iii) *OSMH3KG* (Böckling et al., 2024), an H3-indexed KG spanning the full range of OSM geometries, from POIs to road networks and administrative polygons. We convert each KG into a common entity table of WKT geometries, canonical names, semantic categories mapped from `rdf:type`, and H3 indices (Brodsky, 2018). We resolve ambiguous names (e.g., chain store names), by appending contextual properties such as neighborhood or street name, and coordinates if needed.

Strata Grid. To mitigate the over-representation of high-income and urban regions in OSM- and Wikidata-derived KGs (Herfort et al., 2023; Das et al., 2025), we assign each entity two strat-

Income tier	Density tier			#Questions	#Countries
	low	medium	high		
HIC	5,521	6,009	6,713	18,243 (40%)	76
UMC	1,156	4,012	4,194	9,362 (20%)	56
LMC	2,446	4,095	3,699	10,240 (22%)	44
LIC	502	3,858	3,855	8,215 (18%)	25
Total	9,625 (21%)	17,974 (39%)	18,461 (40%)	46,060	201

Table 2: **Question counts per strata cell and country in MultiGlobeQA.** Rows: World Bank income tiers; columns: H3-resolution-3 density terciles.

ification tags: (i) an income tier (low-, lower-middle-, upper-middle-, or high-income countries; *LIC/LMC/UMC/HIC*) obtained from the World Bank FY2026 classification, and (ii) a density tier (*low/medium/high*) obtained from the WorldPop R2025A 1km population raster (Tatem, 2017), binning values at the global 33rd and 67th percentiles.¹ Combining four income and three density tiers yields a 4×3 grid of 12 strata. We sample entities and generate questions independently within each cell, drawing up to 30 entity tuples per (template, KG, stratum) pair so that every populated cell contributes whenever the template is satisfiable. MultiGlobeQA spans 201 countries and territories across all income and density tiers (Table 2, Fig. 5).

3.3 Ground-truth Computation

We generate MultiGlobeQA with execution-based ground truth: instead of generating questions and verifying answers post-hoc, we pair each question template with a query template executed over the KG. For each (template, KG, strata cell) combination, the executor filters candidate entities by role-specific type (e.g., cafe), evaluates the query, and samples from the results. We validate each result against type-specific constraints (e.g., distances bounded by Earth’s circumference, polygons closed and non-self intersecting) and discard degenerate executions (i.e., divisions by zero, disconnected graphs). We de-duplicate on (question, ground truth) and instantiate each tuple with a uniformly sampled natural-language variant. The resulting English benchmark contains **46,060 verified QA pairs** spanning all **65 TIDs** and **315 variants**, with approximately 6.9k, 24k, and 15k pairs grounded in KnowWhereGraph, OSMH3KG, WorldKG, respectively (Table 7). Appendix A.2 reports per-SFID and answer-format distributions.

¹Appendix A provides further details.

Code	Language	Family / Subgroup	Script	Tier
fra	French	IE / Romance	Lat	H
ita	Italian	IE / Romance	Lat	H
por	Portuguese	IE / Romance	Lat	H
ron	Romanian	IE / Romance	Lat	M
spa	Spanish	IE / Romance	Lat	H
bul	Bulgarian	IE / Balto-Slavic	Cyr	M
rus	Russian	IE / Balto-Slavic	Cyr	H
deu	German	IE / Germanic	Lat	H
eng	English	IE / Germanic	Lat	H
urd	Urdu	IE / Indo-Aryan	PA	L
sqi	Albanian	IE / Albanian	Lat	L
ell	Greek	IE / Graeco-Phrygian	Grk	M
jpn	Japanese	Japonic / Japanesic	JK	M
kat	Georgian	Kartvelian / Georgian-Zan	Geo	M
tur	Turkish	Turkic / Oghuz	Lat	M
vie	Vietnamese	Austro-Asiatic / Vietic	Lat	M
zho	Chinese	Sino-Tibetan / Sinitic	Han	M

Table 3: **The 17 languages in MultiGlobeQA.** Codes are ISO 639-3; *Family/Subgrouping* follow Glottolog (Hammarström et al., 2026); *Tier* (H/M/L = high-/mid/low resource) is adapted from FLORES (Goyal et al., 2022; Team et al., 2022). Scripts: Lat (Latin), Cyr (Cyrillic), Grk (Greek), PA (Perso-Arabic), Han, JK (Japanese Kana/Kanji), Geo (Georgian).

3.4 False-Premise and Multimodal Questions

We include **3,589 false-premise questions** (7.8% of the English benchmark), derived by perturbing a satisfiable (template, entity tuple) pair so that the template’s underlying premise no longer holds, for example, a containment relation that is false or a route passing through no entity of the requested type. These questions probe whether models detect the false premise by refusing to answer, rather than hallucinating plausible-sounding answers.

For a subset of templates constrained to visually salient entities (e.g., landmarks, monuments), we release a separate multimodal slice over OSMH3KG and WorldKG (**946 questions**, 2% of the English benchmark), replacing the entity name with a representative image from Wikidata or Wikimedia Commons via the KG’s image triples.²

3.5 Multilingual Extension

We translate the English benchmark to 16 languages, chosen to span six language families, three resource tiers, and seven scripts (Table 3).

Template Translation. We translate the 315 template variants into each target language with Google Cloud Translation Advanced (v3), an adaptive LLM-based machine translation (MT) system.³

²We exclude KnowWhereGraph because its hazard- and event-centric entities lack canonical image bindings.

³To preserve template slots, we mark placeholders as non-

Two proficient speakers of both English and the target language then post-edit every translated template to ensure semantic precision (e.g., the difference between “within” and “inside”) and fluency, assigning one of three ordinal labels (*Correct* < *Correct – Phrasing Improvement Needed* < *Incorrect*) and supply a corrected string when the label is not *Correct* (Appendix B.1).

Disagreement Resolution. The two annotators per language reach an agreement of 69.7%. For the remaining cases, we consolidate the annotators’ judgments into a single translation per language in two stages: a rule-based pass which handles deterministic cases, and an LLM-based judge (Claude Opus 4.7) for the non-deterministic ones (see Appendix B.2). Rule-based resolution settles 88.6% of items across 16 target languages, and the judge arbitrates the rest. On a sample of four languages, a third annotator agreed with the judge on at least 88.9% of arbitrated questions (Table 9, Fig. 7a).

3.6 Post-editing Instantiated Questions

Instantiating templates with concrete entities introduces language-specific grammatical artifacts (e.g., wrong articles, agreement mismatches) that are tedious to fix manually at scale. We correct them with an ensemble of three LLM post-editors (Gemini-3-Flash, DeepSeek-v4-Flash, and Qwen3.5-27B), each restricted to the grammar surrounding substituted entities and returning typed edits from a 13-category taxonomy. We apply edits proposed by at least two models, falling back to Gemini-3-Flash’s output otherwise. Across all 17 languages, the ensemble edits 16.2% of questions, with at least two models agreeing on 90.1% (see Appendix C).

3.7 Released Variants

We provide MultiGlobeQA in two variants, both spanning all 17 languages. The *large* variant is the full benchmark of 46,060 QA pairs per language. The *small* variant is a canonical subset of *large*, comprising 5,916 questions sampled at up to 200 per (KG, SFID) cell, preserving per-cell coverage while reducing evaluation cost roughly eightfold. We publicly release the benchmark, code, and the KG snapshots (Anonymous, 2026).⁴

translatable. A post-processing pass verifies that every source placeholder appears exactly once in the output.

⁴Code: <https://github.com/andreeaiana/MultiGlobeQA>, data: <https://huggingface.co/datasets/aiana94/MultiGlobeQA>

4 Experimental Setup

Models. We evaluate three open-weight multi-modal LLMs – Qwen3.5-35B and Qwen3.5-27B (Qwen Team, 2026), and Gemma-3-27B-Instruct (Team et al., 2025) – and one closed-source model, Gemini-3-Flash (Google DeepMind, 2025). Gemini-3-Flash and Qwen3.5-35B use native reasoning modes, while Qwen3.5-27B and Gemma-3-27B-Instruct use prompted chain-of-thought. We also report a **majority-class baseline** that predicts each sub-template’s most frequent gold answer as a non-reasoning lower bound. Appendices D and E report model configurations and inference setup.

Evaluation Settings. We evaluate LLMs under three tiers that isolate parametric knowledge, reasoning, and tool use. **Tier 1 (T1, parametric)** directly presents the question. **Tier 2 (T2, reasoning)** adds explicit reasoning over the same prompt. **Tier 3 (T3, agentic)** lets the model write and execute Python in a restricted CODEAGENT interpreter (Roucher et al., 2025), iteratively calling tools to gather evidence. Its variants differ in the retrieval source: spatial KGs (T3a), web search (T3b), and both (T3c). We additionally evaluate four *oracle* conditions that inject the gold KG triples supporting each answer, approximating perfect retrieval. Three differ only in surface form – structured JSON, raw N-Triples, or verbalized prose (T1 oracle-structured/-raw/-verbalized) – to test representation sensitivity. The fourth (T3 oracle) supplies the same structured triples to the T3 agent, retaining the Python interpreter but disabling retrieval tools. These comparisons separate missing knowledge from missing computation (T1-oracle vs. T1) and live from perfect retrieval (T3-oracle vs. T3). Appendix H.2 lists all prompts.

Evaluation Metrics. We report two core metrics: *exact match* (EM) and *normalized error* (NE), both computed by deterministic, answer-type-dependent functions. For EM, we define a correctness function $E_{t_i}(\hat{a}_i, a_i^*) \in \{0, 1\}$ for each answer type t_i , which returns 1 iff the prediction $\hat{a}_i$ matches the gold a_i^* under type-specific criteria: exact equality for categorical and grid-cell answers, geodesic or relative tolerance τ_i for continuous answers (distances, areas, ratios, coordinates, geometries, dates), exact set equality for enumerations, and $\text{IoU} \geq 0.5$ for polygons. NE instead measures error magnitude through $e_t(\hat{a}_i, a_i^*) \in [0, 1]$: $\min(1, d_i/\tau_i)$ for continuous types, $1 - \text{Jaccard}$

	Qwen3.5-27B					Qwen3.5-35B					Gemma-3-27B-Instruct					Gemini-3-Flash				
Tier	EM	NE	Cov.	EM _{cov}	FRR	EM	NE	Cov.	EM _{cov}	FRR	EM	NE	Cov.	EM _{cov}	FRR	EM	NE	Cov.	EM _{cov}	FRR
<i>no external context</i>																				
T1	4.4 (0.1)	95.9 (0.1)	24.7 (2.1)	17.8 (1.1)	75.0 (2.2)	2.0 (0.0)	98.0 (0.0)	6.6 (0.0)	30.8 (0.4)	93.4 (0.0)	4.8 (0.0)	95.8 (0.0)	69.8 (0.0)	6.9 (0.0)	30.2 (0.0)	22.7 (0.1)	78.5 (0.1)	93.5 (0.2)	24.3 (0.1)	6.5 (0.2)
T2	3.8 (0.0)	96.6 (0.0)	31.9 (0.3)	12.0 (0.1)	67.8 (0.3)	0.2 (0.0)	99.8 (0.0)	0.5 (0.0)	49.1 (5.2)	99.5 (0.0)	3.8 (0.0)	96.7 (0.0)	56.0 (0.0)	6.8 (0.0)	44.0 (0.0)	24.6 (0.1)	76.8 (0.1)	91.5 (0.5)	26.9 (0.1)	8.5 (0.6)
<i>retrieval from KG (T3a), web (T3b), or both (T3c)</i>																				
T3a	44.3 (0.4)	57.0 (0.3)	61.2 (0.4)	72.5 (0.4)	1.9 (0.1)	29.3 (10.8)	71.3 (10.6)	39.4 (14.1)	73.9 (1.5)	4.9 (1.2)	23.9 (9.6)	76.7 (9.2)	45.7 (19.2)	53.3 (2.0)	47.9 (23.8)	29.5 (3.5)	71.7 (3.3)	73.3 (6.2)	40.1 (2.2)	25.9 (5.6)
T3b	22.6 (0.7)	77.9 (0.7)	35.0 (0.7)	64.6 (0.9)	0.7 (0.1)	22.8 (0.4)	78.0 (0.4)	36.6 (0.8)	62.3 (0.2)	2.8 (0.3)	21.7 (0.4)	79.5 (0.4)	57.4 (0.6)	37.8 (0.3)	30.8 (0.4)	27.2 [†]	73.8 [†]	74.1 [†]	36.7 [†]	8.2 [†]
T3c	44.1 (0.1)	57.0 (0.2)	59.8 (0.1)	73.7 (0.3)	1.0 (0.0)	37.2 (0.2)	63.5 (0.2)	48.7 (0.1)	76.4 (0.3)	3.8 (0.1)	31.5 (0.5)	69.6 (0.5)	64.0 (1.2)	49.2 (0.3)	22.3 (0.7)	28.0 [†]	73.2 [†]	75.4 [†]	37.1 [†]	24.0 [†]
<i>oracle conditions (gold triples injected)</i>																				
T1-o	35.6 (0.2)	65.0 (0.2)	60.1 (0.3)	59.2 (0.5)	35.8 (0.3)	8.3 (0.1)	91.3 (0.1)	15.2 (0.1)	54.7 (0.5)	80.9 (0.1)	24.9 (0.0)	75.4 (0.0)	64.0 (0.0)	38.9 (0.1)	32.2 (0.0)	57.0 [†]	43.0 [†]	89.2 [†]	63.9 [†]	9.9 [†]
T3-o	61.6 (0.2)	39.2 (0.2)	80.0 (0.3)	77.1 (0.4)	6.7 (0.2)	55.0 (0.2)	45.7 (0.1)	69.6 (0.1)	79.1 (0.2)	8.4 (0.0)	45.3 (0.2)	55.6 (0.1)	61.9 (0.1)	73.2 (0.4)	12.0 (0.4)	60.2 [†]	40.0 [†]	73.9 [†]	81.7 [†]	21.6 [†]
Baseline	28.8	71.2	100.0	28.8	0.0	28.8	71.2	100.0	28.8	0.0	28.8	71.2	100.0	28.8	0.0	28.8	71.2	100.0	28.8	0.0

Table 4: **Results across evaluation tiers** (*small* split, English, text modality; EM ↑, NE ↓, FRR ↓). Evaluation spans 4,979 *true-premise* questions, with refusals counted incorrect. Oracle tiers inject gold triples; T1-o uses *verbalized* triples, the best oracle format (Appendix G.3). EM above baseline is in bold. Parentheses give standard deviation over three seeds; [†] marks cells with fewer seeds, and single-run cells show no standard deviation.

for sets, and $1 - \text{IoU}$ for polygons, with $e_{t_i} = 1$ for refusals and format errors. EM and NE are the means of E_{t_i} and e_{t_i} over the N questions; $\text{NE} = 0$ is a perfect prediction and $\text{NE} = 1$ an error beyond tolerance, with full definitions in Table 12. Unless stated otherwise, we compute EM and NE over *true-premise* questions, counting refusals and non-answers as incorrect. We additionally report coverage (Cov), the share of true-premise questions receiving a parseable answer, and EM_{cov} over those, so that $\text{EM} = \text{Cov} \times \text{EM}_{\text{cov}}$ separates willingness to answer from accuracy when answering. Coverage is reduced by abstention, format errors, harness errors (e.g., context-window overflow) and, at the agentic tiers, by runs that exhaust the step budget. FRR counts abstentions on answerable questions, and the negative rejection rate (NRR) correct rejections of false-premise ones, thus separating premise detection from indiscriminate abstention. Budget exhaustion is not a refusal and is excluded from both.⁵ Unless otherwise specified, we report means and standard deviations over three runs.

5 Results and Discussion

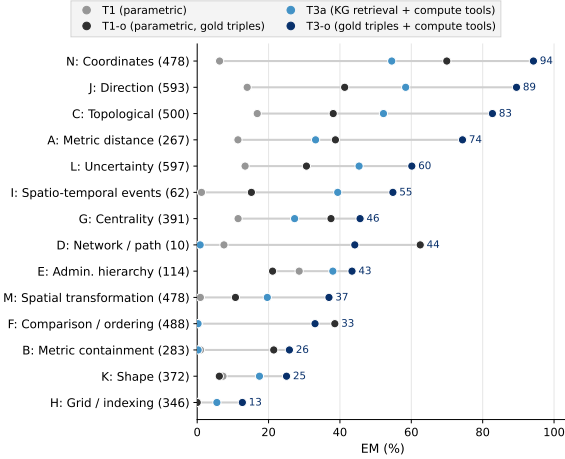
We evaluate all LLMs across evaluation tiers and along the dimensions MultiGlobeQA stratifies. Un-

less stated otherwise, we report performance on the English text portion of the *small* split, which Appendix G.1 shows to be representative of the *large* one; Appendix G.9 reports the multimodal slice.

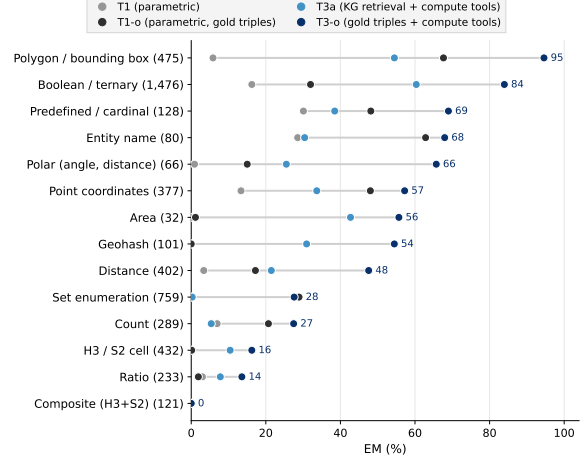
5.1 Performance Across Evaluation Tiers

No model exceeds the 28.8 EM majority baseline without retrieval or injected gold triples (Table 4): Gemini-3-Flash reaches 22.7 at T1, the open-weight models only 2.0 to 4.8. Reasoning before answering (T2) does not close that gap and can widen it (e.g., Gemini-3-Flash +1.9, Qwen3.5-35B -1.8), which follows from the execution-verified construction: gold answers are computed from geometry rather than stated in text, so no amount of thinking recovers a coordinate that the model never memorized. Retrieval reverses this. All open-weight models gain substantially, most of all Qwen3.5-27B (4.4 EM at T1 against 44.3 at T3a), and KG retrieval outperforms web search for every model, by 2.3 (Gemini-3-Flash) to 21.7 (Qwen3.5-27B) points. Combining the two helps the weakest models (Gemma-3-27B-Instruct +7.6, Qwen3.5-35B +7.9 over T3a) and not Gemini-3-Flash (-1.5) or Qwen3.5-27B (-0.2), indicating that web search is largely redundant once the KG supplies the facts. Agentic retrieval is also noticeably less stable. EM standard deviation stays below 0.5 at T1 and T2, but reaches 9.6 (Gemma-3-27B-Instruct) and 10.8

⁵This covers over 90% of the Qwen models’ T3a non-answers but under 10% of the others.



