GEOEXPERTS: SCALING GLOBAL GEOCODING WITH REGION-AWARE LORA EXPERTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Geocoding means converting addresses or natural-language location descriptions into geographic coordinates and supports applications such as map search, route planning, and logistics. However, existing APIs are costly and impose substantial deployment overhead, while current models struggle to combine global coverage with robustness to complex inputs. To address these issues, we propose GeoExperts, an efficient global geocoding framework that requires no external geographic resources at inference time. To accommodate diverse address systems, GeoExperts adopts an expert specialization strategy. It divides the world into regions based on address systems and trains specialized experts on a shared lightweight language model. For address inputs, an address classification expert routes each request to the appropriate geocoding expert, which maps structured addresses directly to latitude and longitude. For natural-language inputs, a semantic understanding expert converts text into normalized addresses, enabling unified support for standard addresses, address variants, and natural-language location queries. We also introduce GeoMosaic, the first large-scale dataset designed for LLM-based geocoding without external resources, and GeoMosaic-Bench, which contains 40K samples and provides a unified evaluation of direct geocoding, input robustness, and semantic location reasoning. On structured address evaluation, GeoExperts reduces the base model's median error from 49.85 km to 1.44 km. Its accuracy approaches that of commercial APIs at the 5 km threshold and exceeds it at larger distance thresholds. The whole dataset and code will be public.

1 INTRODUCTION

Geocoding (Leidner, 2008) aims to convert postal addresses, structured place names, or natural-language location descriptions into geographic coordinates (Speriosu & Baldridge, 2013; Kamalloo & Rafiei, 2018; Wing & Baldridge, 2011). As a fundamental technology linking language to geographic space, with growing global location data and diverse user queries, geocoding systems need to combine global coverage, robustness to diverse inputs, and efficient deployment.

Traditional geocoding systems combine address parsing with retrieval from geographic databases (Speriosu & Baldridge, 2013; Kamalloo & Rafiei, 2018). Commercial APIs (Google, 2026; ope, 2026b; Amap, 2026) incur costs at scale, while local systems require maintaining large address indexes. Meanwhile, their performance depends on database coverage and input quality, making incomplete, multilingual, and implicit location descriptions challenging. To make geocoding more robust to complex inputs, prior studies have explored statistical address parsing, fuzzy matching, and neural address normalization (Cohen et al., 2003; Eisenstein et al., 2010; Yassine et al., 2020; Halterman, 2023). These methods typically improve individual components of traditional pipelines, while the final coordinates still depend on external candidate databases or geocoding services (Cohen et al., 2003; Yassine et al., 2020). In recent years, researchers begin to investigate whether LLMs can directly generate latitude and longitude from city names, points of interest (POIs), addresses, or complex location descriptions (Li et al., 2023; Manvi et al., 2024; Cui et al., 2026). Owing to the generalization capabilities of large language models, these approaches have shown promising performance in certain settings.

However, existing methods still fall well behind established APIs in geocoding accuracy. Moreover, current models are often fine-tuned on data from a single city, country, or a few regions (Li

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

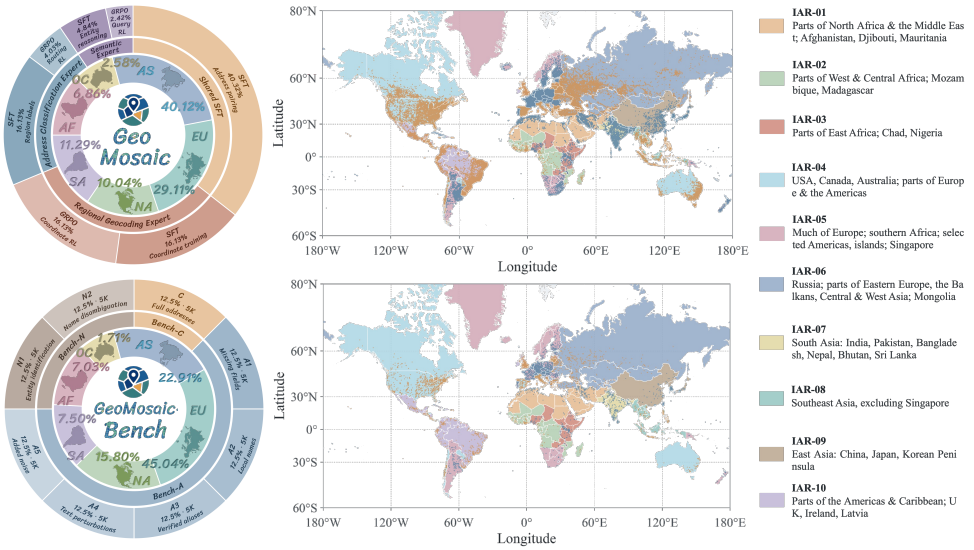


Figure 1: **Data composition and geographic distribution of GeoMosaic and GeoMosaic-Bench.** The left panel shows dataset splits and sample proportions by continent. The right panel shows sample locations, with map background colors indicating the ten IARs.

et al., 2023; Cui et al., 2026). This restricts their applicability and introduces a risk of catastrophic forgetting. Existing work also typically focuses on either converting structured addresses into coordinates (Cui et al., 2026) or extracting addresses from natural language for external systems to resolve. It does not provide a unified framework for standard addresses, non-standard addresses, and natural-language location descriptions. More importantly, current publicly available address datasets typically focus on normalized addresses, or geocoding within a single region (G-NAF, 2026; BAN, 2026). Large-scale datasets with coordinate supervision that jointly cover multiple countries, languages, and input formats are still lacking. Moreover, existing benchmarks rarely evaluate geocoding accuracy, the rate of valid coordinate generation, ambiguity resolution, and inference efficiency together (Truong et al., 2026; Rodrigues et al., 2026).

To address these challenges, we propose GeoExperts, which, to our knowledge, is the first efficient global geocoding framework that supports multiple input formats without relying on map APIs or external databases. Learning heterogeneous address conventions and a vast, long-tailed coordinate distribution with a single lightweight model can lead to negative transfer across regions. We therefore adapt an expert specialization strategy built on a shared lightweight base model. Regional geocoding experts map structured addresses within their assigned regions to coordinates. An address classification expert routes inputs to the appropriate regional expert, while a semantic understanding expert converts natural-language inputs into standard addresses. Together, these experts form a complete and efficient geocoding pipeline which can handle standard addresses, non-standard addresses, and natural-language location descriptions.

We also introduce GeoMosaic, the first dataset designed for LLM-based geocoding without relying on external resources, together with a unified evaluation benchmark, GeoMosaic-Bench. The dataset covers 185 countries and regions and contains approximately 1M samples. The benchmark comprises standard addresses, address variants, and natural-language queries. Evaluation metrics include median distance, accuracy at different distance thresholds, valid output rate, and latency. Experimental results show that our expert specialization strategy effectively balances global geocoding accuracy, input robustness, and computational efficiency. To sum up, our main contributions are summarized as follows:

- We propose GeoExperts, the first efficient global geocoding framework that jointly handles standard addresses, non-standard addresses, and natural-language location descriptions without relying on map APIs, gazetteers, or external databases.

Table 1: **Overview of the GeoMosaic dataset and benchmarks.**

Component	Size	Content		
		Purpose	**Size**	**Explanation**
Dataset	960K	SG SFT	500K	Complete structured addresses paired with coordinates.
		RG SFT	200K	A common entity set partitioned under $K = 10$ for training coordinate experts.
		RG GRPO	200K	Complete-address queries with reward-side frozen coordinates.
		AC SFT	200K	Geocoding inputs paired with regional labels.
		AC GRPO	50K	Routing queries with reward-side regional labels.
		SU SFT	60K	Audited reasoning traces for entity identification and disambiguation.
		SU GRPO	30K	Query-only rollouts optimized using GRPO.
Bench-C	5K	Complete structured addresses without masking, alias substitution, or injected noise.		
Bench-A	25K	A1	5K	Missing locality or street
		A2	5K	Local-language names
		A3	5K	Verified aliases
		A4	5K	Surface-form perturbations
		A5	5K	Redundant or irrelevant noise
Bench-N	10K	N1	5K	Fact-based entity identification
		N2	5K	Intent-guided name disambiguation

- We introduce GeoMosaic, a large-scale global dataset with coordinate supervision for standard addresses, address variants, and natural-language location queries, and GeoMosaic-Bench, a unified benchmark for evaluating geocoding across these input types and multiple performance metrics.
- We conduct systematic comparisons and ablation studies of GeoExperts. The results show that GeoExperts can serve as an alternative to commercial APIs within certain accuracy requirements while handling more complex input formats.