(a) EM by spatial function.



(b) EM by answer format.

Figure 2: **Accuracy (EM) by spatial function (a) and answer format (b) across evaluation conditions** (*small* split, English, text modality; averaged over the four models). Parentheses give the number of *true-premise* questions in each family; the T3-o value is printed at the right of each row. We omit *Date* as it covers only 8 questions (b).

(Qwen3.5-35B) at T3a, as the same query can succeed or exhaust its budget depending on which entities are resolved first.

Abstention. Models differ more in whether they answer than in whether they are right when they do. At T1, coverage ranges from 7% (Qwen3.5-35B) to 94% (Gemini-3-Flash), yet the two are correct on a similar share of what they answer (30.8% vs. 24.3%): Gemini-3-Flash hardly abstains but is frequently wrong, whereas Qwen3.5-35B prefers abstaining. Abstention becomes sensitive to difficulty once models have evidence to assess: ordering the spatial-function families by their accuracy under gold triples, all LLMs refuse more on the harder ones under KG retrieval (Appendix G.2).

Oracle Conditions. Both oracle conditions supply identical gold triples: T1-o in the prompt, T3-o to the tool-using agent. Even with tools, accuracy caps at 61.6, leaving 38% to 55% of questions wrong under perfect retrieval. Injecting the triples lifts parametric accuracy $2.5\times$ (Gemini-3-Flash) to $8.1\times$ (Qwen3.5-27B) and agentic accuracy $1.4\times$ (Qwen3.5-27B) to $2.0\times$ (Gemini-3-Flash), so T3a reaches only 49% to 72% of the accuracy at T3-o. We attribute the remainder to entity resolution and query formulation over a large KG, independently of the spatial reasoning targeted. Adding tools to the same context (T3-o vs. T1-o) raises accuracy by 20.4 to 46.7 points for the open-weight LLMs but 3.2 for Gemini-3-Flash, which computes in context where the others depend on execution, and it closes the gap between them, from Gemini-3-Flash leading Qwen3.5-27B by 21.4 points at T1-o to 60.2

vs. 61.6 at T3-o. The T3-o contexts overflow the open-weight models’ windows on 8.9% to 19.6% of questions against 1.2% for Gemini-3-Flash; excluding harness errors Qwen3.5-35B reaches 68.4 and becomes the best model, so the oracle ranking partly reflects context capacity. Accuracy is nevertheless constrained not by access to the evidence but by computation over it. Tool-augmented evaluation conflates the two (Bao et al., 2026), whereas the oracle conditions separate them.

Compute Cost. Agentic retrieval multiplies median tokens per question by 107 to 437. Comparing T3a with T3-o isolates retrieval costs: the oracle reaches higher accuracy with 1.6 to 6.3 times fewer tokens, and the Qwen models exhaust the step budget on 44% and 59% of retrieval questions against 6% and 5% under T3-o. Most of the agentic budget is thus spent on exploration that never reaches the evidence (Appendix G.4).

5.2 Error Analysis

Under gold context, accuracy spreads by 81 points across the 14 spatial-function families (Fig. 2a): coordinates reach 94, direction 89, shape 25 and grid indexing 13. Administrative hierarchy is the only family substantially answerable from parametric knowledge (28.6 at T1), which we attribute to administrative containment being stated rather than computed. This is driven by Gemini-3-Flash alone (75.2 vs. at most 17.0 for the others), as are the next two families, topological relations (58.1) and direction (36.6). Discrete spatial encodings do not fail uniformly (Fig. 2b): geohash recovers to 54.5

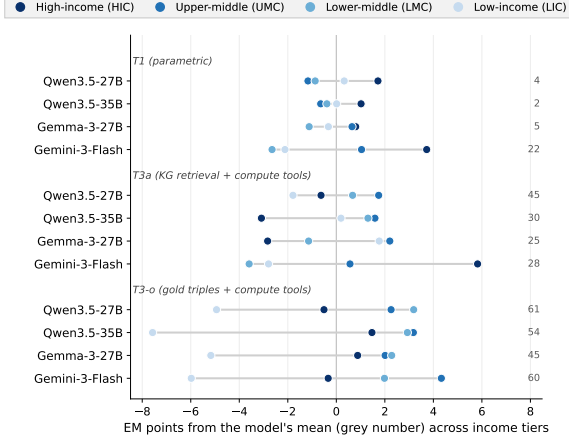


Figure 3: **Accuracy by income tier** (*small split, English, text modality; standardized EM*).

under gold context, H3 and S2 cell indices only 16.2, and their composition is never correct in any condition. Geohash interleaves latitude and longitude bits deterministically, whereas H3 and S2 require projection onto an icosahedral or spherical-cube grid and hierarchical cell arithmetic. Grid indexing ranks last for the same reason: all of its questions require an H3/S2 or composite answer.⁶

Across all models and conditions, NE is 0 for 18.4% and 1 for 72.0%, leaving 9.6% with partial credit: when a model is wrong on MultiGlobeQA, it is typically wrong by a wide margin. NE therefore tracks 1–EM closely (Table 4), lying within one point of 100–EM for eight of the 14 answer formats with at least 10 questions and departing by more than three only for six of the seven graded ones. Among these, set enumerations are near misses, while point coordinates that pass the tolerance retain a large residual error.

5.3 Stratified Analysis

Income Tiers. We next analyse accuracy by income tier, standardizing each tier to the benchmark-wide spatial-function distribution while holding its own per-family accuracy fixed.⁷ Questions about high-income regions are answered more accurately under parametric knowledge by every model (+1.0 Qwen3.5-35B to +5.8 Gemini-3-Flash; Fig. 3), and the margin widens under perfect retrieval (+4.4 Qwen3.5-27B to +9.0 Qwen3.5-35B at T3-o); each of these gaps is at least twice its seed-to-seed varia-

tion. Under KG retrieval the reversals for Qwen3.5-35B (-3.3) and Gemma-3-27B-Instruct (-4.6) stay within seed variation, while Gemini-3-Flash (+8.6) and Qwen3.5-27B (+1.2) exceed theirs. At T3-o the lowest tier is low-income for every model, while the highest is a middle tier, i.e., not a simple ordering by income. The injected oracle contexts are themselves uneven, carrying on average about twice as many gold triples for high-income as for low-income questions, so evidence volume and regional familiarity cannot be separated.⁸

False-Premise Questions. Comparing negative rejection against the false-refusal rate on the same 13 sub-templates (Fig. 20), all four models decline false premises far more often than answerable questions under retrieval, most sharply Qwen3.5-27B (85.7% vs. 13.6%) and least Gemini-3-Flash (63.2% vs. 33.9%). Four of the 20 model-tier cells fall below chance, all without retrieval.⁹

Languages. As MultiGlobeQA’s 17 languages are translations of one fixed question set, a per-language comparison isolates language from question difficulty. Accuracy is stable: at T1 the best-to-worst spread is at most 2.1 points for every model, and 3.9 to 9.0 for Qwen3.5-27B across tiers (Fig. 21). Refusal varies more, up to 11.0 points across languages. These comparisons vary only the question and its entity labels, as the instructions, answer format and tool catalog remain English, and thus isolate sensitivity to question surface form rather than end-to-end multilingual competence. The largest deviations under retrieval are Georgian and Urdu, whose non-Latin entity names pass through an English tool interface.

6 Conclusion

We introduce MultiGlobeQA, a benchmark of 46,060 geospatial questions spanning 14 spatial functions and 15 answer formats, stratified by income and population density over 201 countries and territories, and released parallel in 17 languages with false-premise and multimodal slices. Evaluating four LLMs across parametric, reasoning, and agentic tiers, we find that computation rather than access to data is the bottleneck: given gold triples and compute tools, no model exceeds 61.6 EM, and grid indexing stays below 13. Re-

⁶Appendix G.5 gives the per-model breakdowns.

⁷The benchmark design confounds raw comparison: it equalizes question counts per (income, density) cell but not their composition, and high-income cells with richer map data carry more of the hardest spatial functions.

⁸Fig. 18 shows the same analysis by population-density tier, and Fig. 19 replicates both at T1 on the *large split*.

⁹These rates cover only runs that terminated, a non-random subset: Qwen3.5-27B completes 56% of its T3a questions.

trieval and tools account for nearly all the gains, while low-income regions trail high-income ones under parametric knowledge and under gold triples. Closing this gap calls for models that perform spatial computation reliably, whether in context or through tools, rather than for more parametric geographic knowledge.

Limitations

Template coverage is bounded by query authoring rather than by question writing: each template requires a hand-written SQL shape and a matching gold triple query, so a new spatial function or KG costs a new query. Once that query exists, question count is not a constraint, since further instances follow from sampling entities that satisfy it. In exchange, correctness can be decided by executing a query, not by an LLM judge.

Ground truth is computed against a fixed snapshot of each KG, and is therefore correct with respect to the snapshot rather than the world. Incompleteness affects questions whose answer ranges over a set of entities, such as counts, set enumerations, and nearest-entity questions, since a query over a KG missing those entities executes successfully but returns an incomplete answer. Errors in the source KGs, such as inaccurate geometries or mislabeled feature types, propagate to the gold answers. We release the gold triples for each question so that individual answers remain auditable. Snapshots also become outdated, such that a model retrieving current information in the web-search agentic tier may be scored against a gold answer that no longer holds. Re-executing the queries against a more recent snapshot would quantify this drift.

Frontier closed models have been reported to reason more reliably and to use tools and search more effectively than open-weight ones (Chen et al., 2025; Golikov et al., 2026; Fan et al., 2026). Our academic budget constrains which models we can evaluate: we serve three open-weight LLMs locally and access one closed model from the small, fast tier of its family rather than the frontier tier. Our results do not show closed models to be uniformly stronger: Gemini-3-Flash leads without tools, but the best agentic and oracle scores are Qwen3.5-27B’s. Krechetova and Kochedykov (2025) likewise find that the newest of eight commercial models is not the most accurate on multistep geospatial agent tasks. We therefore cannot rule out that larger or newer models would close the gaps we report.

We constrain model outputs to the gold answer type’s JSON schema, which isolates spatial reasoning from format compliance and reduces format errors. However, it also tells the model the shape of the answer, so MultiGlobeQA does not test whether a model can produce a spatial answer in free-form text, nor whether it can select the appropriate representation itself, which we leave to future work.

Ethical Considerations

MultiGlobeQA is derived from openly licensed sources (Appendix A), and images in the multimodal slice come from Wikidata and Wikimedia Commons through the entities’ own image links. Entities represent public geographic features and points of interest, so neither the questions nor the released triples contain personal data. Annotators for the 16 target languages were recruited through the authors’ academic networks and among professional translators, gave informed consent, and were appropriately compensated (Appendix B.1).

The coverage disparities we analyze are properties of the underlying KGs and are inherited by any benchmark built on them. Stratified sampling balances questions across regions, but the entity pool remains denser in high-income regions. Strong aggregate performance on MultiGlobeQA should therefore not be read as evidence of uniform capability across the world, and per-region results are the more informative signal. This matters because spatial errors carry real consequences in the applications that motivate this work, such as navigation, logistics, and disaster response. Our results show that current models fail on the computations these applications depend on, and we would caution against reading high scores on MultiGlobeQA as evidence of readiness for deployment.

Acknowledgments

This work was supported by the SpatialBenchRAG project grant of the Open Science Office of the University of Mannheim, by the Google Cloud Research Credits program with the award EDU460771163, and by the state of Baden-Württemberg through bwHPC.

References

- Anonymous. 2026. [MultiGlobeQA — knowledge-graph snapshots \(v1\)](https://doi.org/10.5281/zenodo.21786184). Dataset. <https://doi.org/10.5281/zenodo.21786184>.

- Mutian Bao, Qiuyi Qi, Tian Liang, Jinjian Zhang, Wei Zhou, Ming Kong, Linjian Mo, and Qiang Zhu. 2026. [UrbanGeoEval: A City-Scale Benchmark for Evaluating Large Language Models in Geospatial Reasoning](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 40183–40223.
- Prabin Bhandari, Antonios Anastasopoulos, and Dieter Pfoser. 2023. [Are large language models geospatially knowledgeable?](#) In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pages 1–4.
- Martin Böckling, Heiko Paulheim, and Sarah Detzler. 2024. [A Planet Scale Spatial-Temporal Knowledge Graph Based On OpenStreetMap And H3 Grid](#). In *GeoLD2024: 6th Geospatial Linked Data Workshop*.
- Isaac Brodsky. 2018. H3: Uber’s Hexagonal Hierarchical Spatial Index. <https://www.uber.com/blog/h3/>. Uber Engineering Blog.
- Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, and 1 others. 2025. [BrowseComp-Plus: A More Fair and Transparent Evaluation Benchmark of Deep-Research Agent](#). *arXiv preprint arXiv:2508.06600*.
- A.G. Cohn and SM Hazarika. 2001. Qualitative Spatial Representation and Reasoning: An Overview. *Fundamenta Informaticae*, 46(1-2):1–29.
- Paramita Das, Sai Keerthana Karnam, Aditya Bharat Soni, and Animesh Mukherjee. 2025. [Social Biases in Knowledge Representations of Wikidata separates Global North from Global South](#). In *Proceedings of the 17th ACM Web Science Conference 2025*, pages 12–21.
- Mahir Labib Dihan, Md Tanvir Hassan, Md Tanvir Parvez, Md Hasebul Hasan, Md Almash Alam, Muhammad Aamir Cheema, Mohammed Eunus Ali, and Md Rizwan Parvez. 2025. [MapEval: A Map-Based Evaluation of Geo-Spatial Reasoning in Foundation Models](#). In *International Conference on Machine Learning*, pages 13774–13813. PMLR.
- Gabriel Ionut Dorobantu and Ana Cornelia Badea. 2026. [Geospatial reasoning and awareness in large language models: a systematic review](#). *Artificial Intelligence Review*, 59(4):111.
- Alishiba Dsouza, Nicolas Tempelmeier, Ran Yu, Simon Gottschalk, and Elena Demidova. 2021. [WorldKG: A World-Scale Geographic Knowledge Graph](#). In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, page 4475–4484, New York, NY, USA. Association for Computing Machinery.
- Max J Egenhofer and Robert D Franzosa. 1991. [Point-set Topological Spatial Relations](#). *International Journal of Geographical Information System*, 5(2):161–174.
- Fahim Faisal and Antonios Anastasopoulos. 2023. [Geographic and Geopolitical Biases of Language Models](#). In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 139–163.
- Shengda Fan, Xuyan Ye, Yupeng Huo, Zhi-Yuan Chen, Yiju Guo, Shenzhi Yang, Wenkai Yang, Shuqi Ye, Jingwen Chen, Haotian Chen, and 1 others. 2026. [AgentProcessBench: Diagnosing Step-Level Process Quality in Tool-Using Agents](#). *arXiv preprint arXiv:2603.14465*.
- Pavel Golikov, Evgenii Opryshko, Gennady Pekhimenko, and Mark C Jeffrey. 2026. [Robust Reasoning Benchmark](#). *arXiv preprint arXiv:2604.08571*.
- Google DeepMind. 2025. Gemini 3 Flash Model Card. <https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf>. Accessed: 2026-03-18.
- Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. [The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation](#). *Transactions of the Association for Computational Linguistics*, 10:522–538.
- Wes Gurnee and Max Tegmark. 2024. [Language Models Represent Space and Time](#). In *International Conference on Learning Representations*, volume 2024, pages 2483–2503.
- Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2026. [Glottolog 5.3](#). (Available online at <http://glottolog.org>. Accessed on 2026-04-17.).
- Md Hasebul Hasan, Mahir Labib Dihan, Tanzima Hashem, Mohammed Eunus Ali, and Md Rizwan Parvez. 2026. [MapAgent: A Hierarchical Agent for Geospatial Reasoning with Dynamic Map Tool Integration](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 1296–1322.
- Benjamin Herfort, Sven Lautenbach, João Porto de Albuquerque, Jennings Anderson, and Alexander Zipf. 2023. [A spatio-temporal analysis investigating completeness and inequalities of global urban building data in OpenStreetMap](#). *Nature Communications*, 14(1):3985.
- Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. [YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia](#). *Artificial intelligence*, 194:28–61.
- Jaedong Hwang, Kumar Tanmay, Seok-Jin Lee, Ayush Agrawal, Hamid Palangi, Kumar Ayush, Ila Fiete, and Paul Pu Liang. 2025. [Learn Globally, Speak Locally: Bridging the Gaps in Multilingual Reasoning](#). *arXiv preprint arXiv:2507.05418*.

- Krzysztof Janowicz, Pascal Hitzler, Wenwen Li, Dean Rehberger, Mark Schildhauer, Rui Zhu, Cogan Shimizu, Colby Fisher, Ling Cai, Gengchen Mai, and 1 others. 2022. [Know, Know Where, KnowWhere-Graph: A densely connected, cross-domain knowledge graph and geo-enrichment service stack for applications in environmental intelligence](#). *AI Magazine*, 43(1):30–39.
- Bowen Jin, Chulin Xie, Jiawei Zhang, Kashob Kumar Roy, Yu Zhang, Zheng Li, Ruirui Li, Xianfeng Tang, Suhang Wang, Yu Meng, and 1 others. 2024. [Graph Chain-of-Thought: Augmenting Large Language Models by Reasoning on Graphs](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 163–184.
- Nikolaos Karalis, Georgios Mandilaras, and Manolis Koubarakis. 2019. [Extending the YAGO2 Knowledge Graph with Precise Geospatial Knowledge](#). In *International Semantic Web Conference*, pages 181–197. Springer.
- Sergios-Anestis Kefalidis, Dharmen Punjani, Eleni Tsalapati, Konstantinos Plas, Maria-Aggeliki Polali, Pierre Maret, and Manolis Koubarakis. 2024. [The question answering system GeoQA2 and a new benchmark for its evaluation](#). *International Journal of Applied Earth Observation and Geoinformation*, 134:104203.
- Sergios-Anestis Kefalidis, Dharmen Punjani, Eleni Tsalapati, Konstantinos Plas, Mariangela Pollali, Michail Mitsios, Myrto Tsokanaridou, Manolis Koubarakis, and Pierre Maret. 2023. [Benchmarking Geospatial Question Answering Engines Using the Dataset GeoQuestions1089](#). In *International semantic web conference*, pages 266–284. Springer.
- Varvara Krechetova and Denis Kochedykov. 2025. [GeoBenchX: Benchmarking LLMs in Agent Solving Multistep Geospatial Tasks](#). In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Generative and Agentic AI for Multi-Modality Space-Time Intelligence*, pages 27–35.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient Memory Management for Large Language Model Serving with PagedAttention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.
- Jinyang Li, Binyuan Hui, Ge Qu, Jiayi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, and 1 others. 2023. [Can LLM AI-ready Serve as A Database Interface? A Big Bench for Large-Scale Database Grounded Text-to-SQLs](#). *Advances in Neural Information Processing Systems*, 36:42330–42357.
- Wenbin Li, Di Yao, Ruibo Zhao, Wenjie Chen, Zijie Xu, Chengxue Luo, Chang Gong, Quanliang Jing, Haining Tan, and Jingping Bi. 2025a. [STBench: Assessing the Ability of Large Language Models in Spatio-Temporal Analysis](#). In *Companion Proceedings of the ACM on Web Conference 2025*, pages 749–752.
- Zekun Li, Malcolm Grossman, Mihir Kulkarni, Muhao Chen, Yao-Yi Chiang, and 1 others. 2025b. [MapQA: Open-domain Geospatial Question Answering on Map Data](#). *arXiv preprint arXiv:2503.07871*.
- Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, and 1 others. 2024. [On the Opportunities and Challenges of Foundation Models for GeoAI \(Vision Paper\)](#). *ACM Transactions on Spatial Algorithms and Systems*, 10(2):1–46.
- Gengchen Mai, Krzysztof Janowicz, Rui Zhu, Ling Cai, and Ni Lao. 2021. [Geographic Question Answering: Challenges, Uniqueness, Classification, and Future Directions](#). *AGILE: GIScience series*, 2:8.
- Rohin Manvi, Samar Khanna, Marshall Burke, David B Lobell, and Stefano Ermon. 2024. [Large language models are geographically biased](#). In *International Conference on Machine Learning*, pages 34654–34669. PMLR.
- Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. 2021. [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*, pages 4582–4598.
- Roshanak Mirzaee and Parisa Kordjamshidi. 2022. [Transfer Learning with Synthetic Corpora for Spatial Role Labeling and Reasoning](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6148–6165.
- Mazda Moayeri, Elham Tabassi, and Soheil Feizi. 2024. [WorldBench: Quantifying Geographic Disparities in LLM Factual Recall](#). In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1211–1228.
- Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. [POTATO: The Portable Text Annotation Tool](#). In *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337.
- Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024. [YaRN: Efficient Context Window Extension of Large Language Models](#). In *International Conference on Learning Representations*, volume 2024, pages 31932–31951.
- Dharmen Punjani, Kuldeep Singh, Andreas Both, Manolis Koubarakis, Iosif Angelidis, Konstantina Bereta, Themis Beris, Dimitris Bilidas, Theofilos Ioannidis, Nikolaos Karalis, and 1 others. 2018. [Template-Based Question Answering over Linked Geospatial](#)

- Data. In *Proceedings of the 12th Workshop on Geographic Information Retrieval*, pages 1–10.
- Qwen Team. 2026. **Qwen3.5: Towards native multi-modal agents**.
- David A Randell, Zhan Cui, and Anthony G Cohn. 1992. **A Spatial Logic Based on Regions and Connection**. *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*, 92(165-176):40–40.
- Jonathan Roberts, Timo Lüddecke, Sowmen Das, Kai Han, and Samuel Albanie. 2023. **GPT4GEO: How a Language Model Sees the World’s Geography**. *arXiv preprint arXiv:2306.00020*.
- Keonwoo Roh, Yeong-Joon Ju, and Seong-Whan Lee. 2025. **XLQA: A Benchmark for Locale-Aware Multilingual Open-Domain Question Answering**. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28797–28809.
- Aymeric Roucher, Albert Villanova del Moral, Thomas Wolf, Leandro von Werra, and Erik Kaunismäki. 2025. ‘smolagents’: a smol library to build great agentic systems. <https://github.com/huggingface/smolagents>.
- Majid Saeedan, Muhammad Shihab Rashid, Ahmed El-dawy, and Vagelis Hristidis. 2026. **GS-QA: A Benchmark for Geospatial Question Answering**. *arXiv preprint arXiv:2605.22811*.
- Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. 2022. **StepGame: A New Benchmark for Robust Multi-Hop Spatial Reasoning in Texts**. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11321–11329.
- Tetsuhisa Suizu, Shohei Higashiyama, Hiroyuki Shindo, Hiroki Ouchi, and Sakriani Sakti. 2026. **Automatic Generation of a Compositional QA Benchmark for Geospatial Reasoning under Spatial and Entity Constraints**. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 818–830.
- Andrew J Tatem. 2017. **WorldPop, open data for spatial demography**. *Scientific Data*, 4(1):170004.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. **Gemma 3 Technical Report**. *arXiv preprint arXiv:2503.19786*.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. **No Language Left Behind: Scaling Human-Centered Machine Translation**. *arXiv preprint arXiv:2207.04672*.
- Thinh Hung Truong, Jey Han Lau, and Jianzhong Qi. 2026. **GPSBench: Do Large Language Models Understand GPS Coordinates?** *arXiv preprint arXiv:2602.16105*.
- Denny Vrandečić and Markus Krötzsch. 2014. **Wikidata: A Free Collaborative Knowledge Base**. *Communications of the ACM*, 57(10):78–85.
- Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. **Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks**. *arXiv preprint arXiv:1502.05698*.
- Michael F Worboys and Matt Duckham. 2004. **GIS: A Computing Perspective**. CRC Press.
- Shirley Wu, Shiyu Zhao, Michihiro Yasunaga, Kexin Huang, Kaidi Cao, Qian Huang, Vassilis N Ioannidis, Karthik Subbian, James Zou, and Jure Leskovec. 2024. **STaRK: Benchmarking LLM Retrieval on Textual and Relational Knowledge Bases**. *Advances in Neural Information Processing Systems*, 37:127129–127153.
- Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. 2024. **TravelPlanner: A Benchmark for Real-World Planning with Language Agents**. In *Proceedings of the 41st International Conference on Machine Learning*, pages 54590–54613.
- Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze D Gui, Ziran W Jiang, Ziyu Jiang, and 1 others. 2024. **CRAG - Comprehensive RAG Benchmark**. *Advances in Neural Information Processing Systems*, 37:10470–10490.
- Dazhou Yu, Riyang Bao, Ruiyu Ning, Jinghong Peng, Gengchen Mai, and Liang Zhao. 2026. **Spatial-RAG: Spatial Retrieval Augmented Generation for Real-World Geospatial Reasoning Questions**. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 11094–11112.
- Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, and 1 others. 2018. **Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task**. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3911–3921.
- Qianheng Zhang, Song Gao, Chen Wei, Yibo Zhao, Ying Nie, Ziru Chen, Shijie Chen, Yu Su, and Huan Sun. 2025. **GeoAnalystBench: A GeoAI Benchmark for Assessing Large Language Models for Spatial**

Analysis Workflow and Code Generation. *Transactions in GIS*, 29(7):e70135.

A Benchmark Construction

A.1 Full Template Listing

Table 5 lists all 65 templates with an example and answer format.

Sub-function	Template	Example	Answer type
SFID A — Metric Distance & Proximity Computation / Reasoning			
Direct distance	What is the distance between {entity_A} and {entity_B}?	How far is Mbélé from Bilsem?	Distance
	How far {cardinal_direction} is {entity_A} from {entity_B}?	How far west is Hov from Olsby?	Distance
Event-to-location	How far away was the {event_A} from {entity_B} of {entity_C} in {time_period}?	How far away was the "Earthquake from USGS Earthquake Catalog with ID pr2020052043" from Farmacia Asturias LLC of Puerto Rico from 2020-01-01 to 2023-12-31?	Distance
Comparative	Is {entity_A} or {entity_B} {rank_type} to {entity_C}?	Is Behruz or Payawak closer to Sa tapa?	Entity name (single)
	Which is closer to the country that borders the country containing {entity_A}: {entity_B} or {entity_C}?	Which of Uma Boco or Dilor is nearer to a neighboring country of Jadin's country?	Entity name (single)
Aggregation	What is the {aggregation_type} distance between {entity_A} in {entity_B}?	What is the average distance between hotels in Pedasi?	Distance
Optimization	Which {entity_A} in {entity_B} minimizes total distance to all {entity_C}?	Which Pharmacy in Lucas County, Ohio is closest overall to all BPHC_Sites?	Entity name (single)
Distance-based selection	What is the distance between the {rank_position} nearest {entity_A} to the {entity_B} and the {entity_C}?	How far is the second closest Market to the 7-11 from 24 Bar?	Distance
Boundary distance	What is the minimum distance from {entity_A} to the boundary of {entity_B}?	What is the minimum distance from Koibi to the boundary of Metekel?	Distance
SFID B — Metric Containment			
Counting	How many {entity_A} are within {value} km of {entity_B}?	How many pharmacies are within 5 km of Siren?	Count
Enumeration	Which {entity_A} are within {value} km of {entity_B}?	Which pharmacies are within 5 km of KFC?	Set enumeration
Spatial join	What is the total population of the area within {value} km of {entity_A}?	How many people live within 5 km of Choco?	Count
SFID C — Topological Relationships			
Intersection (boolean)	Does {entity_A} intersect {entity_B}?	Do Ocelot and Karon intersect spatially?	Boolean
	Do the geometries between {entity_A} intersect with the geometry of {entity_B}?	Do the geometries of Basco intersect with the geometry of Nusa?	Boolean
Intersection (count)	In how many points do {entity_A} and {entity_B} intersect?	In how many points do Lystvej and Rudbølvej intersect?	Count
	How many {entity_A} intersect {entity_B}?	How many highways intersect Jacros?	Count
Intersection (enumeration)	Which {entity_A} intersect {entity_B}?	Which highway intersects Calle 1?	Set enumeration
	Which {entity_A} intersect {entity_B} but not {entity_C}?	Which highway intersects Itbayat but not Rael?	Set enumeration
Overlap	Do {entity_A} and {entity_B} overlap spatially?	Do Dano and Fafan overlap spatially?	Boolean
	Does {entity_A} touch the boundary of {entity_B}?	Does Kumejima touch the boundary of Suzu?	Boolean
Adjacency	Is {entity_A} adjacent to {entity_B}?	Is Utiroa adjacent to Roreti?	Boolean
Disjointness	Is {entity_A} physically disconnected from {entity_B}?	Are Klepalo and Grad spatially disjoint?	Boolean
Enclosure	Does {entity_A} completely surround {entity_B}?	Does Basey completely surround Maragat?	Boolean
SFID D — Network & Path-Based Spatial Reasoning			
Traversal	Which {entity_A} does the {entity_B} traverse?	Which administrative regions does the Carretera 10 traverse?	Set enumeration
	How many {entity_A} does the {entity_B} traverse?	How many administrative regions does the Promachonas - Agistro pass through?	Count
Path proximity	Which {entity_A} does the {entity_B} pass by?	Which cafe does the Pont d'Andotsy pass by?	Set enumeration
	How many {entity_A} does the {entity_B} pass by?	How many cafes does the General Espejo pass by?	Count
SFID E — Containment & Administrative Hierarchy			
Membership	Is {entity_A} administratively a part of {entity_B}?	Is Itbayat administratively a part of Basco?	Boolean
Counting	How many {entity_A} make up {entity_B}?	How many components make up Jerma?	Count
	How many {collection_of_entity} are in {entity_A}?	How many bars, cafes and restaurants are in Dano?	Count
Full containment	Does the administrative boundary of {entity_A} fully contain the area of {entity_B}?	Does the administrative boundary of Siaton fully contain the area of Salag?	Boolean
	Does the administrative border of {entity_A} fully contain the {rank_type} {entity_B} of {entity_C}?	Does the administrative border of Arege fully contain the longest Yonga of Teungo?	Boolean
SFID F — Spatial Comparison & Ordering			
Extremes	Which {entity_A} lies furthest {cardinal_direction} within {entity_B}?	Which Pharmacy lies furthest west within Thomaston, GA?	Entity name (single)
Ranking	What is the {rank_type} {entity_A} in {entity_B} by {attribute_A}?	What is the biggest ZipCodeArea in Big Rapids, MI by elongation?	Entity name (single)
	What is the {rank_position} {attribute} {entity_A} in {entity_B}?	What is the longest RoadSegment in Alexandria, MN?	Entity name (single)

Continued on next page.

Table 5 continued from previous page.

Sub-function	Template	Example	Answer type
Compound comparison	Which {entity_A} are within {value} km of {entity_B} and {cardinal_direction} of {entity_C}?	Which cafes are within 10 km of OMV and south of Viva?	Set enumeration
SFID G — Centrality & Medial Geometry			
Centroid / medoid	What is the most central point of {entity_A}?	What is the most central point of Ivana?	Point coordinates
Entity centrality	What is the most central point of {entity_A} where {entity_B} is located?	What is the most central point of Walmara where Kolobo is located?	Point coordinates
Scale sensitivity	How does the centroid of {entity_A} change when computed at spatial resolutions {value_1} km and {value_2} km?	How does the centroid of Øyer change when computed at spatial resolutions 10 km and 100 km?	Distance
SFID H — Grid & Spatial Indexing			
Spatial unit containment	Which grid cell contains both {entity_A} and {entity_B}?	Which grid cell contains both DHL and Auto Will?	H3 / S2 cell index
	Which grid cell at resolution {value} contains both {entity_A} and {entity_B}?	Which grid cell at resolution 7 contains both Dy Cafe and KFC?	H3 / S2 cell index
Multi-indexing	Which grid cell contains {entity_A} both in h3 and s2 formats?	Which grid cell contains Grad both in h3 and s2 formats?	Composite (H3+S2)
SFID I — Spatio-Temporal Reasoning: Spatio-Temporal Events			
Event distance	How far away was the {event_A} from {entity_B} in {time_period}?	During 2020-01-01 to 2023-12-31, how far was "Earthquake from USGS Earthquake Catalog with ID ci40019119" from AIPHARM LLC?	Distance
Occurrence count	Which {entity_A} were affected by {event_A}?	Which regions were affected by Earthquake from USGS Earthquake Catalog with ID nn00793499?	Set enumeration
Temporal extent	When did {event_A} hit {entity_A}?	When did "Earthquake from USGS Earthquake Catalog with ID hv357401" hit Hilo, HI?	Date
Spatial impact	How large is the area affected by {event_A}?	How large is the area affected by "Smoke plume snapshot with index 8 on date 2018-12-22"?	Area
	How large is the area affected by {event_A} in {entity_A}?	How large is the area affected by "SELLEM fire (from the MTBS dataset) that occurred in 1999-08-28 with ID NV3829211543619990828" in Pahrump, NV?	Area
SFID J — Directional & Orientation Reasoning			
Cardinal	Is {entity_A} {cardinal_direction} of {entity_B}?	Is BCA east of Duquinha?	Boolean
Relative to path / location	Is {entity_A} to the left or right of the route from {entity_B} to {entity_C}?	Is Fried Chicken to the left or right of the route from PNB to Burger King?	Predefined
	Does {entity_A} lie upstream or downstream of {entity_B}?	Does Goodydays lie upstream or downstream of 7-Eleven?	Predefined
Angular orientation	What direction would you travel from {entity_A} to reach {entity_B}?	What direction would you travel from Lokenie to reach Kiyungi?	Cardinal direction
SFID K — Shape & Geometry Properties			
Area comparison	Is {entity_A} larger in area than {entity_B}?	Is Lurøy larger in area than Røros?	Boolean
	Is {entity_A} more compact than {entity_B}?	Is Afder more compact than Fafan?	Boolean
	Is {entity_A} more elongated than {entity_B}?	Is Cagayan more elongated than Basco?	Boolean
Perimeter	What is the perimeter of {entity_A}?	What is the perimeter of Dishu?	Distance
Reasoning about shape	How irregular is the boundary of {entity_A}, measured by {irregularity_metric} at {scale_value} km {scale_unit}?	How irregular is the boundary of Korahe, measured by shape index at 1.0 km simplification?	Ratio (dimensionless)
SFID L — Uncertainty & Fuzzy Spatial Reasoning			
Approximate distance	Is {entity_A} approximately near {entity_B}?	Is Terrasse approximately near Oasis?	Boolean
	Is {entity_A} far from {entity_B}?	Is La ceiba far from Tienda?	Boolean
	Is {entity_A} more likely near {entity_B} than {entity_C}?	Is La plage more likely near Pukka than Lio Villas Resort?	Boolean
Vagueness	Which {entity_A} are roughly closest to {entity_B}?	Which Bank is roughly closest to Ecobank?	Set enumeration
Probabilistic containment	Is {entity_A} plausibly within {entity_B}?	Is Ugat plausibly within El Nido?	Ternary
SFID M — Spatial Transformation & Reference Change			
Relative position	What is the relative position of {entity_A} in the coordinate system of {entity_B}?	What is the relative position of Pukka in the coordinate system of Caltex?	Polar (angle, distance)
Conversion	Convert the location of {entity_A} to {grid_type} at {resolution_type} {value} {metric}.	Convert the location of Ons to S2 at level 15.	Grid cell / coords
SFID N — Spatial Reference & Geometric Representation			
Absolute coordinate retrieval	What are the latitude and longitude of the centroid of {entity_A}?	What are the latitude and longitude of the centroid of Ruse?	Point coordinates
Spatial extent representation	What is the bounding box of {entity_A}?	What is the bounding box of Josue?	Polygon / bounding box

Table 5: The 65 top-level templates in MultiGlobeQA, grouped by Spatial Function ID (SFID). Per template we report the sub-function category, canonical English template with placeholders, one example instantiation, and answer type. Each template expands into one or more parameter-bound sub-templates.