2 Related Work

2.1 Geocoding Model

Traditional commercial geocoding APIs typically retrieve candidate entities from gazetteers (Speriosu & Baldridge, 2013; Gritta et al., 2018) or classify inputs into predefined geographic regions (Wing & Baldridge, 2011; Roller et al., 2012; Wing & Baldridge, 2014). They rely on large external databases and struggle to handle complex natural-language inputs (Amap, 2026; Baidu, 2026; OpenCage, 2026). To address these limitations, researchers have begun exploring end-to-end geocoding systems based on large language models. However, existing methods typically focus on a single country or region (Li et al., 2023; Cui et al., 2026), while those supporting global inputs still depend on calls to external services (Zhang et al., 2024). More importantly, no existing LLM-based geocoding framework jointly supports standard addresses worldwide, robustness to address variations, and complex natural-language inputs.

2.2 Global Geocoding Datasets and Benchmarks

Geographic databases such as OpenStreetMap (Haklay & Weber, 2008), OpenAddresses (ope, 2026a), and GeoNames (geo, 2026) provide addresses, place names, and coordinates, but have uneven coverage and lack standardized splits and evaluation protocols for model development. Other resources focus on address parsing and normalization (parserator, 2026; Deepparse, 2026), or toponym disambiguation and text-based geolocation (Kamalloo & Rafiei, 2018; Lieberman et al., 2010; Gritta et al., 2018). More recently, ReaGeo (Cui et al., 2026) explores end-to-end address-to-coordinate generation mainly in Beijing and Singapore, while MESSY STREETS (Gaere & von Wangenheim, 2026) evaluates non-standard addresses from European and American web pages. However, neither provides global coverage of standard addresses, non-standard addresses, and natural-language location queries, leaving a need for unified datasets and benchmarks for retrieval-free geocoding across diverse address systems and input formats.

3 A new Geocoding Dataset and Benchmark: GeoMosaic

We construct GeoMosaic, a dataset for LLM-based geocoding without map APIs or external geographic retrieval at inference time, together with GeoMosaic-Bench, a unified evaluation benchmark. Specifically, we first collect approximately 6.13 million candidate records from Open-

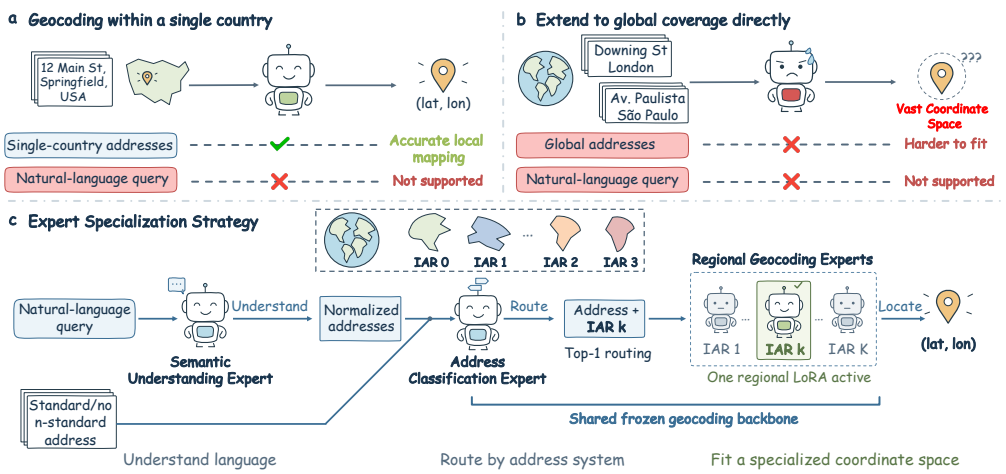


Figure 2: **Pipeline of GeoExperts.** Geocoding works well within individual countries, but simply adding global data makes the expanded coordinate space difficult to model. GeoExperts uses a semantic understanding expert to normalize natural-language queries, while standard and non-standard addresses go directly to the address classification expert. Addresses are then routed to the appropriate regional geocoding expert to generate latitude and longitude.

StreetMap (OpenStreetMap, 2026) that meet basic address and geometry quality requirements. We then further clean the data and adopt a water-filling sampling strategy. This strategy selects candidate records from different countries and regions in a round-robin manner to improve geographic balance. The resulting candidate set contains 2 million distinct geographic entities across 185 countries and regions, each with a complete structured address and corresponding coordinates.

3.1 GeoMosaic Dataset

As shown in Tab. 1, GeoMosaic provides training data for the shared backbone and three expert types. From the 2M candidate pool balanced through country-level water-filling sampling, we select 500K address-coordinate pairs in proportion to each country or region's share for backbone training. For regional geocoding and address classification, we define IARs based on writing systems, administrative hierarchies, and regional balance. IAR-level water-filling sampling yields a balanced 650K training pool with equal numbers of samples per partition. Samples from each IAR are split at the same ratio into two training sets, supervised by coordinates and IAR labels, respectively. For the semantic expert, we construct 90K query-structured address pairs using geographic facts from Wikidata (Wikidata, 2026). 30K are used directly for GRPO reinforcement learning, while the rest are augmented with reasoning traces from GPT5.6-Sol (OpenAI, 2025) for SFT.

3.2 GeoMosaic Bench

GeoMosaic-Bench consists of three complementary evaluation components, Bench-C, Bench-A, and Bench-N, which assess standard address geocoding, robustness to non-standard addresses, and natural-language location understanding, respectively. This design distinguishes basic geocoding accuracy from robustness to address variations and semantic understanding, helping identify the sources of model errors and the capabilities that require further improvement. Bench-C is sampled from the 2 million candidate records through a sampling process separate from training-set construction and contains structured addresses with complete fields and standard formatting. Bench-A evaluates missing fields, multilingual names, aliases, character perturbations, and redundant information. A1, A4, and A5 are paired transformations of Bench-C, while A2 and A3 use independently verified name variants. Both Bench-C and its derived Bench-A variants undergo explicit non-overlap checks against the training set to ensure train-test separation. Bench-N follows the same construction process as the training data for the semantic understanding expert. It selects geographic entities from the candidate pool and uses related geographic facts from Wikidata to construct natural-language location queries. This component includes N1 and N2, which evaluate the model's ability to infer

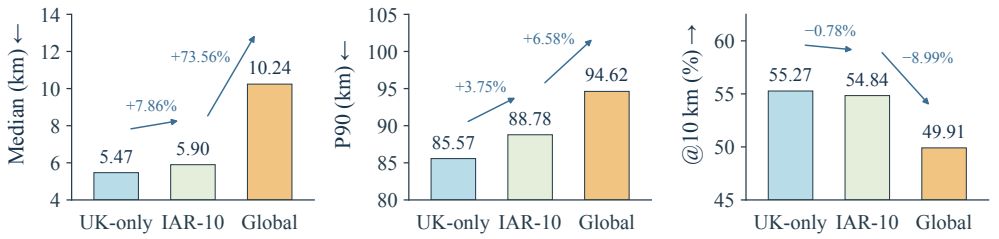


Figure 3: **The need for expert specialization. UK-Only**: Trained only on UK data. **IAR-10**: Regional experts trained separately for each of the ten IARs. **Global**: Trained jointly on global data.

locations from factual clues and user needs. Detailed definitions and input formats for each subtask are provided in Tab. 1.