	WorldKG	KnowWhereGraph	OSMH3KG
Source	OSM PBF ^a	GraphDB endpoint ^b	OSM PBF ^a
License	ODbL	CC-BY 4.0	ODbL
# Node types	137	17	175
# Nodes	745,194,840	308,071,129	280,078,239
# Geographic nodes	14,304,791	5,026,581	9,505,778
# Leaves	560,306,023	264,469,472	187,064,086
# Predicates	1,645	358	34,038
# Edges	1,308,464,629	860,069,872	838,673,446
Mean / max out-degree	3.72 / 42	5.59 / 1,163,503	4.69 / 8,968,560
Mean / max in-degree	2.30 / 176,096,861	2.82 / 67,810,074	3.02 / 41,102,070

^a Constructed in-house from a single OSM planet PBF snapshot (2025-12-08).

^b Crawled from <https://stko-kwg.geog.ucsb.edu/graphdb/repositories/KWG>; no longer publicly accessible (upstream GraphDB license lapsed). Data remains CC-BY 4.0; we redistribute the cached snapshot used for generation.

Table 6: **The three KGs used to construct MultiGlobeQA.** *ODbL*: Open Data Commons Open Database License; *CC-BY 4.0*: Creative Commons Attribution 4.0.

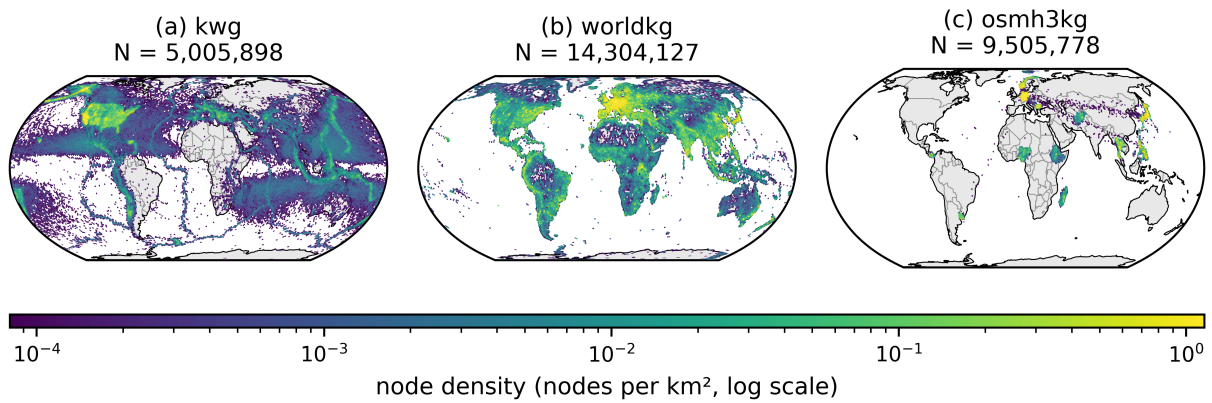


Figure 4: Geographic density of geometries over (a) KnowWhereGraph (KWG), (b) WorldKG, and (c) OSMH3KG.

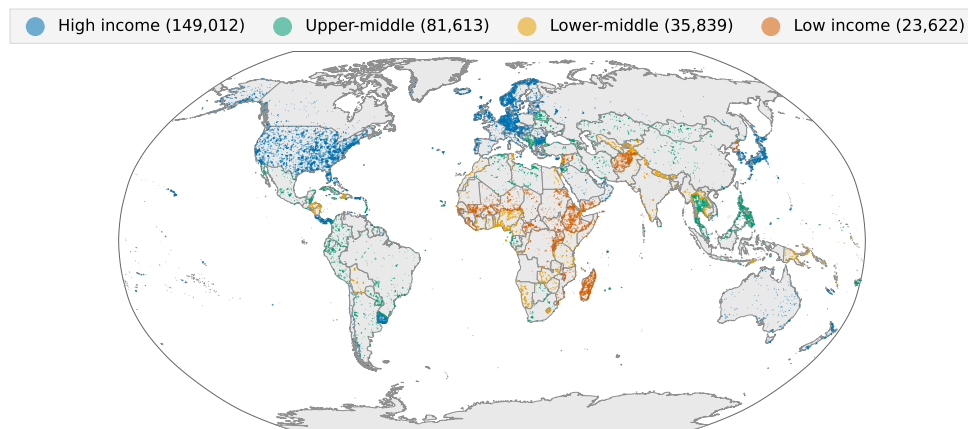


Figure 5: **Global distribution of MultiGlobeQA entities.** 290,086 distinct entity sites across 201 countries and territories, colored by World Bank income tier.

(SFID) Spatial Function	WorldKG	KnowWhereGraph	OSMH3KG
(A) Metric Distance	4,537	1,225	4,046
(B) Metric Containment	329	365	1,193
(C) Topological	1,108	925	3,632
(D) Network / Path	–	–	645
(E) Admin Hierarchy	–	348	3,686
(F) Comparison / Ordering	2,295	954	2,178
(G) Centrality	–	214	616
(H) Grid / Indexing	718	210	745
(I) Spatio-Temp. Events	–	477	–
(J) Direction	2,782	550	2,303
(K) Shape	–	680	2,433
(L) Uncertainty	1,071	378	1,251
(M) Spatial Transformation	1,669	479	1,302
(N) Coordinates	331	94	291
Total	14,840	6,899	24,321

Table 7: **Row counts per (KG, SFID) cell**, including the multimodal slice for OSMH3KG and WorldKG. Dashes (–) mark SFIDs with no data in that KG.

A.2 Data Sources

Table 6 reports provenance and statistics of each source KG, Table 7 the SFID distribution per KG, and Fig. 6 the answer-format distribution. Fig. 4 shows geometry density across the KGs.

Entity-name Canonicalization. On the English benchmark, we use each entity’s English label where available, falling back to `rdfs:label` otherwise. We also preserve aliases from language-specific labels on the entity row and surface them in the alias store used for evaluation.

Geographic Strata and Coverage. Income tiers are assigned by spatially joining each entity’s geometry with Natural Earth 10m country polygons¹⁰ and mapping the resulting ISO-3 to the World Bank FY2026 classification¹¹. Density tiers aggregate the WorldPop R2025A 1km population raster (Tatem, 2017) to H3 resolution 3, and are binned at the global 33rd and 67th percentiles. Fig. 5 shows the global distribution of distinct entity sites.

B Multilingual Extension

B.1 Human Annotation Protocol

Recruitment and Compensation. For each of the 16 target languages, we recruited two annotators through the authors’ academic networks, predominantly Master’s and PhD students in data science, computer science, or related fields, plus a few professional translators. All had C2-level proficiency in both the target language and English. We compensate externally recruited annotators (those not affiliated with the authors’ labs) with €14 per hour.

¹⁰www.naturalearthdata.com/10m-cultural-vectors/

¹¹<https://api.worldbank.org/country>

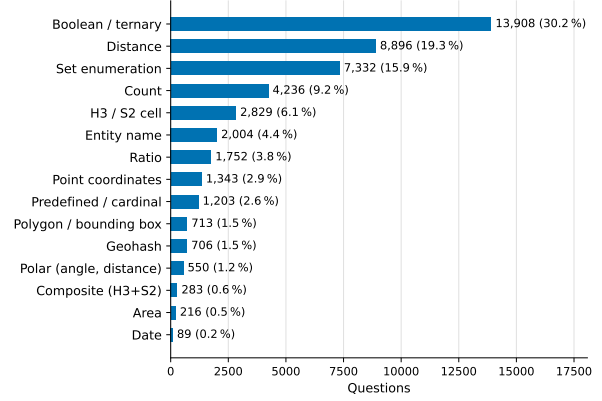


Figure 6: **Distribution of MultiGlobeQA questions over the 15 answer formats.**

Annotation Workflow. We use the Potato annotation tool (Pei et al., 2022). The annotation task comprises six phases.

(1) *Welcome and Consent.* The welcome page summarizes the project, task, and expected workload. Annotators must agree to three consent items before proceeding: that they are at least 18, that their annotations will be used solely for anonymized academic research with no personally identifiable information published, and that they understand these terms. We collect no personally identifiable information apart from a self-chosen annotator ID.

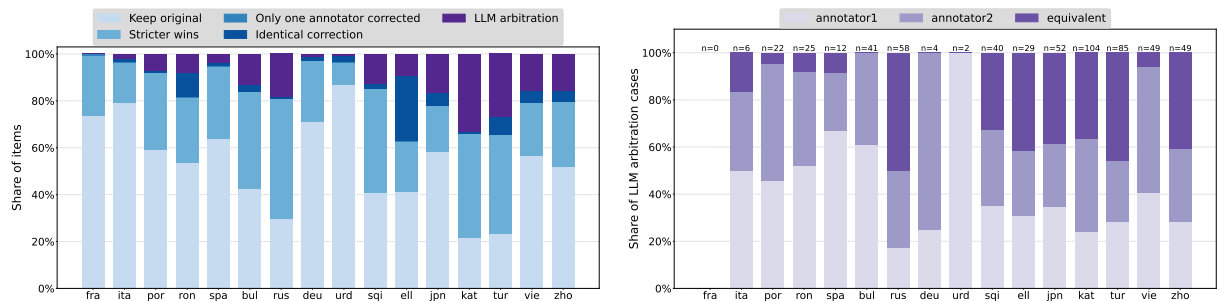
(2) *Annotation Guidelines.* Each annotator must read the full guidelines page before any item is shown; the GUIDELINES link remains accessible throughout. We frame the task as translation quality assessment: judging whether a MT of an English template faithfully preserves the original meaning, structure, and spatial intent. The labels are:

- **Correct:** the spatial relationship, interrogative structure, entity placeholders, and meaning are all preserved, and the phrasing reads naturally to a native speaker.
- **Correct – Phrasing Improvement Needed:** the spatial meaning is preserved and the phrasing is grammatical but unnatural, register-mismatched, or non-idiomatic.
- **Incorrect:** the translation alters, omits or uses the wrong spatial relation; merges, renames, reorders, or omits entity placeholders; adds or removes critical information; or changes the question type (e.g., turning a yes/no question into a *what* or *where* question).

For any non-*Correct* label, the annotator must provide a corrected translation and select at least one

Reason	Description
<i>Correct – Phrasing Improvement Needed</i>	
Word choice / vocabulary	The translation uses a word or phrase that is technically correct but not the most natural or commonly used in the target language (e.g., a less idiomatic synonym).
Grammar / syntax	The sentence structure or grammar is acceptable but awkward, unusual, or not preferred in standard usage.
Register / formality	The translation is too formal, too informal, or does not match the neutral register expected for a factual spatial question.
Word order	The word order, while grammatically correct, is unusual or less natural. Reordering improves readability.
<i>Incorrect</i>	
Meaning changed or lost	The translation alters, omits, or distorts the original meaning of the question.
Information added or removed	The translation includes details not present in the original or leaves out important parts.
Wrong spatial relationship	The spatial relation in the translation differs from the original (e.g., "between" → "near", "north of" → "above").
Entity placeholders altered	One or more {entity_X} placeholders are missing, renamed, merged, split, or otherwise modified.

Table 8: Correction-reason taxonomy non-*Correct* items. Annotators select all that apply.



(a) **Resolution source by language.** Share of items resolved by each deterministic rule (keep original MT, stricter judgment, one-sided resolution, identical corrections) or escalated to LLM arbitration.

(b) **LLM arbitration outcomes by language.** For items escalated to the blind LLM judge (Claude Opus 4.7), distribution of final selections across annotator 1, annotator 2, and equivalent.

Figure 7: **Resolution pipeline outcomes across languages.** Left: resolution source over all items. Right: outcomes for LLM-arbitrated cases only.

reason from the fixed taxonomy in Table 8. Two additional rules prevent over-correction. First, entity placeholders (e.g. {entity_A}) must appear verbatim; annotators are told *not* to translate, capitalize, or otherwise modify them, though they may reorder placeholders to match natural word order. Second, annotators must *not* correct for entity-dependent grammatical variation, e.g., a Romanian masculine participle that would shift to feminine for a specific placeholder is marked *Correct*, as such morphology is handled by the post-editing LLM ensemble (Appendix C). The instructions also include language-specific rules (diacritics, non-Latin scripts, RTL handling, agglutinative-suffix attachment).

(3) *Annotation Items.* Each item appears on its own page showing the English template, the machine translation, an entity-constraint description, and one or two English example instantiations. The annotator selects one of the three labels; non-*Correct*

labels require a corrected translation and one or more reasons from the multi-select taxonomy.

(4) *Attention Checks.* We interleave attention checks with real items throughout each session to detect inattentive responding. Each check is an item whose text contains an explicit instruction with the expected label, the required correction reasons (when applicable), and the corrected string.

(5) *Review & Feedback.* Before submission, annotators may revise any completed item. They also answer three mandatory feedback questions on task difficulty, guideline clarity, and translation quality, plus an optional free-text field.

(6) *Completion.* The final page confirms completion and provides a completion code.

B.2 Disagreement Resolution Pipeline

The first stage is a rule-based pass over deterministic cases: if both annotators label the MT *Correct*,

Metric	fra	ita	por	ron	spa	bul	rus	deu	urd	sqi	ell	jpn	kat	tur	vie	zho
% Correct	86.9	87.9	74.4	66.5	79.0	60.4	50.9	83.4	91.6	58.8	50.7	66.6	38.6	43.4	66.7	64.8
% Phrasing	3.5	12.1	17.5	28.4	19.2	34.9	40.9	13.7	5.0	35.6	45.2	8.7	45.6	27.6	28.5	12.9
Corr. rate	25.6	20.8	41.0	46.1	36.6	57.7	70.0	28.7	12.9	59.3	59.0	42.0	77.9	76.3	43.2	47.9
Agreement	74.3	82.7	66.9	72.2	68.8	58.7	49.2	74.1	90.5	55.2	78.5	80.1	56.1	58.0	77.6	72.2
α	-0.10	0.19	0.16	0.37	0.09	0.23	0.16	0.15	0.48	0.20	0.59	0.69	0.37	0.52	0.57	0.60
κ	0.01	0.19	0.19	0.38	0.09	0.28	0.23	0.20	0.48	0.21	0.59	0.69	0.39	0.52	0.57	0.61
chrF	92.9	95.8	91.7	94.2	95.0	88.2	83.4	94.4	97.0	89.7	93.6	92.8	85.5	85.8	92.1	90.5
% LLM arb.	0.0	1.9	6.9	7.9	3.8	13.2	18.3	1.3	0.6	12.6	9.2	16.4	32.8	26.8	15.5	15.5

The negative α for French reflects an annotator-calibration mismatch (one annotator never used *Incorrect*); the stricter-wins rule routes these items to the conservative correction with no LLM escalation, so the released data stays coherent.

Table 9: **Per-language annotation statistics** ($N = 315$ items). *% Correct*: share of annotator-item decisions labeled *Correct* ($\% \text{ Incorrect} = 100 - \% \text{ Correct} - \% \text{ Phrasing}$). *Corr. rate (%)*: share of items where at least one annotator corrected original MT. **Agreement (%)**: exact agreement on the three-level label scale; α : Krippendorff’s ordinal α ; κ : Cohen’s linear-weighted κ on the same scale. *chrF* is computed between the original MT and the final resolved translation. *% LLM arb.*: share sent to LLM arbitration.

we retain it; if their corrections agree, we take either; if the labels differ, we defer to the stricter annotator, since a stricter label is less likely to be a false-positive *Correct*. We validated this rule against a third annotator on a sample of six languages, observing near-perfect agreement. All text comparisons use NFC Unicode normalization followed by whitespace collapse, so typographically-equivalent corrections resolve via rules.

In the second stage, the LLM judge (Claude Opus 4.7) receives the English template, the original MT, and the two anonymized candidate corrections in randomized order to mitigate positional bias.¹² For each item it returns one of the three choices (*annotator 1*, *annotator 2*, or *equivalent*), with a one-sentence rationale naming the decisive criterion. The output is strict JSON, and the pipeline rejects any response failing schema validation. The final correction is retrieved via the chosen annotator label, so the judge never produces translation content. Fig. 7 shows the distribution of items across deterministic rules per language, and the judge’s selections for escalated questions.

Decision Criteria. The prompt instructs the judge to choose between candidate corrections by the following priority criteria, consulting the next only when the current one does not decide:

1. *Placeholder integrity*: every placeholder must appear verbatim, matching the count and identity in the English template. A correction that renames, merges, splits, or loses a placeholder is worse than one that keeps them intact.

2. *Semantic fidelity*: the spatial relation, quantifier, and question type (yes/no vs. what/where) must match the English; the correction preserving meaning more faithfully is preferred.
3. *Grammatical correctness*: agreement, case, verb form, and word order must be valid in the target language; the correction with fewer grammar errors is preferred.
4. *Technical terminology*: for domain terms (e.g., centroid, variance, fractal dimension), the correction using established vocabulary over approximations is preferred.
5. *Natural phrasing*: among otherwise-equal options, the more idiomatic phrasing is preferred.

To prevent spurious choices, the prompt disallows four dimensions as decision drivers: (i) gender and case agreement hard-coded on a placeholder (these vary with the entity filler and are handled downstream by the post-editing LLM ensemble), (ii) whitespace and trivial-punctuation differences, (iii) correction length, and (iv) register choice (formal/informal), provided it is internally consistent.

Language-specific Instructions. We expand the system prompt with short per-language instructions injected into the `{language_specific_notes}` slot. Each instruction comprises 5-8 points capturing three categories of target-language knowledge not covered by the universal criteria: (i) orthographic conventions and variations that should not influence the choice (e.g., the legacy-cedilla vs.

¹²We choose a frontier closed-source model to reduce training-data overlap with the models under evaluation.