4 METHODOLOGY

GeoExperts consists of three types of experts with complementary roles: Regional Geocoding Experts, an Address Classification Expert, and a Semantic Understanding Expert. For standard and non-standard addresses, the address classification expert first identifies the appropriate region based on its address system. The corresponding regional geocoding expert is then activated to map the original address directly to latitude and longitude. For natural-language inputs, the semantic understanding expert first interprets location clues and generates a normalized, structured address. This address then follows the same classification and regional geocoding process.

4.1 EXPERT SPECIALIZATION STRATEGY

Prior studies (Cui et al., 2026) have shown that regional fine-tuning can give lightweight language models strong geocoding capabilities. However, with limited model capacity, jointly learning heterogeneous address conventions worldwide and large-scale address-coordinate mappings may reduce regional geocoding accuracy. As shown in Fig. 3, the model trained only on UK data (UK-only) performs well within the UK. However, its accuracy on UK addresses drops substantially when training is expanded to global data that include UK samples. By contrast, IAR-10 uses the same amount of training data as the jointly trained global baseline but adapts each regional geocoding expert only on data from its corresponding IAR. UK addresses are handled by the expert for the IAR containing the UK. These results show that expert specialization mitigates the regional performance degradation caused by joint global training, preserving high regional accuracy while extending global coverage.

We therefore group countries and regions into K Institutional Address Regimes (IARs) based on shared characteristics, including address formats, administrative structures, and writing systems. Within the limited sample space of each IAR, we use LoRA (Hu et al., 2021) to train a regional geocoding expert specialized in that regime. We also train an address classification expert to identify the IAR of each input address and route the request to the corresponding regional expert for geocoding. Only one regional LoRA adapter is activated for each input, so the number of active parameters does not grow linearly with the number of experts. Furthermore, when new address data become available for a region, only the corresponding LoRA adapter needs to be updated, without retraining the entire global model. We then introduce a semantic understanding expert to convert natural-language queries into structured addresses that the downstream address classification and regional geocoding modules can process.

4.2 REGIONAL GEOCODING EXPERTS

In GeoExperts, each regional geocoding expert learns the conditional mapping from address representations within a specific IAR to latitude and longitude in decimal degrees. Let $\mathcal{D}_k = \{(a_i, y_i)\}$ denote the training data for the k-th IAR, where a_i is an input address and $y_i = (\phi_i, \lambda_i)$ represents the reference coordinates. The regional expert models

$$p_k(y \mid a) = p\big(y \mid a; \theta_g, \Delta\theta_k\big), \tag{1}$$

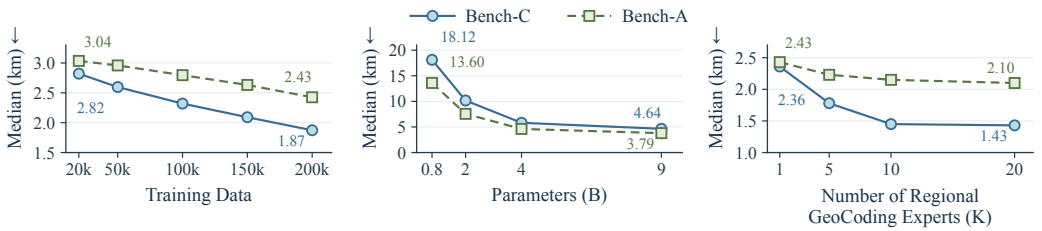


Figure 4: **Scaling analysis.** From left to right, the panels show how the number of training samples, model parameter count, and number of regional experts (K) affect median location error on Bench-C and Bench-A. Lower values are better.

where θ_{g} denotes the frozen parameters of the global geocoding backbone, and $\Delta\theta_k$ refers to the trainable parameters of the k-th regional LoRA adapter. Regional LoRA experts share a global backbone to retain transferable knowledge while specializing in narrower address distributions, reducing training and storage costs relative to separate full models and enabling efficient regional updates.

4.3 Address Classification Expert

Since the shared backbone cannot determine which regional adapter to load on its own, we train a separate address classification expert to map each complete address to the corresponding IAR ID. This expert shares the same frozen global geocoding backbone with the regional geocoding experts but uses its own classification LoRA parameters $\Delta\theta_{\mathrm{cls}}$. Given an address $\tilde{a}$, the classification expert estimates

$$p_{\mathrm{cls}}(z \mid \tilde{a}) = p\big(z \mid \tilde{a}; \theta_{\mathrm{g}}, \Delta\theta_{\mathrm{cls}}\big), \qquad z \in \{1, \ldots, K\}. \tag{2}$$

It then selects the IAR through a top-1 decision:

$$\hat{z} = \arg\max_{z} p_{\mathrm{cls}}(z \mid \tilde{a}). \tag{3}$$

During inference, the system first loads the classification LoRA adapter onto the shared backbone to predict the IAR. It then unloads this adapter and dynamically loads $\Delta\theta_{\hat{z}}$ for a second inference pass that maps the address to coordinates. Since only low-rank adapters are switched rather than full model weights, this two-stage process maintains low storage and model-switching overhead.

4.4 Semantic Understanding Expert

To extend GeoExperts from address inputs to natural-language location descriptions, we introduce a semantic understanding expert. Natural-language inputs may contain redundant text, indirect references, or location clues that require contextual interpretation. The address classification expert therefore cannot reliably assign such inputs directly to an IAR. To avoid interference between open-ended semantic reasoning and fine-grained coordinate fitting under a limited parameter budget, we use a language model separate from the geocoding backbone and adapt it with a semantic understanding LoRA adapter. Given a natural-language query q_{NL}, this expert generates

$$\mathcal{E}_{\mathrm{sem}}(q_{\mathrm{NL}}) = (r, a), \tag{4}$$

where r is the reasoning process used to resolve location clues, and a is the structured address passed to subsequent modules. This address is then passed to the address classification expert and the corresponding regional geocoding expert in sequence,

$$\hat{z} = \arg\max_{z} p_{\mathrm{cls}}(z \mid a), \qquad (\hat{\phi}, \hat{\lambda}) = \mathcal{E}_{\mathrm{geo}}^{(\hat{z})}(a). \tag{5}$$

This cascade separates natural-language interpretation and address-to-coordinate mapping into two clearly defined sub-tasks and forms a unified geocoding framework for standard addresses, non-standard addresses, and natural-language inputs.

4.5 Training of Experts

Supervised expert adaptation. We first perform full-parameter supervised fine-tuning of the base model on address–coordinate pairs from around the world to establish basic global geocoding capabilities. The resulting model serves as the shared backbone for the regional geocoding experts and

Table 2: **Overall geocoding results.** Best and second-best results are in **bold** and underlined. Each invalid output is assigned the largest error observed among the valid outputs.