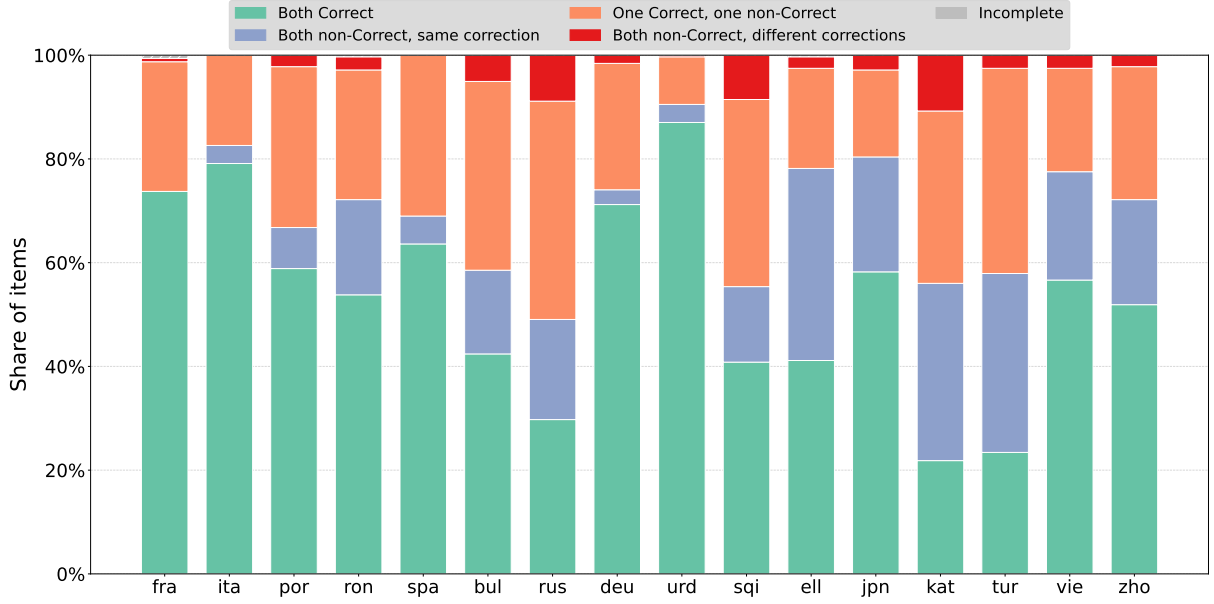


Figure 8: **Pairwise human-annotator agreement on translation post-editing, per target language.** Two annotators independently labeled every translated question as *Correct* (no edit needed) or *non-Correct*, supplying an edited string in the latter case.

comma-below distinction in Romanian), (ii) terminology preferences for domain-specific vocabulary, and (iii) typological anti-patterns, in particular reminders not to penalize corrections for failing to hard-code gender or case agreement on a placeholder. All are available in the project’s repository.

System Prompt Template. Fig. 22 shows the system prompt template, with two expandable slots: {target_language} (the human-readable language name), and {language_specific_notes} (the per-language instructions).

B.3 Per-language Annotation Statistics

Table 9 summarizes the annotation effort and reliability signals per target language.

B.4 Pairwise Annotator Agreement

Fig. 8 shows pairwise agreement on the translation post-editing task. Each item falls into one of five mutually exclusive categories combining the label decision (*Correct* vs. *non-Correct*) with the change type for items corrected by at least one annotator.

B.5 Correction Reason Distribution

Annotators labeling an item as *non-Correct* selected at least one reason from a predefined taxonomy. Fig. 9 reports the normalized correction-reason distribution per language and label. Correction reasons are not mutually exclusive.

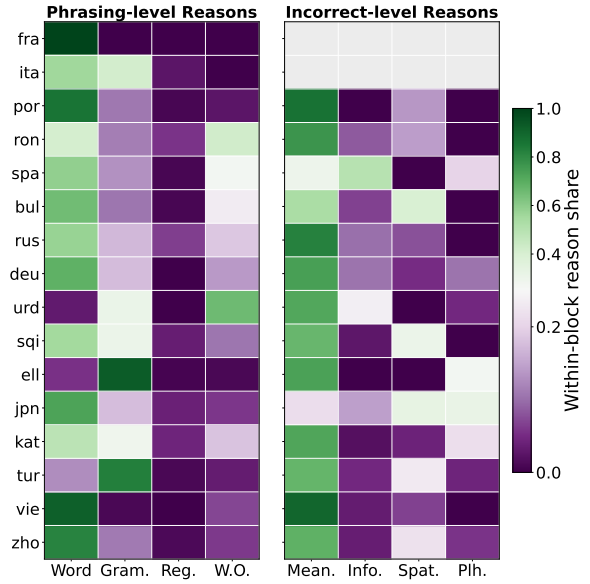


Figure 9: **Normalized multi-select correction-reason distributions per language.** Phrasing-level reasons (left block): *Word*=word choice/vocabulary; *Gram.*=grammar/syntax; *Reg.*=register/formality; *W.O.*=word order/sentence structure. Incorrect-level reasons (right block): *Mean.*=meaning changed or lost; *Info.*=information added or removed; *Spat.*=wrong spatial relationship; *Plh.*=entity placeholders altered.

C Post-editing Ensemble

Edit Taxonomy. Table 10 lists the full 13-category taxonomy used by the post-editing LLM ensemble.

Operation	Description	Example (before → after)	Typological trigger
NO_CHANGE	The instantiated question is already correct; no edit needed.	EN: What is the distance between London and Paris? (unchanged).	—
ART_ADD	Add an article that is missing because the template did not anticipate the entity’s type.	EN: ...distance between <i>Thames</i> and <i>Seine</i> ... → ...distance between <i>the Thames</i> and <i>the Seine</i> ...	Article system
ART_DEL	Delete an article that becomes redundant with the instantiated entity.	DEU: in <i>the Tokyo</i> → <i>in Tokyo</i>	Article system
ART_CHG	Change one article for another, typically to match the entity’s gender or number.	FRA: <i>le</i> Tour Eiffel → <i>la</i> Tour Eiffel	Article system + gender
ADP_ADD	Add an adposition (e.g., preposition, postposition, or case / topic particle) required when the entity is substituted in.	ENG: the capital <i>France</i> → the capital <i>of France</i>	Adposition system
ADP_DEL	Delete an adposition (e.g., preposition, postposition, or case/topic particle) that is now redundant.	ENG: within <i>of 5 km</i> of Paris → within <i>5 km</i> of Paris	Adposition system
ADP_CHG	Change an adposition to another, including case / topic particles.	FRA: <i>à Italie</i> → <i>en Italie</i>	Adposition system
DEC_CHG	Change the declension / case ending of a word adjacent to the entity.	RON: capitala <i>țara</i> Franța → capitala <i>țării</i> Franța	Case system
AGR_CHG	Change gender, number, or person agreement on a modifier or verb.	ITA: Roma <i>è stato fondato</i> → Roma <i>è stata fondata</i>	Gender / number agreement
CONTR_CHG	Add, remove, or change a contraction triggered by the entity.	FRA: <i>de le Louvre</i> → <i>du Louvre</i>	Article–preposition fusion
ELISION	Elide a vowel before a vowel-initial entity.	ITA: <i>de Italie</i> → <i>d’Italie</i>	Vowel-hiatus phonotactics
PUNC_CHG	Fix punctuation or spacing around the entity.	FRA: la distance entre Paris et Rome? → la distance entre Paris et Rome ?	Orthographic convention
OTHER	Any other minor fix (e.g., diacritic correction, compound-split, stray casing) that does not fit the categories above.	Language specific	—

Table 10: **The 13-category taxonomy of instantiation-artifact edits for the post-editing ensemble.** *Operation*: the acronym the models output; *Description*: the explanation given to the model; *Example*: a minimal before/after pair in a language where the edit is typical; *Typological trigger*: the morpho-syntactic feature prompting the edit.

System Prompt. Fig. 23 shows the system prompt for the ensemble LLMs. The change operations ({_TAXONOMY_BLOCK}) are generated from Table 10. Placeholders such as {{entity_A}} mark proper nouns to be preserved verbatim.

User Prompt Template. Fig. 24 shows the per-question user prompt, specifying the target-language ISO 639-3 code, the original English template, and the instantiated question.

Expected Output Scheme. Each model returns strict JSON formatted to the required number of edits (25). Per-language prompts and model-specific parameters are in the code release.

D Model Details

Table 11 summarizes the model configurations used in evaluation, post-editing, and annotation.

E Inference Setup

We serve open-weight models with vLLM (Kwon et al., 2023) on one 141 GB H200, and query Gemini-3-Flash through its official API. All conditions use greedy decoding (T = 0, top-p = 1). For T3 runs, we cap each question at 10 agent steps with a 4096-token budget per step.¹³ Qwen3.5-27B’s T3a/T3c runs extend context to 64K via YaRN (Peng et al., 2024) to accommodate cumulative tool outputs; other variants use their native 32K window. We constrain all model outputs to the per-question gold answer-type JSON schema, separating format compliance from spatial reasoning.

¹³Steps are code generation/execution rounds, not individual tool calls. Across T3 variants (*small* split, English, one seed), correct answers average 4.3 tool calls vs. 6.5 for incorrect ones, suggesting extra steps often signal non-convergence.

Model Name	Huggingface / API Identifier	# Parameters	Provider	Cutoff	Serving Notes
Gemini-3-Flash	gemini-3-flash-preview	n/d	Google	Jan 2025	API; thinking minimal/high (T1/T2)
Gemma-3-27B-Instruct	google/gemma-3-27b-it	27B	Google	Aug 2024	vLLM; bfloat16
Qwen3.5-27B	Qwen/Qwen3.5-27B-FP8	27B	Alibaba Cloud	n/d	vLLM; FP8; YaRN 64K context (T3a/T3c)
Qwen3.5-35B	Qwen/Qwen3.5-35B-A3B-FP8	35B (3B active)	Alibaba Cloud	n/d	vLLM; FP8; sparse MoE
DeepSeek-v4-Flash	deepseek-v4-flash	284B (13B act.)	DeepSeek AI	n/d	API; sparse MoE
Claude-Opus-4.7	claude-opus-4-7	n/d	Anthropic	Jan 2026	–

Table 11: **Models used across evaluation, post-editing, and annotation.** “n/d” = not disclosed (closed model) or not on the model card (Qwen3.5 released Feb 2026; DeepSeek v4 released Apr 2026). Cutoff is the training-data/knowledge cutoff per the model card.

Answer type	Distance d_i	Tolerance τ_i
<i>Discrete</i>		
Boolean / ternary (T/F/plausibly)	$\mathbb{1}[\hat{a} \neq a^*]$	exact ($\tau=1$)
Entity name (single)	$\mathbb{1}[\text{Resolve}(\hat{a}) \notin G^*]^a$	exact ($\tau=1$)
Predefined / cardinal direction	$\mathbb{1}[\hat{a} \neq a^*]$	exact ($\tau=1$)
Set enumeration	$1 - \text{Jaccard}(\hat{A}, A^*)$	exact set ^b
<i>Grid-cell</i>		
H3 / S2 cell index	$\mathbb{1}[\hat{a} \neq a^*]$	exact ^c
Geohash string	$\mathbb{1}[\hat{a} \neq a^*]$	exact
Composite (H3+S2)	$\mathbb{1}[\hat{a} \neq a^*]$ (component-wise)	exact ^c
<i>Numeric</i>		
Distance	$ \hat{y} - y^* $	$\max(\delta, \alpha y^*)$, $\delta=0.5 / 5 / 10$ km
Area	$ \hat{y} - y^* $	$\max(\delta, \alpha y^*)$, $\delta=0.25 / 25 / 100$ km ²
Count	$ \hat{y} - y^* $	$\max(1, \alpha y^*)$
Ratio (dimensionless)	$ \hat{y} - y^* $	$\alpha y^* $ (relative, $\delta=0$)
<i>Geometric / spatial</i>		
Point coordinates	geodesic d_{geo}	$\max(\delta, \alpha d_{\text{scale}})$, $\delta=5$, $d_{\text{scale}}=200$ km
Polar (angle, distance)	angular + radial error	5° angular and radial within numeric τ
Polygon / bounding box	$1 - \text{IoU}$	$\text{IoU} \geq 0.5^b$ (EPSG:6933)
<i>Temporal</i>		
Date	$ \hat{d} - d^* $ days	1 / 15 / 365 days (day/month/year)

^a G^* is the gold-acceptable answer set (including tied entities).

^b Exceptions to $E_{t_i} = \mathbb{1}[d_i < \tau_i]$: $E_{t_i} = 1$ requires exact set equality ($e_{t_i} = 1 - \text{Jaccard} / \text{IoU} \geq 0.5$) ($e_{t_i} = 1 - \text{IoU}$).

^c For H3 we additionally credit resolution-agnostic spatial agreement (predicted cell’s ancestor at the gold resolution matches the gold cell).

Table 12: **Per-answer-type distance d_i and tolerance τ_i used to compute EM and NE.** $E_{t_i} = \mathbb{1}[d_i < \tau_i]$ and $e_{t_i} = \min(1, d_i/\tau_i)$ (so $\text{EM} = \frac{1}{N} \sum_i E_{t_i}$ and $\text{NE} = \frac{1}{N} \sum_i e_{t_i}$), except where marked. Exact-match types use $d_i = \mathbb{1}[\hat{a} \neq a^*]$ with $\tau_i=1$. For numeric answers $\tau_i = \max(\delta, \alpha|y^*|)$, where $\alpha = 0.05$ and δ is a magnitude-dependent floor: distances use $\delta = 0.5 / 5 / 10$ km for urban (< 50 km), regional (50–500 km) and national (≥ 500 km) scales; areas use $\delta = 0.25 / 25 / 100$ km² over the analogous bands (< 2500 , 2500–250000, ≥ 250000 km²); counts $\delta = 1$; ratios $\delta = 0$. Coordinates use $\tau_i = \max(\delta, \alpha d_{\text{scale}})$ with $\delta = 5$, $d_{\text{scale}} = 200$ km; dates use precision-dependent day tolerances.

F Evaluation Metrics

Table 12 gives the per-answer-type distance d_i and tolerance τ_i used to score each of the 15 answer formats.

G Additional Results

G.1 Representativeness

We conduct the majority of the analyses in §5 on the *small* split of MultiGlobeQA, which is a strict subset of the *large* one. Table 13 compares the

two splits across all metrics for the *true-premise* questions: differences in EM are at most 0.8 points before and 2.2 points after re-weighting the small split to the large split’s spatial-function distribution. The majority baseline is likewise nearly identical (28.4% large against 28.8% small).

G.2 Abstention

Fig. 10 plots each model’s false-refusal rate per spatial-function family against how hard that family is. We measure difficulty as the accuracy mod-

	Qwen3.5-27B					Qwen3.5-35B					Gemma-3-27B-Instruct					Gemini-3-Flash				
Split	EM	NE	Cov.	EM _{cov}	FRR	EM	NE	Cov.	EM _{cov}	FRR	EM	NE	Cov.	EM _{cov}	FRR	EM	NE	Cov.	EM _{cov}	FRR
large	4.7	95.4	21.4	21.8	78.4	2.8	97.2	7.1	39.3	92.9	5.6	95.0	66.4	8.4	33.6	23.2	77.7	88.9	26.1	5.9
small	4.5	95.8	26.4	17.0	73.2	2.1	98.0	6.6	31.2	93.4	4.8	95.8	69.7	6.9	30.3	22.6	78.6	93.7	24.1	6.3
Δ	-0.2	+0.4	+5.1	-4.8	-5.2	-0.7	+0.8	-0.5	-8.1	+0.5	-0.8	+0.8	+3.3	-1.5	-3.3	-0.6	+0.9	+4.8	-1.9	+0.4
Δ^{rw}	+1.1	-1.0	-0.8	+5.9	+0.8	+0.7	-0.7	+0.3	+3.3	-0.3	+0.7	-0.7	+0.6	+1.0	-0.6	+2.2	-2.1	+5.8	+0.9	-0.6

Table 13: **Representativeness of the *small* MultiGlobeQA split** (English, text modality, Tier 1, single seed; EM $\uparrow$, NE $\downarrow$, FRR $\downarrow$). Metrics are computed over *true-premise* questions only, 41,540 of 45,114 for the large split and 4,979 of 5,636 for the small one. Δ is *small* – *large*, and Δ^{rw} is the same difference after re-weighting each small-split per-function rate by the large split’s distribution of spatial functions, which the two splits do not share.

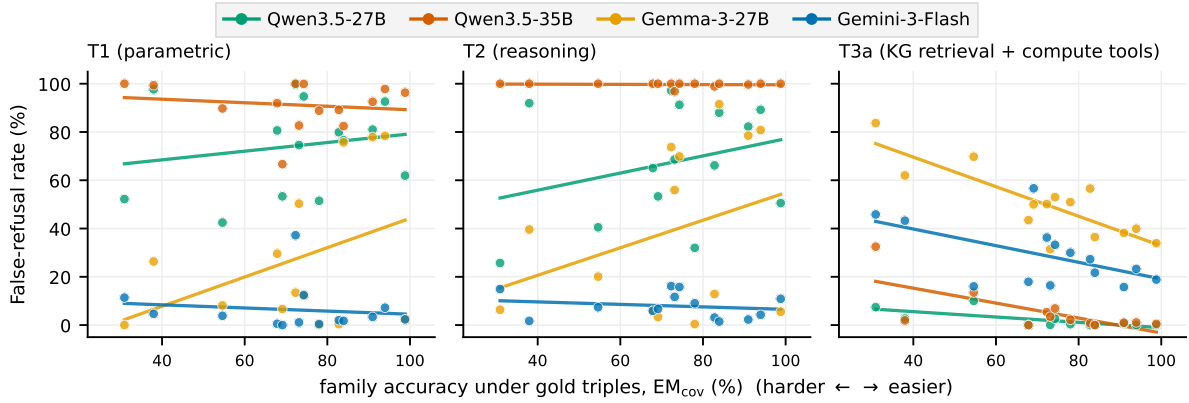


Figure 10: **False-refusal rate against spatial-function family difficulty** (*small* split, English, *true-premise* questions, three seeds). Difficulty is measured as the accuracy that family permits under gold triples (EM_{cov} at T3-o, averaged over models). Lines are least-squares fits shown for orientation; reported coefficients are Spearman.

els reach on the questions they attempt when gold triples are injected (EM_{cov} under T3-o, averaged over the four LLMs), from 30.9% for grid indexing to 98.8% for coordinate questions. Because this ordering is computed once, from a different condition, and excludes refused items, it shares no denominator with the refusal rates plotted against it. Under KG retrieval all four models slope downward (ρ from -0.40 to -0.69), declining most on the families that remain hard even with gold facts. Without retrieval the sign is inconsistent (-0.25 to $+0.30$): Gemma-3-27B-Instruct refuses most on the easiest families, Gemini-3-Flash rarely refuses at either tier, and Qwen3.5-35B refuses 97–100% of true-premise questions at T2 across all 14 families. Ordering families by oracle EM instead, which counts refusals as incorrect, ranks them almost identically ($\rho = 0.82$).