Model	Valid (%)	Bench-C					Bench-A					Bench-N				
		Med.↓	P90↓	@1↑	@5↑	@25↑	Med.↓	P90↓	@1↑	@5↑	@25↑	Med.↓	P90↓	@1↑	@5↑	@25↑
Commercial Geocoding APIs																
OpenCage	68.16	0.00	19266.35	74.84	75.42	75.76	0.04	19799.17	54.03	54.77	55.19	–	–	–	–	–
Geoapify	90.17	0.01	583.83	67.00	72.80	76.40	0.21	1428.50	52.56	59.52	65.20	–	–	–	–	–
TomTom	59.47	34.60	15176.97	33.40	43.00	48.80	30.09	16763.71	37.80	45.64	49.60	–	–	–	–	–
Closed-source models or large open-source models																
GPT-4.1	97.69	2.09	19.71	27.80	71.24	89.30	1.43	19.51	39.00	74.46	87.92	453.45	8868.77	7.17	17.08	23.74
DeepSeek-V4-Flash	99.99	2.29	25.59	27.20	69.66	89.88	1.58	24.31	38.78	75.09	90.11	519.63	8783.10	5.10	14.12	20.58
Doubao-Seed-2.0-Lite	99.90	2.76	33.22	20.98	65.34	88.12	1.90	33.03	32.41	71.98	88.74	527.12	8745.63	5.05	14.23	20.03
Qwen3-4B	**100.00**	132.92	740.44	0.48	4.34	20.16	76.32	616.20	2.35	12.50	32.72	1144.65	**9751.59**	0.71	3.89	8.51
Qwen3.5-4B	**100.00**	49.85	386.45	2.96	17.48	40.56	21.56	322.46	9.59	29.88	51.91	2044.35	10586.54	1.51	6.30	11.39
InternLM3-8B	99.99	83.06	598.14	0.62	5.38	25.40	46.06	446.68	4.03	16.36	39.34	1131.09	10078.63	0.88	5.09	10.21
Gemma3-4B	99.99	132.30	773.88	0.40	3.82	18.36	76.82	615.05	2.25	11.47	28.87	1245.92	10128.02	0.49	3.46	7.95
Gemma3-12B	99.99	43.43	341.46	1.38	11.26	36.94	27.58	326.66	4.78	20.54	47.98	885.60	10158.17	1.36	6.20	12.88
Llama3.2-3B	**100.00**	70.91	554.04	0.70	7.54	30.26	40.09	470.40	4.21	18.38	42.39	1137.79	10289.88	0.70	3.50	8.20
Llama3.2-8B	**100.00**	11.95	200.53	3.26	27.54	65.16	7.86	162.60	10.07	39.61	70.21	963.50	9751.89	1.77	7.65	13.99
MiniCPM5-2B	**100.00**	339.31	3289.34	0.32	1.14	6.04	223.53	3350.05	1.72	5.44	14.45	4128.62	10483.16	0.26	1.41	3.43
GeoExperts-SFT (Ours)	99.54	1.69	104.09	39.58	68.72	82.36	2.50	150.63	31.68	62.55	77.86	841.82	10575.36	**2.27**	**9.24**	14.61
GeoExperts (Ours)	**100.00**	**1.44**	**92.35**	**43.16**	**70.50**	**83.40**	**2.19**	**140.00**	**34.54**	**64.50**	**78.92**	**806.25**	10236.20	**2.27**	9.22	**14.79**

the address classification expert. We then freeze this backbone and train separate LoRA adapters for these experts. Each regional geocoding expert uses address–coordinate pairs from its corresponding IAR to learn fine-grained geocoding mappings within that partition. The address classification expert takes addresses as input and uses the corresponding IAR IDs as supervision labels, learning to route requests to the appropriate regional expert. Natural-language semantic reasoning and precise coordinate fitting have different objectives. To reduce interference from geocoding adaptation with semantic understanding, the semantic understanding expert uses the original base model as a separate backbone and undergoes supervised fine-tuning with LoRA. Given a natural-language query, it generates a reasoning process that resolves location clues, along with the structured address.

Reinforcement Learning with GRPO. To align the reinforcement learning objective with final geocoding accuracy, we design a multiscale distance reward based on the Haversine distance between predicted and ground-truth coordinates. Let d denote the location error in kilometers. We define the reward as

$$R_{\text{dist}}(d) = \sum_{j=1}^{3} w_j \exp\left(-\frac{d}{\tau_j}\right) + w_{\text{g}} \,\text{clip}\left(1 - \frac{\log(1+d)}{\log(1+D)},\, 0,\, 1\right), \tag{6}$$

where τ_j controls the distance scale of each exponential term, and D is a global distance normalization constant. The operator $\text{clip}(\cdot, 0, 1)$ clips values to $[0, 1]$. All weights are nonnegative and satisfy $\sum_{j=1}^{3} w_j + w_{\text{g}} = 1$. The exponential terms provide feedback on location errors at different spatial scales, while the logarithmic term preserves distinguishable reward signals even for large errors. We then use GRPO (Guo et al., 2025) to further optimize each expert's LoRA adapter, starting from its SFT checkpoint. For each input, the model generates multiple candidate outputs and updates its parameters based on relative rewards within the group, while keeping the backbone frozen. The regional geocoding experts optimize coordinate predictions on address samples from their respective IARs using both distance and format rewards. The address classification expert combines an exact-match reward for the IAR ID with a format reward to improve partition selection accuracy. Once the regional and classification experts are trained, we apply GRPO to the semantic understanding expert, using the frozen downstream geocoding pipeline to compute rewards. We compute R_{dist} shown in Eq. (6), supplemented by format and output-type rewards. During this stage, the classification expert, regional experts, and all backbone parameters remain frozen, and only the semantic expert's LoRA adapter is updated.

Table 3: **Component ablation.** SU, SG, AC, and RG denote Semantic Understanding Expert, Shared Geocoding Backbone, Address Classification Expert, and Regional Geocoding Experts, respectively. "/" indicates that the expert is not required for the task.

Model	Components				Bench-C			Bench-A			Bench-N		
	SU	SG	AC	RG	Med. ↓	P90 ↓	@5 ↑	Med. ↓	P90 ↓	@5 ↑	Med. ↓	P90 ↓	@5 ↑
Qwen3.5-4B													
Using Qwen3.5-4B for all experts					48.39	386.45	17.56	21.78	325.87	29.94	1029.73	12682.22	8.65
Using SG for all experts	/	✓			3.42	224.56	57.12	3.92	245.14	54.36	–	–	–
Using Qwen3.5-4B for RG	/	✓	✓		51.11	390.77	16.70	23.90	335.15	28.60	–	–	–
Using Qwen3.5-4B for AC	/	✓		✓	1.46	95.44	70.28	2.21	138.94	64.48	–	–	–
GeoExperts + Qwen3.5-4B for SU	/	✓	✓	✓	**1.44**	**92.35**	**70.50**	**2.19**	**140.00**	**64.50**	–	–	–
Using Qwen3.5-4B for other experts	✓				–	–	–	–	–	–	820.32	10283.38	9.13
Using SG for other experts	✓	✓			–	–	–	–	–	–	818.13	10277.62	8.78
Using Qwen3.5-4B for RG	✓	✓	✓		–	–	–	–	–	–	816.83	10253.32	8.95
Using Qwen3.5-4B for AC	✓	✓		✓	–	–	–	–	–	–	816.54	**10217.11**	9.15
GeoExperts (Ours)	✓	✓	✓	✓	–	–	–	–	–	–	**806.25**	10236.20	**9.22**

Table 4: **Ablation study on AC and RG.** Best and second-best results are in **bold** and underlined

Training stage					Bench-C					Bench-A				
SG	AC		RG		Med.↓	P90↓	@1↑	@5↑	@25↑	Med.↓	P90↓	@1↑	@5↑	@25↑
	SFT	GRPO	SFT	GRPO										
					49.85	386.45	2.96	17.48	40.56	21.56	322.46	9.59	29.88	51.91
✓					3.38	225.04	26.64	57.16	74.20	3.92	246.91	22.98	54.34	71.89
✓	✓		✓		1.69	104.09	39.58	68.72	82.36	2.50	150.63	31.68	62.55	77.86
✓	✓	✓	✓		1.52	94.37	42.66	70.18	**83.52**	2.24	140.99	34.18	64.26	78.83
✓	✓		✓	✓	1.45	**92.29**	43.10	**70.70**	83.44	**2.17**	**138.86**	**34.69**	**64.73**	**79.02**
✓	✓	✓	✓	✓	**1.44**	92.35	**43.16**	70.50	83.40	2.19	140.00	34.54	64.50	78.92

5 EXPERIMENTS

All components of GeoExperts are fine-tuned from Qwen3.5-4B on 8× NVIDIA A40 GPUs. The evaluation is mainly on 2× NVIDIA GeForce RTX 3090 GPUs.

5.1 MAIN RESULTS

Performance on GeoMosaic-Bench. As shown in Tab. 2, GeoExperts substantially improves geocoding for both standard and non-standard addresses. Relative to Qwen3.5-4B, it reduces median error on Bench-C (49.85 → 1.44 km) and Bench-A (21.56 → 2.19 km), while maintaining a 100% valid output rate on both tasks. On the natural-language task Bench-N, GeoExperts reduces the base model's median error from 2044.35 km to 806.25 km. Although commercial APIs retain an advantage in median error, GeoExperts achieves accuracies of 83.40% and 78.92% at the 25 km threshold on Bench-C and Bench-A, respectively, exceeding all compared commercial APIs. Together, these results show that GeoExperts supports diverse input formats within a unified framework and offers a viable alternative to commercial APIs in applications with greater tolerance for location errors.

Scaling Analysis. As shown in Fig. 4, GeoExperts improves both standard address geocoding and robustness to non-standard inputs as training data and model capacity increase. Increasing training samples from 20K to 200K reduces median error on Bench-C (2.82 → 1.87 km) and Bench-A (3.04 → 2.43 km). Scaling from 0.8B to 9B parameters further demonstrates consistent benefits across the two tasks, with median error reductions of 74.4% and 72.1%, respectively. Expert specialization also improves accuracy, reducing median errors by 38.6% and 11.5% as K increases from 1 to 10, while keeping only one regional LoRA adapter active per request. Increasing K to 20 yields only another 0.02 km and 0.05 km improvement, respectively, so we use $K = 10$.

Replacing Commercial APIs in Real-World Tasks. We evaluate GeoExperts on CRM address geocoding and OCR logistics document geocoding using clean and noisy inputs. CRM inputs mix addresses with contact details, department names, and delivery instructions, while OCR inputs contain field labels, line breaks, and recognition errors. As shown in Tab. 6, GeoExperts maintains a

Table 5: **Ablation study on SU.**

Training stage				N1		N2	
SFT			GRPO	Med.↓	@25↑	Med.↓	@25↑
Base	SG	AC					
Qwen3.5-4B				8662.86	1.62	7996.78	1.64
✓				710.41	21.50	943.31	7.72
	✓			1158.58	8.00	2103.17	2.80
		✓		1836.28	1.20	5704.69	0.20
✓			✓	**668.77**	**21.74**	**906.80**	**7.84**

Table 6: **Applications in tasks requiring geocoding.**

Scenario	Input	System	Valid	Med. km ↓	@5 km ↑	@25 km ↑
CRM/CSV	Clean	OpenCage	85.2%	**0.000**	**79.0**	**79.2**
	Clean	Ours	**100.0%**	2.428	68.0	78.0
	Noisy	OpenCage	18.0%	6325.461	2.0	2.0
	Noisy	Ours	**100.0%**	**3.127**	**60.0**	**72.0**
OCR document	Clean	OpenCage	38.0%	6059.804	5.8	9.4
	Clean	Ours	**100.0%**	**2.561**	**64.0**	**76.0**
	Noisy	OpenCage	37.8%	6115.880	5.8	9.4
	Noisy	Ours	**100.0%**	**2.346**	**64.0**	**74.0**

100% valid output rate across all settings and achieves median errors of 3.13 km and 2.35 km on noisy CRM and OCR inputs, respectively. Compared with direct OpenCage calls, it substantially improves accuracy at the 5 km threshold on noisy CRM ($2.0\% \rightarrow 60.0\%$) and OCR ($5.8\% \rightarrow 64.0\%$) inputs. These results show that GeoExperts can handle noisy business text and serve as an alternative to direct commercial API calls in these workflows when kilometer-level accuracy is sufficient.

5.2 Ablation Analysis

Effectiveness of Components. As shown in Tab. 3, the shared backbone and specialized experts complement one another to improve geocoding performance. Compared with using the shared backbone alone, adding regional and classification experts substantially reduces median error on Bench-C ($3.42 \rightarrow 1.44$ km) and Bench-A ($3.92 \rightarrow 2.19$ km), demonstrating the benefits of regional specialization for fine-grained geocoding. The dedicated classification expert further improves median error and accuracy at the 5 km threshold on address tasks, helping the full inference pipeline make better use of the regional experts. On Bench-N, the semantic expert provides the main improvement in median error ($1029.73 \rightarrow 820.32$ km), which decreases further to 806.25 km when combined with the downstream classification and regional experts. The full model achieves the lowest median error and highest accuracy at the 5 km threshold in this ablation on Bench-N. These results support assigning semantic parsing, regional routing, and coordinate prediction to specialized experts to better handle diverse forms of location input.

Ablation on Training Strategies. As shown in Tab. 4, shared backbone training and expert SFT progressively improve geocoding performance, while GRPO provides further gains in accuracy. Applying GRPO to the regional experts yields larger reductions in median error than applying it to the classification expert, but optimizing both does not consistently provide additional gains. GRPO provides only modest gains for the classification expert, largely because its accuracy already exceeds 95% after SFT. For the semantic expert, Tab. 5 shows that GRPO reduces median error and improves accuracy at the 25 km threshold on both N1 and N2. These results show that GRPO complements supervised training, but its benefits depend on the expert being optimized and the task.

Discussion of Decoupling. As shown in Tab. 5, we initialize the semantic understanding expert from the original base model, the shared geocoding backbone (SG), or the address classification expert (AC). After semantic SFT, initialization from the original base model performs best on both N1 and N2, reducing N2 median error relative to initialization from SG ($2103.17 \rightarrow 943.31$ km) or AC ($5704.69 \rightarrow 943.31$ km). This suggests that adaptation for coordinate prediction and regional classification may interfere with subsequent semantic learning. We therefore train the semantic expert independently from the original base model and connect it to downstream experts through structured addresses, decoupling semantic reasoning from geocoding adaptation.

6 Conclusions

We introduce GeoExperts, which combines regional expert specialization on a shared backbone with an independent semantic understanding module. It jointly handles standard addresses, non-standard addresses, and natural-language location descriptions without external geographic resources at inference time. We also construct GeoMosaic and GeoMosaic-Bench to provide training resources and unified evaluation for global geocoding. Experiments demonstrate the value of expert specialization for geocoding accuracy and complex input handling. They also show that GeoExperts has the potential to replace commercial APIs in noisy business scenarios where kilometer-level location errors are acceptable, offering a viable approach to locally deployable global geocoding.

AI USE STATEMENT

Generative AI tools were used to support literature searches, improve the wording of the manuscript, and make the text more concise. The authors checked and revised all AI-assisted text to ensure its accuracy and consistency with the methods and experimental findings. All cited references were verified against their original sources. The authors assume full responsibility for the final content of this work, including all text, claims, and artifacts developed with the assistance of generative AI.

ETHICS STATEMENT

This study uses address and geometry data from OpenStreetMap and geographic entity facts from Wikidata to investigate the mapping from addresses and location descriptions to geographic coordinates. We plan to release the relevant data with clear documentation of their sources, processing procedures, and applicable licenses. Before release, we will review records that may contain personal information or sensitive locations. Model outputs should be treated as location estimates, not as factual evidence of an individual's identity, activities, or whereabouts. They should not be used to locate or track individuals without authorization. The IAR partitioning in this work is used solely for address modeling and expert routing and does not imply any political stance or value judgment regarding any country, region, or group.

REPRODUCIBILITY STATEMENT

This paper describes the construction of GeoMosaic, the expert partitioning scheme, SFT and GRPO training, and the distance reward design. We plan to release code for data construction, training, inference, and evaluation, along with data and model resources where permitted by the applicable licenses. We will also provide model versions, sample counts for each stage, prompt templates, training and reward configurations, random seeds, checkpoint selection criteria, and the runtime environment. The evaluation materials will specify adapter loading and routing settings, decoding parameters, coordinate parsing rules, and how invalid predictions are scored.

REFERENCES

Geonames. https://www.geonames.org/, 2026. Accessed: 2026-9-20.

Openaddresses. https://openaddresses.io/, 2026a. Accessed: 2026-9-20.

Opencage. https://opencagedata.com/, 2026b. Accessed: 2026-9-20.

Amap. Amap Web Service API: Geocoding and Reverse Geocoding, 2026. URL https://lbs.amap.com/api/webservice/guide/api/georegeo. In Chinese. Accessed: 2026-09-21.