G.3 Oracle Context Representation

In T1-o, we inject the gold triples in three surface forms: raw triples, structured JSON, and verbalized prose (§4). Fig. 11 shows the results. We find that verbalized prose is best for the two mid-size open-

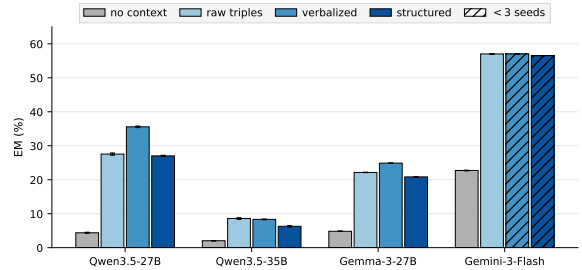


Figure 11: **Oracle ablation (T1)** (*small* split, English, text modality; EM over *true-premise* questions, refusals counted incorrect; mean $\pm$ std. dev. over three seeds).

weights LLMs, by 8.6 EM points over structured JSON for Qwen3.5-27B and 4.1 for Gemma-3-27B-Instruct. The choice is consequential only for the open-weight models, where the three formats span 2.3 to 8.6 points, against 0.5 for Gemini-3-Flash, so we use verbalized prose throughout.

G.4 Compute Cost

Table 14 and Fig. 12 report median tokens, tool calls and tokens per correct answer for every model and condition. Web search is the most expensive

Model	EM	#tokens/question	#tokens/correct answer	#tool calls	cap hit (%)	Context overflow (%)
<i>T1 (parametric)</i>						
Qwen3.5-27B	4.4	134	3.1k	0.0	—	0.0
Qwen3.5-35B	2.0	125	6.1k	0.0	—	0.0
Gemma-3-27B-Instruct	4.8	139	2.9k	0.0	—	0.0
Gemini-3-Flash	22.7	193	850	0.0	—	0.0
<i>T2 (reasoning)</i>						
Qwen3.5-27B	3.8	308	8.0k	0.0	—	0.0
Qwen3.5-35B	0.2	295	128k	0.0	—	0.0
Gemma-3-27B-Instruct	3.8	264	7.0k	0.0	—	0.0
Gemini-3-Flash	24.6	232	943	0.0	—	0.0
<i>T3a (KG retrieval)</i>						
Qwen3.5-27B	44.3	56k	127k	6.7	43.6	1.7
Qwen3.5-35B	29.3	55k	186k	7.5	58.6	8.2
Gemma-3-27B-Instruct	23.9	22k	94k	4.8	5.5	3.1
Gemini-3-Flash	29.5	21k	70k	2.9	1.1	0.0
<i>T3b (web search)</i>						
Qwen3.5-27B	22.6	85k	374k	8.6	67.1	4.7
Qwen3.5-35B	22.8	88k	387k	8.4	62.3	8.5
Gemma-3-27B-Instruct	21.7	30k	138k	5.6	11.5	4.3
Gemini-3-Flash	27.2	63k	230k	5.8	19.5	0.0
<i>T3c (KG + web)</i>						
Qwen3.5-27B	44.1	61k	139k	6.8	45.5	1.4
Qwen3.5-35B	37.2	65k	174k	7.1	49.2	10.4
Gemma-3-27B-Instruct	31.5	33k	104k	5.8	12.4	5.1
Gemini-3-Flash	28.0	22k	78k	2.2	0.8	0.0
<i>T3-o (gold triples injected)</i>						
Qwen3.5-27B	61.6	8.9k	14k	2.5	6.0	8.9
Qwen3.5-35B	55.0	9.0k	16k	2.3	4.5	19.6
Gemma-3-27B-Instruct	45.3	4.7k	10k	2.6	16.2	16.2
Gemini-3-Flash	60.2	13k	22k	2.1	2.3	1.2

Table 14: **Compute costs** (*small* split, English, text modality). EM is over *true-premise* questions. Cost columns are medians over the answered questions, excluding harness errors, whose share is given in the last column; tokens per correct answer is the median token count divided by EM, and *cap hit* the share of questions where the agent exhausts its step budget. Tool calls and cap hit do not apply without an agent (—).

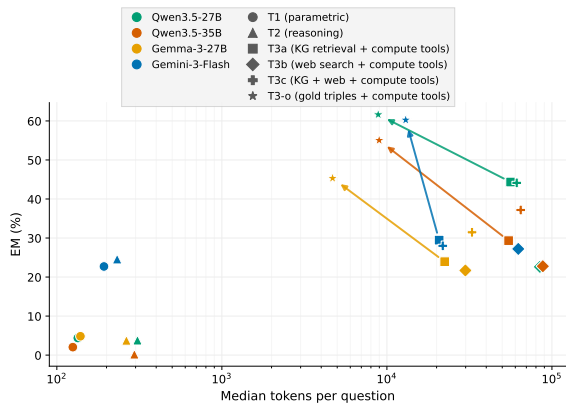


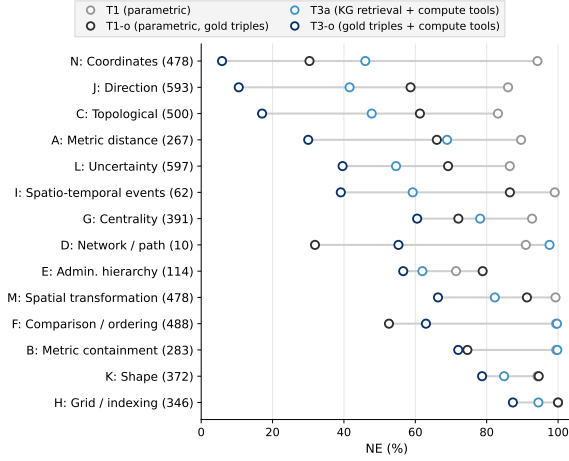
Figure 12: **Accuracy vs. compute cost** (*small* split, English, text modality; EM over *true-premise* questions against median tokens per question, log scale). Color denotes model, shape denotes tier. Arrows run from T3a to T3-o.

source at 30k-88k median tokens, and adding it to

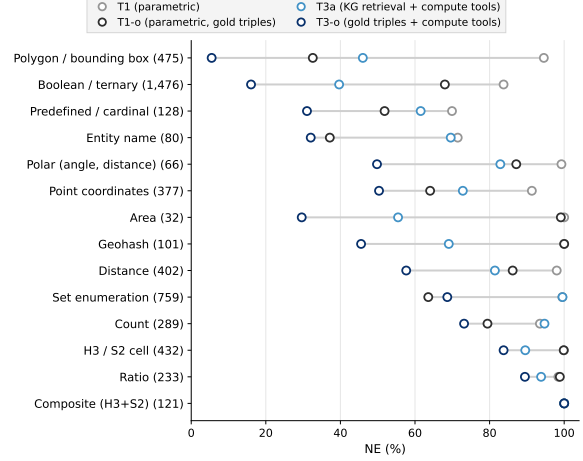
the KG (T3c) raises cost over KG retrieval alone (T3a) while improving accuracy only for Gemma-3-27B-Instruct and Qwen3.5-35B. The oracle runs the same agent with the same tools and step budget, differing only in whether the evidence must be retrieved, and it is cheaper for all four models, by $6.3\times$ (Qwen3.5-27B), $6.1\times$ (Qwen3.5-35B), $4.8\times$ (Gemma-3-27B-Instruct) and $1.6\times$ (Gemini-3-Flash), while scoring 17 to 31 points higher. Tool calls fall from 2.9-7.5 per question to 2.1-2.6. The agent therefore spends its budget on search that does not terminate in the evidence, rather than on longer reasoning over evidence it has found.

G.5 Performance by SFID and Answer Format

Fig. 13 reports NE by spatial function and answer format, and Figs. 14 - 17 break both EM and NE down by model.



(a) NE by spatial function.



(b) NE by answer format.

Figure 13: **Normalized error (NE) by spatial function (a) and answer format (b) across evaluation conditions** (*small* split, English, text modality; averaged over the four models). Parentheses give the number of *true-premise* questions in each family.

G.6 Income & Density Equity

Fig. 18 repeats the equity analysis from §5.3 over the three population-density tiers. On the *small* split the high-minus-low gap is positive in all twelve model condition cells (+0.2 for Qwen3.5-35B at T1 to +5.6 for Gemini-3-Flash at T3-o) and monotone in six, which would suggest that densely mapped regions are better served. However, this does not replicate: at T1 on the *large* split, which carries 41,540 questions against 4,979 and a more balanced composition, the same gaps fall to within ± 0.2 for Qwen3.5-27B, Qwen3.5-35B and Gemini-3-Flash, and +1.1 for Gemma-3-27B-Instruct. Since the *small* split samples up to 200 questions per (KG, SFID) cell, its density composition differs from the benchmark’s. The income gap does replicate on the *large* split, staying positive for all four models on both splits (+1.1 for Qwen3.5-35B to +3.1 for Gemini-3-Flash on *large*; Fig. 3 vs. Fig. 19).

G.7 False-premise Questions

False premises occupy 13 of the 124 sub-templates in the *small* split, and models refuse far more often on those than on the benchmark as a whole: Qwen3.5-27B declines 13.6% of their answerable questions at T3a against 1.9% overall (Fig. 20). A false-refusal rate computed over the full benchmark would therefore overstate selectivity, so we restrict the comparison to answerable questions from the same sub-templates.

G.8 Multilinguality

Fig. 21 plots each language’s EM relative to English, for all four models at T1 (Fig. 21a) and for Qwen3.5-27B across tiers (Fig. 21b). English anchors the axis at 2.1 to 22.6 EM at T1 and at 61.5 for Qwen3.5-27B at T3-o, so equal deltas are not comparable across tiers. The spread is widest at T1-o (8.3 points against 4.5 at T3a and 3.9 at T3-o). The injected triples are in English regardless of the question language, and T1-o is the only condition in which the model must read them directly rather than compute over them, which suggests the sensitivity lies in processing a cross-lingual context rather than in the question itself.

G.9 Multimodal Slice

Table 15 reports performance across the tiers on the multimodal slice of *small* split. Replacing the entity name with an image lowers the ceiling: the best model reaches 54.6 on the 269 answerable image questions against 61.6 on text. Gemini-3-Flash again leads at T1 by a wide margin (39.8 vs. 3.1 to 7.7), as it does on text. Nearly half the slice is topological, so it probes visual entity grounding more than the spatial functions the text benchmark isolates.

H Prompts

H.1 Benchmark Construction Prompts

Fig. 22 shows the prompt for the judge in the annotator-disagreement resolution pipeline; Figs. 23-25 the prompts for the post-editing ensemble.

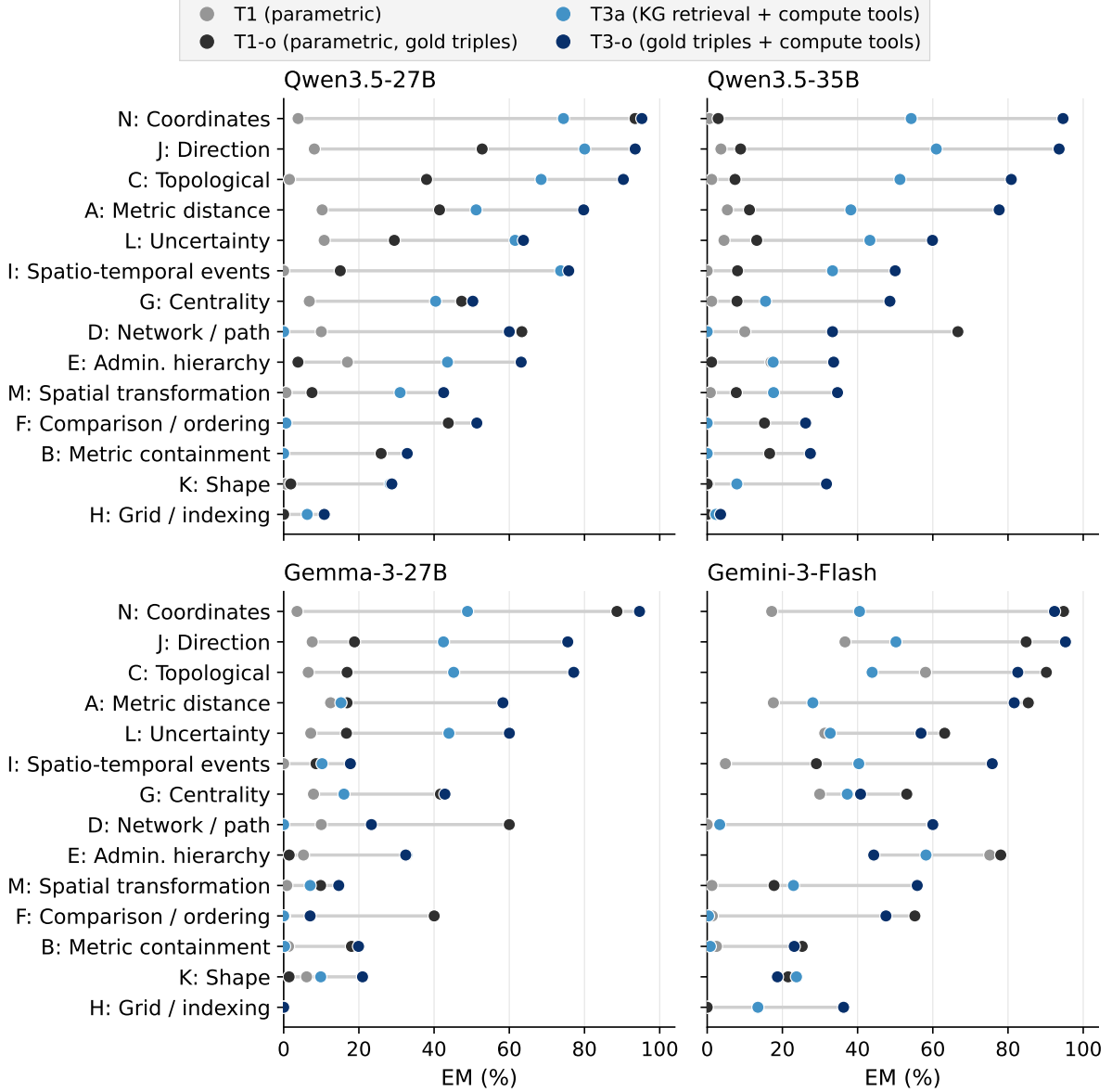


Figure 14: **Accuracy (EM) by spatial function and model across evaluation conditions** (*small* split, English, text modality). Parentheses give the number of *true-premise* questions in each family.

Model	T1	T2	T1-o	T3a	T3-o
Qwen3.5-27B	7.7	5.0	19.2	25.2	50.6
Qwen3.5-35B	3.5	0.0	2.6	17.3	32.6
Gemma-3-27B-Instruct	3.1	4.1	4.3	34.8	18.2
Gemini-3-Flash	39.8	46.5	45.7	33.6	54.6

Table 15: **Accuracy (EM) on the multimodal slice** (*small* split, English; EM over the 269 *true-premise* image questions).

H.2 Model Prompts

We use a tier-specific template per evaluated model. The templates share the task framing, question, and expected answer type, with the following placeholders substituted per item at inference:

- `$question`: the question text (English or target language);
- `$answer_type_description`: a natural language description of the expected answer type (e.g., *a distance*, *a set of entity names*);
- `$unit_clause`: an appended unit hint for numeric answers (e.g., *in kilometers*);
- `$distance_clause`: an optional scale hint for distance answers;
- `$tools_section`: the callable tools available to the agent (retrieval and geospatial functions), for the agentic tiers;

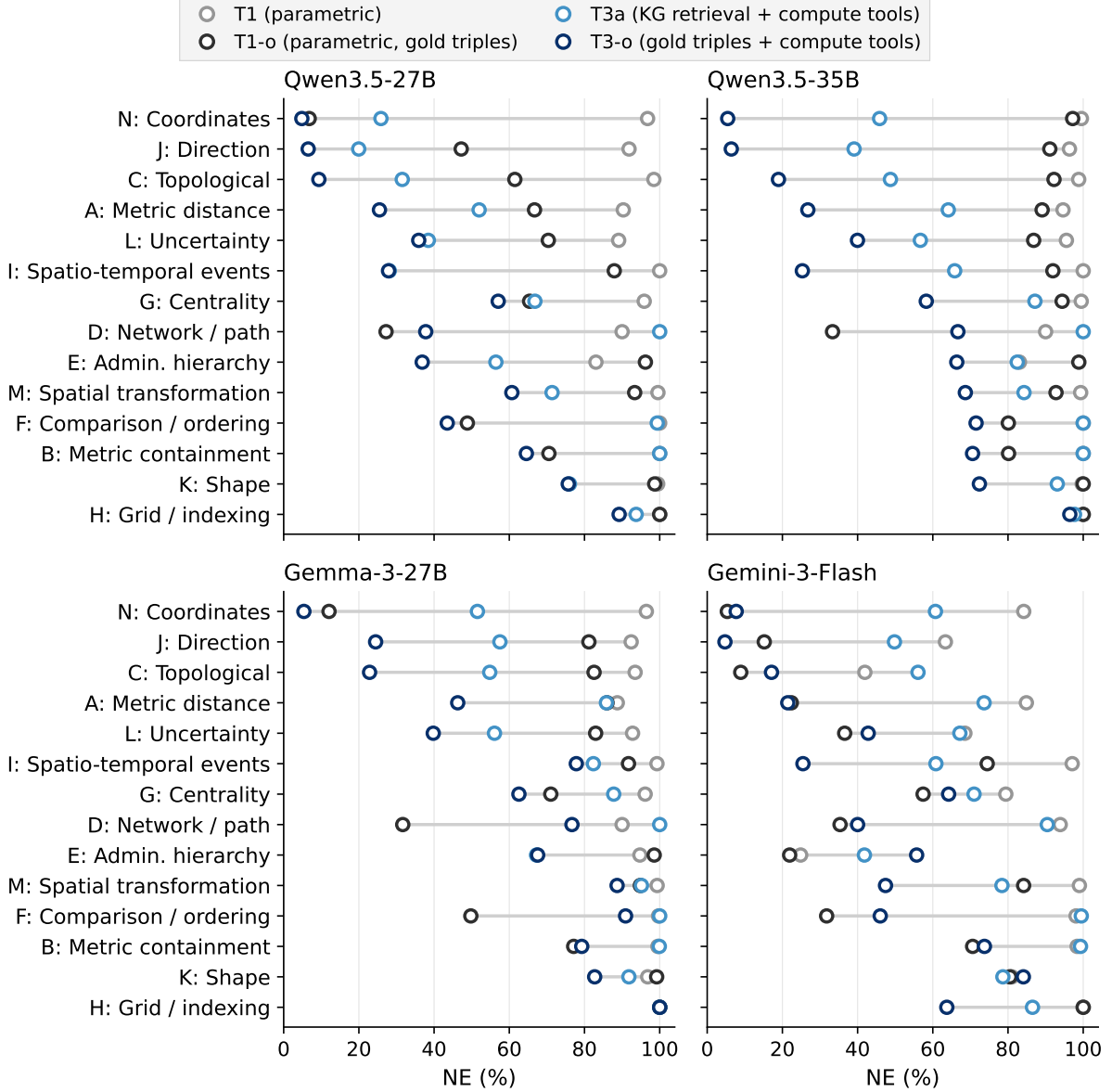


Figure 15: **Normalized error (NE) by spatial function and model across evaluation conditions** (*small* split, English, text modality). Parentheses give the number of *true-premise* questions in each family.