Baidu. Baidu Maps Open Platform: Geocoding API, 2026. URL https://lbsyun.baidu.com/docs/webapi?title=geocoding/guide/webservice-geocoding. In Chinese. Accessed: 2026-09-21.

BAN. La Base Adresse Nationale, 2026. URL https://adresse.data.gouv.fr/. Accessed: 2026-09-22.

William W Cohen, Pradeep Ravikumar, Stephen E Fienberg, et al. A comparison of string distance metrics for name-matching tasks. In *IIWeb*, volume 3, pp. 73–78, 2003.

Jian Cui, Zhiyuan Ren, Desheng Weng, Yongqi Zhao, Gong Wenbin, Yu Lei, and Zhenning Dong. Reageo: Reasoning-enhanced end-to-end geocoding with llms. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 28012–28023, 2026.

Deepparse. Deepparse: a library for parsing multinational street addresses, 2026. URL https://deepparse.org/. Accessed: 2026-09-22.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pp. 1277–1287, 2010.

G-NAF. Geoscape Geocoded National Address File, 2026. URL https://data.gov.au/data/dataset/geocoded-national-address-file-g-naf. Accessed: 2026-09-22.

Edward Gaere and Florian von Wangenheim. MESSY STREETS: A benchmark for geocoding real-world addresses. In *Proceedings of the 34th ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '26)*, Riverside, CA, USA, 2026. ACM. doi: 10.1145/3841645.3843362.

Geofabrik. Geofabrik: free geodata database, 2026. URL https://www.geofabrik.de/. Accessed: 2026-09-24.

Google. Google Maps Platform: Geocoding API Overview, 2026. URL https://developers.google.com/maps/documentation/geocoding/overview. Accessed: 2026-09-21.

Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. What's missing in geographical parsing? *Language Resources and Evaluation*, 52(2):603–623, 2018.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Mordechai Haklay and Patrick Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive computing*, 7(4):12–18, 2008.

Andrew Halterman. Mordecai 3: A neural geoparser and event geocoder. *arXiv preprint arXiv:2303.13675*, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Ehsan Kamalloo and Davood Rafiei. A coherent unsupervised model for toponym resolution. In *Proceedings of the 2018 world wide web conference*, pp. 1287–1296, 2018.

Jochen L Leidner. *Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names*. Universal-Publishers, 2008.

Zekun Li, Wenxuan Zhou, Yao-Yi Chiang, and Muhao Chen. Geolm: Empowering language models for geospatially grounded language understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 5227–5240, 2023.

Michael D Lieberman, Hanan Samet, and Jagan Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *2010 IEEE 26th international conference on data engineering (ICDE 2010)*, pp. 201–212. IEEE, 2010.

Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. Geollm: Extracting geospatial knowledge from large language models. In *International Conference on Learning Representations*, volume 2024, pp. 38791–38807, 2024.

OpenAI. Introducing GPT-5, 2025. Accessed: 2025-11-07.

OpenCage. OpenCage Geocoding API Documentation, 2026. URL https://opencagedata.com/api. Undated online documentation; year denotes access year. Accessed: 2026-09-21.

OpenStreetMap. OpenStreetMap: an open-source geographic database, 2026. URL https://wiki.openstreetmap.org/wiki/Downloading_data. Accessed: 2026-09-22.

parserator. U.S. address parser, 2026. URL https://parserator.us/usaddress/. Accessed: 2026-09-22.

Rodrigo Ferreira Rodrigues, Karim Radouane, Jose G Moreno, and Lynda Tamine. Geobench-llm: A comprehensive benchmark for evaluating llms on geo-related tasks. *arXiv preprint arXiv:2608.07411*, 2026.

Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pp. 1500–1510, 2012.

Michael Speriosu and Jason Baldridge. Text-driven toponym resolution using indirect supervision. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1466–1476, 2013.

Thinh Hung Truong, Jey Han Lau, and Jianzhong Qi. Gpsbench: Do large language models understand gps coordinates? *arXiv preprint arXiv:2602.16105*, 2026.

Wikidata. Wikidata: an knowledge base, 2026. URL https://www.wikidata.org/wiki/Wikidata:Main_Page. Accessed: 2026-09-24.

Benjamin Wing and Jason Baldridge. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 955–964, 2011.

Benjamin Wing and Jason Baldridge. Hierarchical discriminative classification for text-based geolocation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 336–348, 2014.

Marouane Yassine, David Beauchemin, François Laviolette, and Luc Lamontagne. Leveraging Subword Embeddings for Multinational Address Parsing, 2020.

Yifan Zhang, Cheng Wei, Zhengting He, and Wenhao Yu. Geogpt: An assistant for understanding and processing geospatial tasks. *International Journal of Applied Earth Observation and Geoinformation*, 131:103976, 2024.

SUPPLEMENTARY MATERIALS OF GEOEXPERTS: SCALING GLOBAL GEOCODING WITH REGION-AWARE LORA EXPERTS

In order to provide a more complete illustration of the capabilities of GeoExperts, a comprehensive appendix has been developed.

CONTENTS

A More Experiments **14**

A.1 Task-level Comparison 14

A.2 Training with other model families 14

A.3 Comparison of geocoding across IARs 14

A.4 Analysis of inference efficiency 14

B Detailed Information About GeoMosaic **15**

B.1 Data Source 15

B.2 Dataset Construction 15

B.3 Benchmark Construction 16

B.4 More Cases 17

C More Information About Training and Evaluation **18**

C.1 More Details of format reward 18

C.2 Hyperparameter settings 19

C.3 Training and evaluation prompts 19

D Case Study **19**

E Discussion **20**

E.1 Limitations 20

E.2 Broader Impacts 20

A More Experiments

A.1 Task-level Comparison

As shown in Tab. 7, GeoExperts also outperforms the original base model when handling missing fields, multilingual expressions, and alias substitution, although performance varies across tasks. The median errors on A2 and A3 are 1.20 km and 2.71 km, respectively, compared with 10.75 km on A1. This indicates that missing key address fields remains a major challenge.

A.2 Training with other model families

To assess the applicability of our training method across model families, we train nine base models from five families for geocoding. All experiments use $K = 10$, 100K training samples, and ground-truth IAR routing. As shown in Tab. 8, all models improve consistently on Bench-C and Bench-A. Median and 90th-percentile (P90) errors decrease, while accuracy improves at all distance thresholds, showing that the method works across different architectures and model sizes. For example, the median errors of Gemma3-4B-IT and Llama3.2-3B-Instruct on Bench-C decrease from 132.30 km and 70.91 km to 15.43 km and 10.92 km, respectively, representing reductions of 88.3% and 84.6%. After training, Qwen3.5-9B performs best in this comparison, with median errors of 5.88 km on Bench-C and 4.40 km on Bench-A. This indicates that base model capabilities still affect final geocoding accuracy. These results validate the effectiveness of geocoding adaptation across model families, but primarily reflect regional expert capabilities because ground-truth IAR routing is used.

A.3 Comparison of geocoding across IARs

To analyze regional expert specialization, we construct a diagnostic set of 1,000 samples, with 100 samples per IAR, and evaluate each regional expert across all partitions. Tab. 9 reports the mean $\log(1+d_{\mathrm{Hav}})$, with rows indicating the deployed expert and columns indicating the samples' ground-truth IAR. Lower values are better. The corresponding regional expert achieves the lowest error in nine of the ten partitions. Matching expert–IAR pairs have a mean score of 1.7471, compared with 2.1163 for mismatched pairs, indicating that regional adaptation develops geocoding capabilities suited to the target partition. This advantage is particularly clear for IAR-02 and IAR-03, where the corresponding experts score 2.2461 and 1.5914, compared with the best scores of 3.5707 and 2.7123 among other experts, respectively. Overall, these results support IAR-based regional geocoding and show that accurate routing helps realize the benefits of regional expert adaptation.