- `$triple_context`: the injected gold KG facts, for the oracle conditions;
- `$max_tool_calls`: the agent step budget per question (10).

Tiers 1 and 2 return a single JSON object (Figs. 26-30). The agentic tiers (T3) instead require Python code terminating in a `final_answer` call. The three T3 variants share the template in Fig. 31, differing only in the first-sentence retrieval source – *a spatial knowledge graph* (T3a), *web search* (T3b), or *a spatial knowledge graph, web search* (T3c) – and in the `$tools_section`: T3a receives the KG tools, T3b the web tools, and T3c both. Tools are

exposed as callable Python functions in a shared sandbox of pre-imported geospatial libraries. We abbreviate their descriptions in Fig. 32.

For Tier 1 oracle conditions (Fig. 27-29), we present the gold triples in one of three verbalizations: *raw* RDF triples, *structured* triples, or *verbalized* natural-language facts. For Tier 3 oracle, we inject the triples, but provide no retrieval tools (Fig. 33); the model parses values from the context and computes the answer in Python.

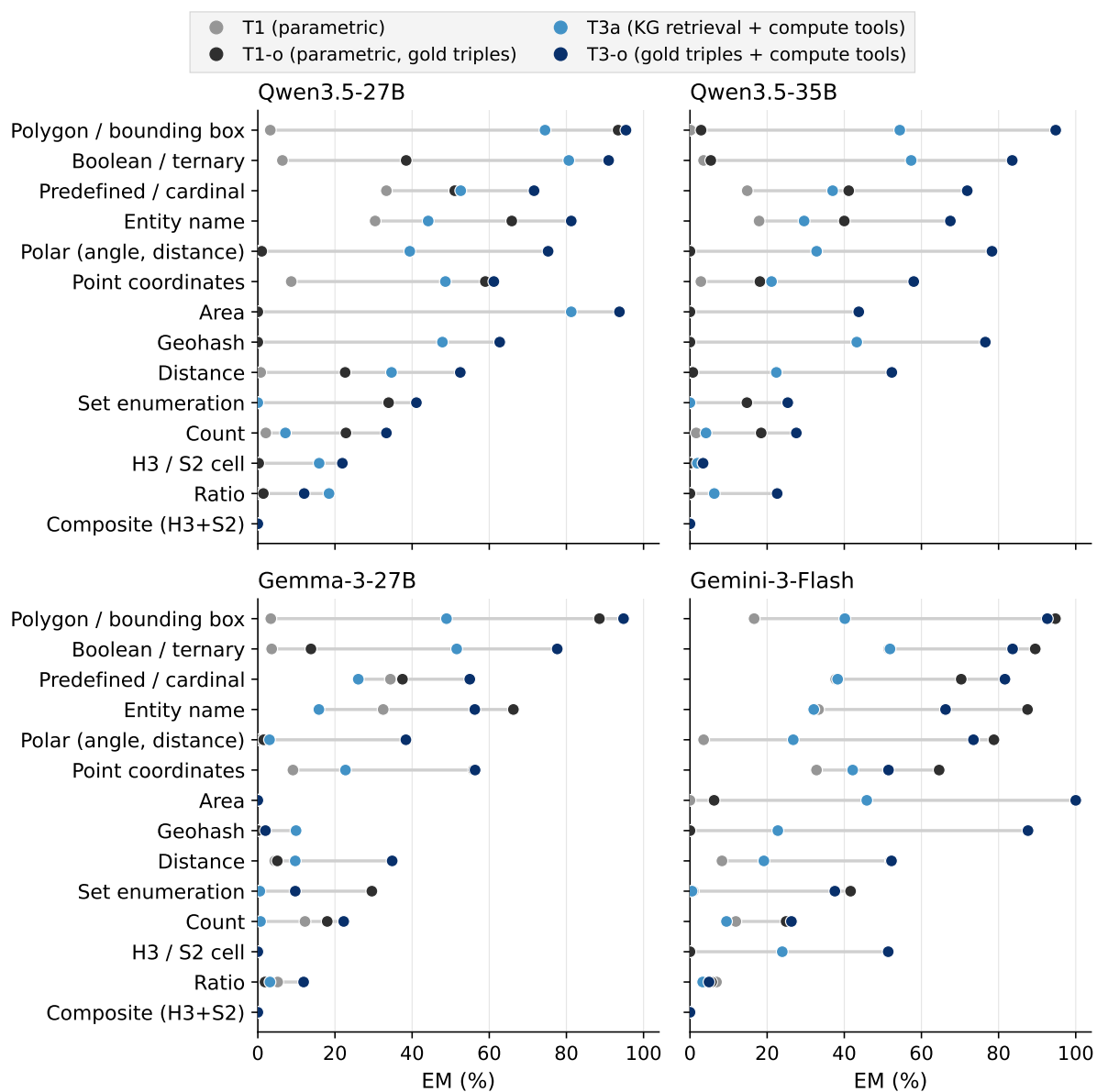


Figure 16: **Accuracy (EM) by answer format and model across evaluation conditions** (*small* split, English, text modality). Parentheses give the number of *true-premise* questions in each family.

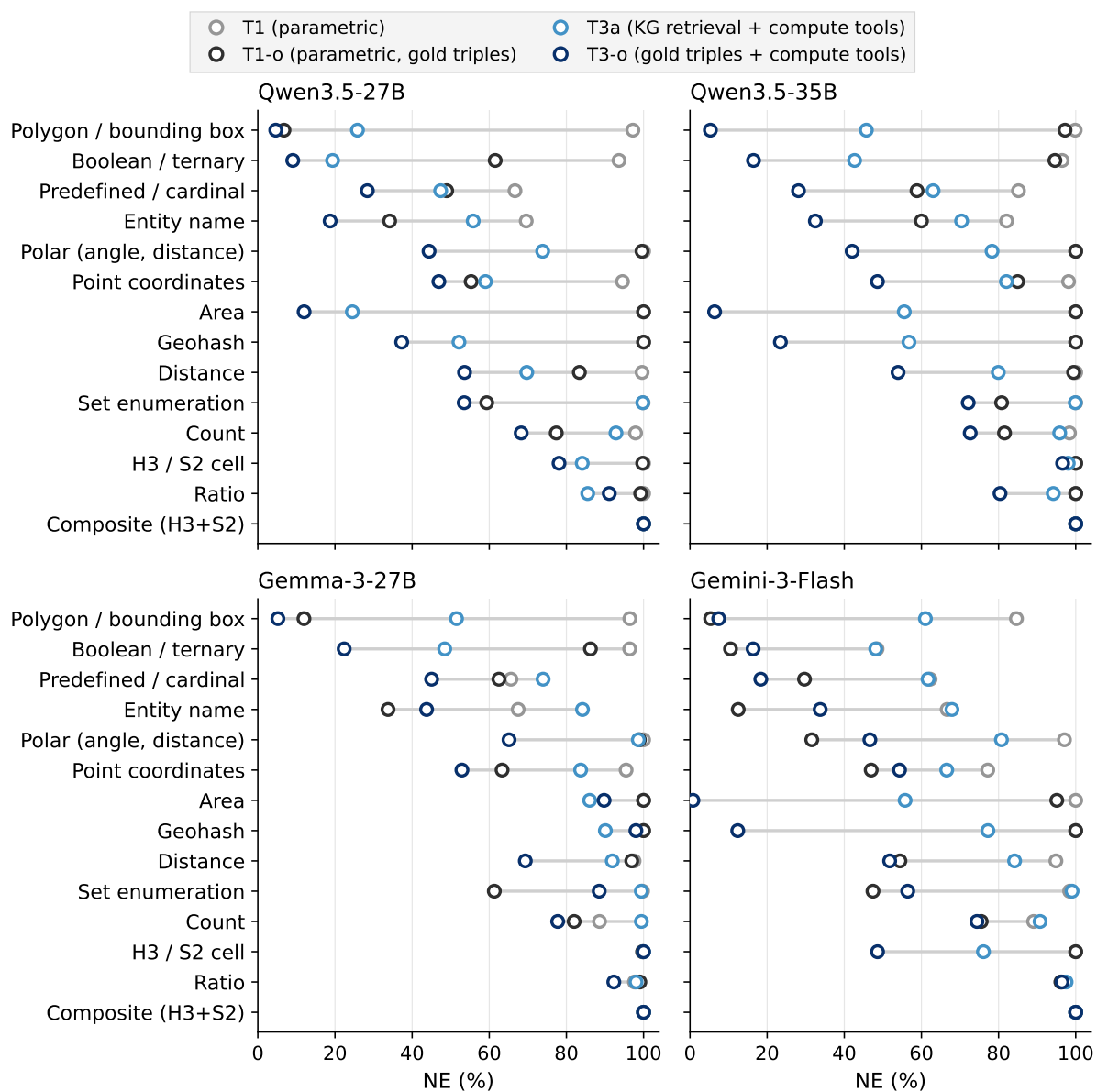


Figure 17: **Normalized error (NE) by answer format and model across evaluation conditions** (*small* split, English, text modality). Parentheses give the number of *true-premise* questions in each family

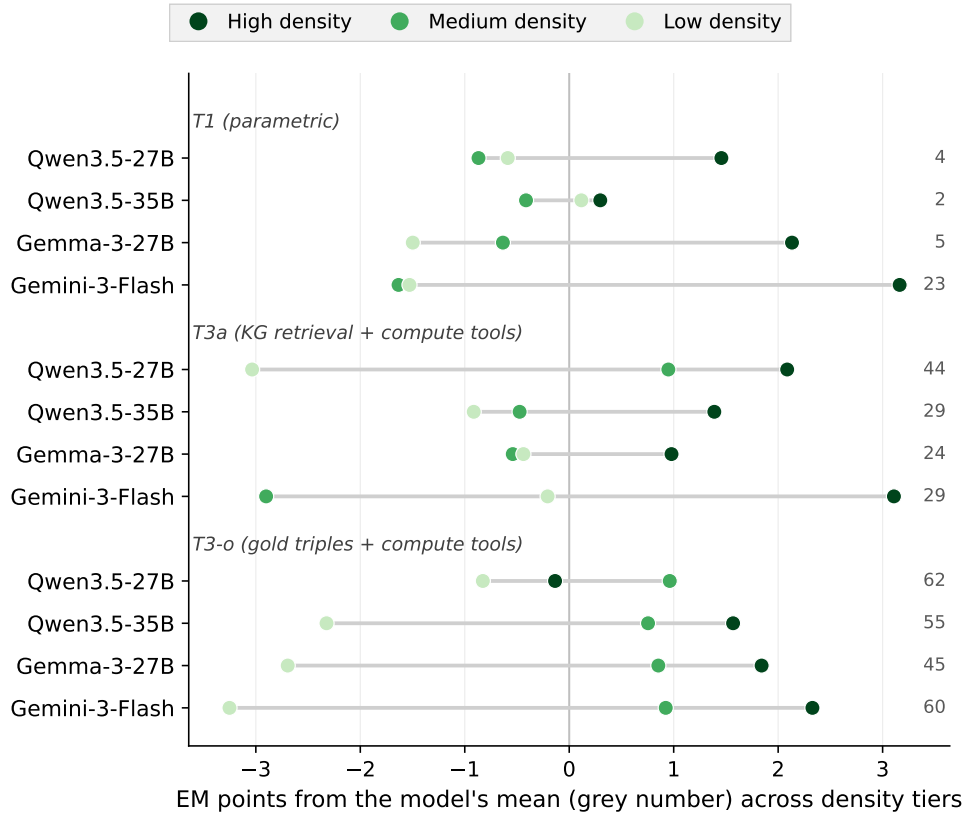


Figure 18: **Accuracy (EM) by population-density tier** (*small* split, English, text modality; standardized EM). Each row is one model under one condition, with every tier placed at its distance from that model's mean across tiers. Standardization re-weights each tier to the benchmark-wide spatial-function distribution while holding its own per-family accuracy fixed, because the strata equalize question counts but not question composition.

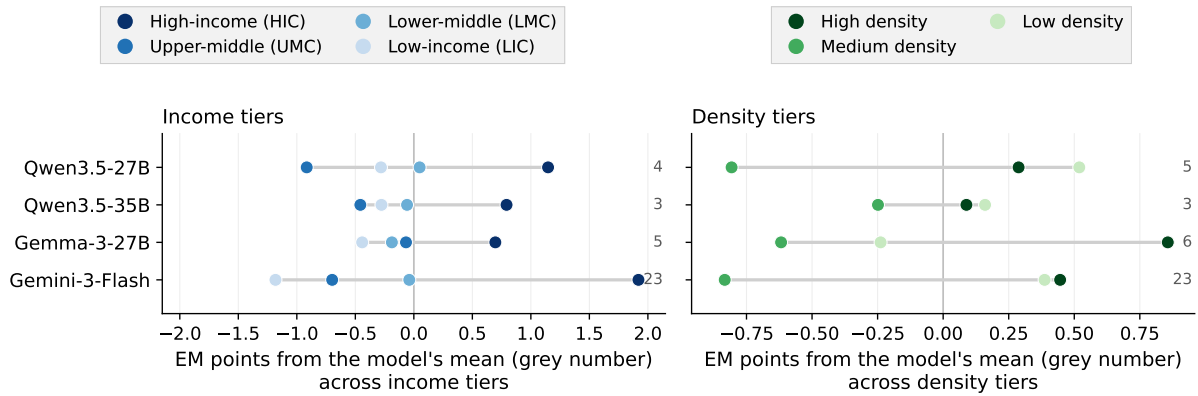


Figure 19: **Accuracy (EM) by income and density tier on the large split** (English, text modality, T1, one seed).

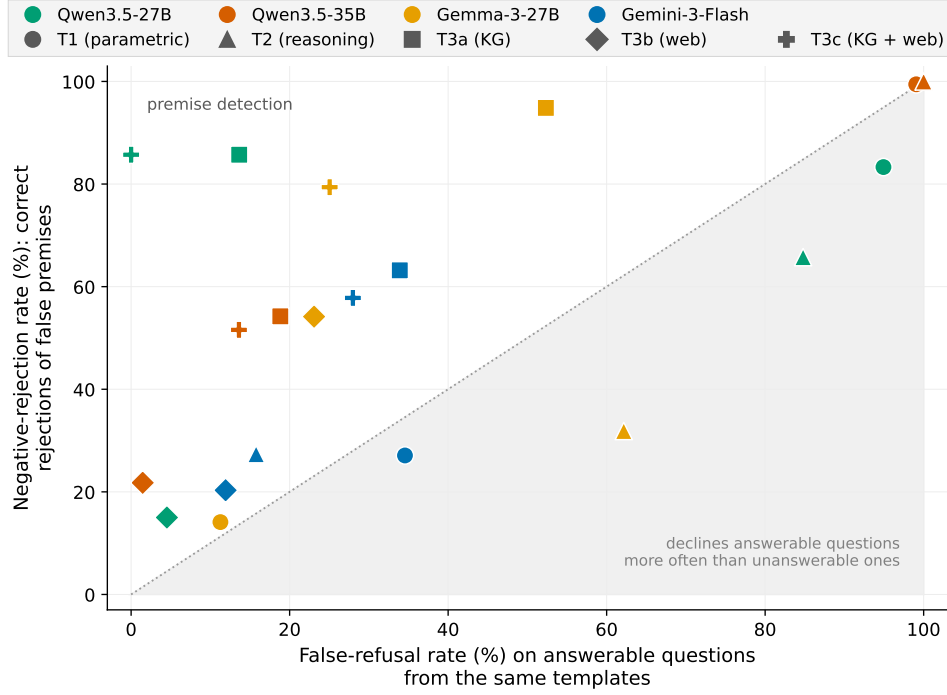


Figure 20: **Premise detection against indiscriminate abstention** (*small split, English, text modality*). Color denotes model, shape denotes tier. False refusal is measured on answerable questions from the 13 templates that carry false premises, not on the whole benchmark. The dotted line is chance.

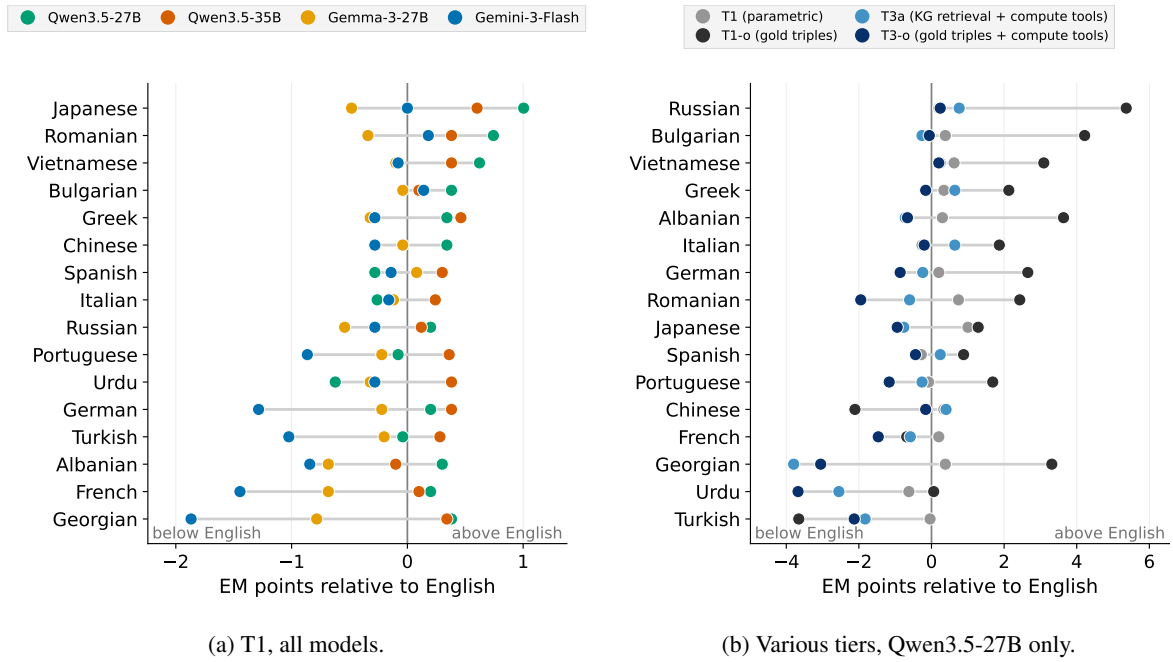


Figure 21: **Per-language accuracy relative to English** (*small split, text modality; EM over true-premise questions*).

Prompt: LLM Judge for Annotator-disagreement Resolution

TASK

You are arbitrating between two human annotators who post-edited a machine translation. For each item, you **must** pick one of the two annotator corrections – you are choosing which human edit the benchmark will use.

The source is English; both annotators corrected a machine translation into **{target_language}**. Neither annotator is a ground-truth source, but the benchmark policy is to use one of **their** edits. You are **not** a third annotator: even if you would prefer a different phrasing, suppress that and pick the better (or less wrong) of the two humans' corrections.

INPUT FORMAT

For each item you receive:

- ``en_template``: the English question template (contains placeholders like ``{entity_A}``, ``{value}``).
- ``original_translation``: the machine-generated translation into `{target_language}`. This is the text the annotators were asked to evaluate.
- ``correction_ann1``: annotator 1's rewrite (may be empty).
- ``correction_ann2``: annotator 2's rewrite (may be empty).

Placeholders are strings in curly braces (``{entity_A}``, ``{value_1}``, etc.). They must appear verbatim and unchanged in the chosen correction.

DECISION CRITERIA

Rank the two corrections using this priority, top-first. Only move to the next criterion if the current one does not decide. These criteria are **relative**: you are comparing the two humans' edits against each other, not against an ideal. If both have issues on a criterion, pick the one with fewer or smaller issues.

1. **Placeholder integrity.** Placeholders must be present verbatim, in matching quantity to the English template. A correction that renames, merges, splits, or loses a placeholder is worse than one that keeps them intact.
2. **Semantic fidelity.** The spatial relation, quantifier, and question type (yes/no vs wh-) must match the English. Prefer the correction that preserves meaning more faithfully.
3. **Grammatical correctness.** Agreement, case, verb form, and word order must be valid `{target_language}`. The correction with fewer grammar errors wins.
4. **Technical terminology.** For domain terms (centroid, variance, fractal dimension, perimeter, etc.), prefer the correction that uses established technical vocabulary over approximations.
5. **Natural phrasing.** Among otherwise-equal options, prefer the more idiomatic phrasing a native speaker would write.

WHAT TO IGNORE

- **Gender or case agreement** that depends on the entity that fills a placeholder. Do not prefer one correction solely because it adds a gender- or case-specific possessive / article / inflection before or after a placeholder. These variations are handled downstream. A correction that hard-codes one gender for a placeholder is slightly **worse**, not better.
- **Whitespace and trivial punctuation differences.** Extra spaces, missing trailing punctuation, and ``{ entity }`` vs ``{entity}`` are minor – do not let them drive the pick.
- **Length.** A shorter correction is not automatically better, and a longer one is not automatically more thorough.
- **Register (formal vs informal).** The benchmark text is not addressed to a specific person. Interrogatives are naturally impersonal; imperatives can be written in any consistent neutral register (e.g., German ``Sie`` vs ``du``, French ``vous`` vs ``tu``, Italian ``Lei`` vs ``tu``, Japanese ``desu/masu`` vs plain form). Do not prefer one correction over another solely because it uses a different register – only penalise register choices that are internally inconsistent within the correction itself.

`{language_specific_notes}`

EDGE CASES

- **Both corrections are identical** after trimming whitespace → output ``equivalent``.
- **Both corrections are acceptable** and neither is clearly better after applying the criteria above → output ``equivalent``. Do not pick arbitrarily.
- **Both corrections have real issues** → pick the one that is **less wrong** by the decision criteria. You must still output ``ann1`` or ``ann2``. "Both are bad" is not an allowed conclusion – the policy is to ship one human edit per item.

OUTPUT FORMAT

For each item, return one JSON object on its own line with exactly these keys: `{"item_id": "<echoed from input>", "pick": "ann1" | "ann2" | "equivalent", "rationale": "<one short sentence, under 20 words, naming the decisive criterion>"}`

Only these three values for ``pick`` are allowed. No other output. No markdown fences. No prose between JSON lines. Do not include the chosen text – the pipeline looks it up from the label.

Figure 22: System prompt shared across the post-editing LLM ensemble. The prompt is abbreviated: we omit the examples section due to space constraints. The full prompt is available in the project's repository.

Post-editing Ensemble: System Prompt

You are a grammar corrector for spatial-reasoning benchmark questions. You receive a TEMPLATE (with placeholder variables in {...}) like {{entity_A}} and an instantiated QUESTION where the placeholders have been replaced with real entity names, also wrapped in {...}. The {...} markers in the QUESTION identify proper nouns that MUST be preserved exactly: never modify the text inside {...}, and the braces must remain in your output.

Fix ONLY minor grammar/syntax/fluency issues caused by substituting entity names into the template: articles, prepositions, declension, agreement, elision, contractions, punctuation. Do NOT rephrase. Do NOT change meaning. Do NOT add or remove content words.

CRITICAL RULES:

- Do NOT change the verb form, mood, or tense. Imperative ("Calculate") must stay imperative; interrogative ("What is") must stay interrogative; declarative must stay declarative. Do NOT convert one to another.
- Do NOT remove, alter, or relocate any {...} markers that wrap entity names. Do NOT change the text inside those markers, and do NOT add or remove braces anywhere in the question.
- Do NOT change the question wording. Only fix grammar AROUND the entities. If the question would be perfectly grammatical with the entities removed, the answer is {"needs_edit": false}.
- The test for whether an edit is needed is whether the surrounding grammar broke when the entities were instantiated. If the sentence reads naturally and grammatically with the entities in place, no edit is needed -- even if the wording is not the most stylish.

If the sentence is already correct, return exactly:
{"needs_edit": false}

If corrections are needed, return STRICT JSON in this schema:
{"needs_edit": true, "edits": [{"op": "<OP>", "orig": "<substring>", "fix": "<replacement>"}],
"corrected_question": "<the complete corrected question with {...} markers preserved>"}

Each edit's "orig" is the exact substring of the input question being changed; "fix" is its replacement. If "orig" or "fix" includes {...}, the content inside the braces must appear identically in both -- the braces and their contents are context, not editable.

Allowed ops (use the most specific one that applies; fall back to OTHER):
{_TAXONOMY_BLOCK}

Below are four worked examples. Apply the same logic to the QUESTION you receive, regardless of its language.

EXAMPLE 1 (NO_CHANGE -- imperative form is valid):
LANGUAGE: en
TEMPLATE: Calculate the distance between {{entity_A}} and {{entity_B}}.
QUESTION: Calculate the distance between {{Tarzanseil}} and {{Friedrich-Guenther-Park}}.
RESPONSE: {"needs_edit": false}

EXAMPLE 2 (NO_CHANGE -- French interrogative form is valid):
LANGUAGE: fr
TEMPLATE: Quelle est la distance entre {{entity_A}} et {{entity_B}} ?
QUESTION: Quelle est la distance entre {{Berlin}} et {{Paris}} ?
RESPONSE: {"needs_edit": false}

EXAMPLE 3 (ART_CHG -- feminine entity needs feminine article):
LANGUAGE: fr
TEMPLATE: Quelle est la distance entre le {{entity_A}} et le {{entity_B}} ?
QUESTION: Quelle est la distance entre le {{Tour Eiffel}} et le {{Louvre}} ?
RESPONSE: {"needs_edit": true, "edits": [{"op": "ART_CHG", "orig": "le {{Tour Eiffel}}", "fix": "la {{Tour Eiffel}}"}], "corrected_question": "Quelle est la distance entre la {{Tour Eiffel}} et le {{Louvre}} ?"}

EXAMPLE 4 (ELISION -- vowel-initial entity triggers elision):
LANGUAGE: fr
TEMPLATE: Quelle est la distance de {{entity_A}} a {{entity_B}} ?
QUESTION: Quelle est la distance de {{Italie}} a {{Espagne}} ?
RESPONSE: {"needs_edit": true, "edits": [{"op": "ELISION", "orig": "de {{Italie}}", "fix": "d'{{Italie}}"}], "corrected_question": "Quelle est la distance d'{{Italie}} a {{Espagne}} ?"}

EXAMPLE 5 (ELISION - vowel-initial entity triggers elision):
LANGUAGE: fr
TEMPLATE: Quelle est la distance de {{{entity_A}}} à {{{entity_B}}} ?
QUESTION: Quelle est la distance de {{{Italie}}} à {{{Espagne}}} ?
RESPONSE: {"needs_edit": true, "edits": [{"op": "ELISION", "orig": "de {{Italie}}", "fix": "d'{{Italie}}"}], "corrected_question": "Quelle est la distance d'{{Italie}} à {{Espagne}} ?"}

Output STRICT JSON, no markdown, no explanations.

Figure 23: System prompt shared across the post-editing LLM ensemble.

Post-editing Ensemble: User Prompt

```
LANGUAGE: <iso_code>
TEMPLATE: <base_template_with_{{entity_X}}_placeholders>
QUESTION: <instantiated_question_with_{{entity_name}}_markers>
```

Figure 24: User prompt for the post-editing LLM ensemble.

Expected JSON output schema

```
# No edit needed
{"needs_edit": false}

# Edits proposed
{"needs_edit": true,
 "edits": [
   {"op": "<OP>",           // one of the 13 taxonomy acronyms
    "orig": "<substring>",  // verbatim substring being changed
    "fix": "<replacement>"} // the replacement substring
 ],
 "corrected_question": "<full corrected sentence, {{...}} markers intact>"}
```

Figure 25: Expected JSON output schema for the LLM ensemble models.

Prompt: Tier 1 (parametric)

You are answering a spatial question about real-world locations and entities.

Question:
\$question

Expected answer: \$answer_type_description\$unit_clause
\$distance_clause
Use {"answer": null} to indicate that you do not have the answer.

Respond with a single JSON object containing the "answer" field (and a "unit" field for numeric answers). Do not add any text outside the JSON object.

Figure 26: Tier 1 (parametric) prompt.

Prompt: Tier 1 oracle raw

You are answering a spatial question about real-world locations and entities. You have been given the relevant RDF triples from a spatial knowledge graph.

Relevant facts:
\$triple_context

Question:
\$question

Expected answer: \$answer_type_description\$unit_clause
\$distance_clause
Use {"answer": null} to indicate that you do not have the answer.

Respond with a single JSON object containing the "answer" field (and a "unit" field for numeric answers). Do not add any text outside the JSON object.

Figure 27: Tier 1 oracle (raw) prompt.

Prompt: Tier 1 oracle structured

You are answering a spatial question about real-world locations and entities. You have been given the relevant information from a spatial knowledge graph as structured triples.

Relevant facts:
\$triple_context

Question:
\$question

Expected answer: \$answer_type_description\$unit_clause
\$distance_clause

Use {"answer": null} to indicate that you do not have the answer.

Respond with a single JSON object containing the "answer" field (and a "unit" field for numeric answers). Do not add any text outside the JSON object.

Figure 28: Tier 1 oracle (structured) prompt.

Prompt: Tier 1 oracle verbalized

You are answering a spatial question about real-world locations and entities. You have been given the relevant facts from a spatial knowledge graph.

Relevant facts:
\$triple_context

Question:
\$question

Expected answer: \$answer_type_description\$unit_clause
\$distance_clause

Use {"answer": null} to indicate that you do not have the answer.

Respond with a single JSON object containing the "answer" field (and a "unit" field for numeric answers). Do not add any text outside the JSON object.

Figure 29: Tier 1 oracle (verbalized) prompt.

Prompt: Tier 2 (reasoning)

You are answering a spatial question about real-world locations and entities.

Question:
\$question

Expected answer: \$answer_type_description\$unit_clause
\$distance_clause

Think step by step. Reason through the question carefully, then commit to a final answer.

Use {"reasoning": "<your reasoning>", "answer": null} to indicate that you do not have the answer.

Respond with a single JSON object containing the "reasoning" field and the "answer" field (and a "unit" field for numeric answers). Do not add any text outside the JSON object.

Figure 30: Tier 2 (reasoning) prompt.

Prompt: Tier 3a (agentic)

You are answering a spatial question about real-world locations and entities. You write Python code to solve this using a spatial knowledge graph and geospatial computation. When images are attached as inputs, you may reason over them directly without retrieval.

Question:
\$question

Expected answer: \$answer_type_description\$unit_clause
\$distance_clause

\$tools_section

You have up to \$max_tool_calls steps to solve this. Write ONLY valid Python code. Do NOT include JSON blocks, markdown formatting, or any non-Python text in your code.

When you have computed the answer, call `final_answer` with a dict:
`final_answer({"answer": 12.5, "unit": "km"})`

If you cannot determine the answer, call:
`final_answer({"answer": None})`

IMPORTANT: `final_answer` is a built-in function. Never assign to it (do not write `final_answer = ...`). Always CALL it directly as shown above, e.g. `final_answer({"answer": ...})`. Use a different variable name for the dict if you want to build it first, e.g. `result = {...}; final_answer(result)`.

IMPORTANT: Call `final_answer` at the TOP LEVEL of your code block, NOT from inside a function definition. The agent only registers a final answer when `final_answer` is invoked at the top level. Do NOT do this:

```
def solve():  
    final_answer({"answer": ...})  
solve()
```

Do this instead – write the call at the top level directly:

```
result = ... # compute  
final_answer({"answer": result, "unit": "km"})
```

Figure 31: Tier 3a (agentic) prompt. T3b/T3c differ only in the first-sentence retrieval source: *a spatial knowledge graph* (T3a), *web search* (T3b), or *a spatial knowledge graph, web search* (T3c).

Agent tools (\$tools_section)

Available tools (call as Python functions):

Knowledge-graph tools (T3a, T3c)

- find_entity(name: string) -> string
Search the KG by name; returns a JSON list of up to 5 ranked matches, each {subject, name, matched predicate}.
- find_entities_by_type(category: string, region: string | None) -> string
Search entities whose rdf:type matches a category (e.g. "hospital", "park"), optionally within a region; JSON list of {subject, name, matched_type}.
- get_entity(entity_id: string) -> string
Full record for a KG identifier: JSON {subject, name, names, types, category, geometry_wkt (auto-simplified), geometry_node, raw_triples}.
- list_categories() -> string
Up to 50 rdf:type values present in the KG: JSON {types, shown, has_more}; pass one to find_entities_by_type.

Web tools (T3b, T3c)

- web_search(query: string) -> string
Web search (Google via Serper when available, else DuckDuckGo); returns ranked result snippets.
- visit_webpage(url: string) -> string
Fetch a URL and return its main content as Markdown.

Always available

- final_answer(answer) -> submit your final answer and stop.

Pre-imported libraries you can use in your code:

- geopy (geopy.distance.geodesic for distance; geopy.geocoders.Nominatim for geocoding)
- shapely (shapely.wkt.loads for WKT parsing; contains, intersects, distance, area, centroid)
- pyproj (coordinate projection for area/perimeter in metric units)
- h3 (H3 spatial indexing)
- math, json, re (standard library)

Figure 32: Agent tools for Tier 3.

Prompt: Tier 3 oracle

You are answering a spatial question about real-world locations and entities. You have been given the relevant information from a spatial knowledge graph, and you write Python code to compute the answer using geospatial libraries. When images are attached as inputs, you may reason over them directly without retrieval.

Knowledge graph context:
\$triple_context

Question:
\$question

Expected answer: \$answer_type_description\$unit_clause
\$distance_clause

You do NOT have access to any retrieval tools (no knowledge graph queries, no web search). All the information you need is already provided above as structured triples. Your task is to parse the relevant values from the context and compute the answer in Python.

Pre-imported libraries you can use in your code:

- geopy (geopy.distance.geodesic for distance, geopy.geocoders.Nominatim for geocoding)
- shapely (shapely.wkt.loads for parsing WKT geometry strings, spatial operations like contains, intersects, distance, area, centroid)
- pyproj (coordinate projection for area/perimeter in metric units)
- h3 (H3 spatial indexing)
- math, json, re (standard library)

You have up to \$max_tool_calls steps to solve this. Write ONLY valid Python code. Do NOT include JSON blocks, markdown formatting, or any non-Python text in your code.

When you have computed the answer, call final_answer with a dict:
final_answer({"answer": 12.5, "unit": "km"})

If you cannot determine the answer, call:
final_answer({"answer": None})

IMPORTANT: final_answer is a built-in function. Never assign to it (do not write `final_answer = ...`). Always CALL it directly as shown above, e.g. `final_answer({"answer": ...})`. Use a different variable name for the dict if you want to build it first, e.g. `result = {...}; final_answer(result)`.

IMPORTANT: Call final_answer at the TOP LEVEL of your code block, NOT from inside a function definition. The agent only registers a final answer when final_answer is invoked at the top level. Do NOT do this:

```
def solve():  
    final_answer({"answer": ...})  
solve()
```

Do this instead - write the call at the top level directly:

```
result = ... # compute  
final_answer({"answer": result, "unit": "km"})
```

Figure 33: Tier 3 oracle prompt.