A.4 Analysis of inference efficiency

Tab. 10 presents the end-to-end inference throughput of OpenCage and GeoExperts, together with additional efficiency metrics for GeoExperts. Both systems process 30,000 queries. OpenCage achieves an observed throughput of 11.67 samples/s, whereas GeoExperts achieves 8.81 samples/s, corresponding to approximately 75.5% of the observed OpenCage throughput. GeoExperts runs in BF16 precision on four NVIDIA A40 GPUs and sequentially invokes the Address Classification Expert and a Regional Geocoding Expert, resulting in two serial model requests per sample. Under this multi-stage inference pipeline, the system processes 17.62 model requests/s and achieves prompt-prefill and output-generation throughputs of 10,064.57 tok./s and 281.18 tok./s, respectively. The substantially lower generation throughput indicates that autoregressive decoding remains the primary performance bottleneck. Loading the model weights requires 5.36 GiB per GPU, excluding the memory occupied by the KV cache and CUDA graphs. The four A40 GPUs provide a theoretical peak BF16 throughput of 598.80 TFLOP/s. Overall, despite performing two-stage inference for routing and regional coordinate prediction, GeoExperts achieves an aggregate throughput close to that of the external geocoding API, demonstrating the practical batch-serving efficiency of the proposed expert-based architecture.

Table 7: **Task-level geocoding results.** Median Haversine distance error is reported.

Model	Bench-C	Bench-A					Bench-N	
	C1	A1	A2	A3	A4	A5	N1	N2
Commercial Geocoding APIs								
OpenCage	0.00	0.00	0.01	0.01	5589.69	5772.44	–	–
Geoapify	0.01	0.01	0.02	0.02	64.23	69.65	–	–
TomTom	0.69	7.31	0.25	0.27	38.39	–	–	–
Closed-source models or large open-source models								
GPT-4.1	2.09	3.85	0.48	0.83	2.13	2.10	436.84	466.59
DeepSeek-V4-Flash	2.29	4.83	0.50	0.90	2.31	2.30	453.60	575.02
Doubao-Seed-2.0-Lite	2.76	5.33	0.76	1.14	2.78	2.67	448.46	590.30
Open-source general models								
Qwen3.5-4B	49.85	73.63	3.13	7.49	49.87	47.66	1579.79	2627.44
InternLM3-8B	83.06	110.35	8.29	18.15	83.83	80.32	1009.84	1239.78
Gemma3-12B	43.43	65.76	6.66	13.85	43.23	42.44	807.97	951.17
Llama3.2-8B	11.95	23.40	2.38	4.50	12.09	13.00	976.90	947.62
GeoExperts-SFT(Ours)	1.69	12.45	1.38	3.01	1.67	2.11	710.41	943.31
GeoExperts(Ours)	**1.44**	**10.75**	**1.20**	**2.71**	**1.44**	**1.85**	**668.77**	**906.80**

B Detailed Information About GeoMosaic

B.1 Data Source

We obtain global OpenStreetMap data from Geofabrik and store them in Protocolbuffer Binary Format (PBF) (Geofabrik, 2026). We then use Osmium to locally filter nodes, ways, and relations with name tags, covering administrative areas, settlements, natural geographic features, landmarks, points of interest, and transportation facilities. For each candidate object, we use its OSM object type and numeric identifier to locate it in the corresponding source PBF file, then re-extract and verify its geometry. For point entities, we use the original node coordinates directly. For polygon entities, we deterministically compute a representative point inside the polygon and use it as the coordinate annotation. Source region labels and candidate bounding-box centers are retained only as auxiliary metadata and are not used to determine coordinate annotations.

B.2 Dataset Construction

Water-Filling Sampling To mitigate the substantial geographic imbalance in raw OpenStreetMap coverage, we adopt deterministic, capacity-constrained water-filling sampling. The sampler first covers every country with eligible candidates and then progressively raises the sampling level of all countries that retain sufficient supply. Once a country exhausts its eligible candidates, its allocation is fixed, while the remaining quota continues to be distributed among countries with available supply. Within each country, candidates are selected in rotation across localities, target categories, and geometry types, with a fixed random seed and versioned ranking keys ensuring reproducibility. This strategy reduces the dominance of data-rich regions and improves global geographic coverage without duplicating samples, synthesizing entities, or relaxing quality and leakage constraints.

Deduplication Because distinct OSM objects may share the same coordinates or name, while a single object may have multilingual names, aliases, and multiple task templates, we do not use coordinates or POI names as primary deduplication keys. Instead, each target is identified by the combination of its OSM object type and numeric identifier, reverified directly from the original OpenStreetMap PBF. A node represents a point object, a way represents a line or closed area such as a road, building, or park, and a relation represents a composite object such as an administrative boundary or multipolygon. During construction, we re-extract each target from the source PBF and verify its country, administrative areas, locality, street, named feature, and geometry. Point targets retain their native node coordinates, compact-area targets use deterministic interior representative points, and candidates that exceed the area or compactness constraints, contain invalid geometry, cross geographic boundaries, or lack a stable representative location are discarded. The surviving records must additionally pass address-completeness checks, visible-input collision resolution, and frozen-benchmark exclusion before entering the unified candidate pool. From this pool, we derive the SG, RG, AC, and SU SFT views and the corresponding RG, AC, and SU GRPO views using task-specific filters, fixed seeds, stable rankings, and, where appropriate, water-filling over countries,

Table 8: **Base-model-family comparison before and after geocoding training on Bench-C and Bench-A.** Base denotes the original model checkpoint. Trained denotes the matched $K = 10$, 100k-training-row, assigned-oracle setting. Median and P90 are computed over all coordinate predictions; hit rates use all samples.

		Bench-C					Bench-A				
Model	**Stage**	**Med.↓**	**P90↓**	**@1↑**	**@5↑**	**@25↑**	**Med.↓**	**P90↓**	**@1↑**	**@5↑**	**@25↑**
MiniCPM5-1B	Base	7230.096	14429.597	0.10	0.16	0.38	7645.404	14436.849	0.088	0.332	0.828
	Trained	114.405	827.681	1.12	7.02	25.10	99.215	705.020	1.112	8.208	25.208
MiniCPM5-2B	Base	339.311	3289.338	0.32	1.14	6.04	223.534	3350.049	1.724	5.436	14.448
	Trained	46.920	560.503	2.92	16.66	41.08	29.512	405.878	5.020	23.520	47.956
Qwen3.5-0.8B	Base	358.762	1784.561	0.38	2.76	9.56	223.306	1390.713	2.768	9.964	18.884
	Trained	35.050	561.350	4.92	25.16	46.66	21.621	426.558	7.760	31.664	51.436
Qwen3.5-2B	Base	156.571	814.946	0.92	6.64	19.60	89.242	645.009	5.028	17.276	32.336
	Trained	17.024	373.206	7.62	31.72	53.86	10.338	310.736	10.984	38.796	59.104
Qwen3.5-4B	Base	49.854	386.453	2.96	17.48	40.56	21.562	322.462	9.588	29.880	51.912
	Trained	7.979	215.679	11.52	41.78	64.76	5.659	207.684	16.576	47.680	68.172
Qwen3.5-9B	Base	24.936	319.269	5.34	24.92	50.10	11.268	253.731	12.624	37.272	59.948
	Trained	5.880	138.605	13.60	46.52	71.14	4.403	129.893	19.652	52.612	73.756
InternLM3-8B	Base	83.058	598.136	0.62	5.38	25.40	46.059	446.678	4.032	16.356	39.336
	Trained	26.421	283.773	2.26	17.36	48.88	16.532	248.971	5.576	27.152	56.984
Gemma3-4B-IT	Base	132.300	773.885	0.40	3.82	18.36	76.822	615.052	2.252	11.468	28.868
	Trained	15.429	212.736	4.38	28.48	59.04	11.203	187.743	6.536	33.844	63.532
Llama3.2-3B-Instruct	Base	70.907	554.037	0.70	7.54	30.26	40.086	470.398	4.212	18.384	42.388
	Trained	10.924	178.429	5.52	31.68	64.94	7.886	166.427	8.788	39.528	68.992

IARs, or feature types. Every frozen benchmark target and its declared leakage closure, including associated members, name variants, geometries, and target points, is excluded before sampling.

Data Reuse GeoMosaic permits controlled entity reuse both within individual datasets and across training objectives. Within-dataset reuse occurs primarily in the SU data, where multiple queries and reasoning templates are intentionally derived from the same geographic entity to improve robustness to different clue formulations. Across tasks, the same location may supervise address-to-coordinate prediction in SG or RG SFT, IAR assignment in AC SFT, coordinate or routing optimization in RG or AC GRPO, and clue composition or name disambiguation in SU SFT and SU GRPO. SG SFT remains entity-disjoint from the principal regional-coordinate and routing-RL datasets, whereas audited overlap is retained among RG SFT, AC SFT, RG GRPO, and AC GRPO when the same entity provides complementary supervision. AC SFT and AC GRPO remain mutually disjoint, while SU SFT and SU GRPO intentionally share a subset of entities to support a progression from supervised reasoning to reinforcement optimization. Such reuse does not amount to training repeatedly on identical examples, because the stages differ in their model-visible inputs, supervision targets, reward functions, and optimized components. Moreover, all frozen benchmark entities and their leakage closures remain excluded throughout construction; consequently, controlled reuse across training tasks improves the utility of verified geographic data without introducing evaluation leakage.

B.3 Benchmark Construction

GeoMosaic-Bench is constructed from the same pool of 2M candidates that have undergone raw OSM PBF re-extraction, address-completeness validation, geometry verification, and visible-input collision resolution. Before task generation, we apply stratified, capacity-constrained water-filling sampling over countries, IARs, and feature types to prevent a small number of data-rich countries or frequent facility categories from dominating the evaluation. The formal benchmark comprises three parts. Bench-C contains 5,000 direct-geocoding examples, each requiring a model to predict coordinates from a verified complete structured address containing the country, administrative areas, locality, street, and named feature. Bench-A contains five robustness tasks, A1–A5, with 5,000 examples per task. A1, A4, and A5 are paired transformations of the same Bench-C targets and preserve their entities, geometries, and coordinate labels; every transformed query must remain uniquely answerable in both the typed structured-address domain and the common-text serialization domain. Specifically, A1 removes either the locality or street while retaining the country, remaining administrative fields, and named feature; A4 deterministically composes at least two surface perturbations, including adjacent-character transposition, source-supported abbreviation, punctuation

Table 9: **Transfer matrix of the 10-partition experts.** Rows are deployed experts and columns are true IARs. Each cell is mean $\log(1 + d_{\mathrm{Hav}})$ on a 1,000-row diagnostic subset with 100 samples per IAR; lower is better. Bold marks the matched expert on the diagonal.

Expert / True	01	02	03	04	05	06	07	08	09	10
IAR-01	**1.6370**	3.5918	2.7432	2.3218	2.0990	1.6920	1.7922	1.7943	1.0842	2.0008
IAR-02	1.9677	**2.2461**	2.8198	2.3941	2.3091	1.7295	1.7758	1.8708	1.0030	2.1395
IAR-03	1.9662	3.6244	**1.5914**	2.3259	2.1513	1.7414	1.7873	1.7514	1.0418	2.0576
IAR-04	1.9255	3.6223	2.7123	**2.2900**	2.1185	1.7649	1.7768	1.8559	1.0440	2.0447
IAR-05	1.8598	3.6879	2.7153	2.3026	**2.0045**	1.7411	1.7704	1.8032	1.0286	2.0886
IAR-06	1.8962	3.7005	2.7633	2.3101	2.1351	**1.6200**	1.7521	1.7798	1.0479	2.0518
IAR-07	1.8785	3.6511	2.7380	2.2829	2.0948	1.6846	**1.6546**	1.7534	1.0424	1.9730
IAR-08	1.8584	3.5707	2.7845	2.3634	2.1009	1.7337	1.7856	**1.5830**	1.0391	2.0470
IAR-09	1.8933	3.7503	2.7473	2.3031	2.1601	1.7385	1.7343	1.7824	**0.9239**	2.0794
IAR-10	1.8901	3.6614	2.7799	2.3476	2.0591	1.7452	1.7479	1.7573	1.0309	**1.9205**

Table 10: **Inference efficiency.**

System	GPU	Prec.	Samples /s↑	Prompt tok./s↑	Output tok./s↑	Model mem. GiB/GPU↓	GPU util. (%) ↑	Peak BF16 TFLOP/s↑
OpenCage	–	–	**11.67**	–	–	–	–	–
GeoExperts	4×A40	BF16	8.81	10064.57	281.18	5.36	92	598.80

replacement, casing changes, duplicated whitespace, and Unicode decomposition; and A5 appends auxiliary form fields, with half of the examples repeating consistent address information and the remainder introducing target-independent content such as recipients, phone numbers without geographic country codes, generic organizations or departments, floors, rooms, and delivery instructions. Any A5 noise that introduces a new place name, conflicting address component, coordinate, entity identifier, or target hint is rejected. In contrast, A2 and A3 are constructed from independently verified source-backed name variants rather than synthetic transformations of Bench-C: A2 substitutes exact localized names from raw OSM `name:*` fields, whereas A3 uses verified official names or aliases from fields such as `official_name`, `int_name`, `alt_name`, `loc_name`, and `short_name`; synthetic translation, transliteration, rewriting, and alias fabrication are prohibited. Bench-N contains two geographic-reasoning tasks, N1 and N2, with 5,000 examples each. N1 identifies a target from two to four non-spatial facts without revealing its name, address, or coordinates, whereas N2 begins with a genuine shared-name group and provides only the ambiguous name together with a positive type or usage constraint for disambiguation. A2, A3, N1, and N2 admit only source-verifiable names, facts, and entities. After materialization, each benchmark undergoes independent structural, provenance, geometry, uniqueness, and quota checks. All subsequently constructed SFT and GRPO datasets must exclude the benchmark targets and their declared leakage closures, including associated members and name variants, thereby maintaining strict separation between training and evaluation.

B.4 More Cases

Fig. 5 presents training examples from Shared Geocoding Backbone SFT, Regional Geocoding Experts SFT, and Regional Geocoding Experts GRPO. These datasets use complete structured addresses as model inputs and coordinates as supervision or reward targets. Regional data are further partitioned by IAR to train distinct Regional Geocoding Experts. Address Classification Expert SFT and Address Classification Expert GRPO use the same type of address input but replace the coordinate target with a single partition label ranging from `IAR-01` to `IAR-10`. Fig. 6 presents examples from Semantic Understanding Expert SFT and Semantic Understanding Expert GRPO. Semantic Understanding Expert SFT is supervised using audited reasoning traces and canonical structured addresses, whereas Semantic Understanding Expert GRPO exposes only the query to the policy model and optimizes it according to the end-to-end geolocation result. Fig. 7 presents representative evaluation samples from the three components of GeoMosaic-Bench.

C More Information About Training and Evaluation

C.1 More Details of format reward

All GRPO experiments use a binary format reward with expert-specific output constraints. Let y denote the model output and $\mathcal{F}_e$ the set of valid outputs for expert $e \in \{\mathrm{RG}, \mathrm{AC}, \mathrm{SU}\}$. The format reward is defined as

$$R_{\mathrm{fmt}}^{(e)}(y) = \mathbb{I}[y \in \mathcal{F}_e]\,, \tag{7}$$

where $\mathbb{I}[\cdot]$ is the indicator function, giving a reward of 1 only when the output satisfies all format requirements for the expert and 0 otherwise. For the regional geocoding expert (RG), the output must be a valid JSON object containing exactly two fields, `latitude` and `longitude`, with no missing or additional fields. Both values must be finite numbers, with latitude in $[-90, 90]$ and longitude in $[-180, 180]$. The format reward weight is set to 0.2. For the address classification expert (AC), the output must be exactly one label from `IAR-01` to `IAR-10`, without explanations, JSON wrapping, or any other text. The format reward weight is set to 0.1. For the semantic understanding expert (SU), the output must be a valid JSON object containing only `reasoning` and `resolved_location`. The `reasoning` field must be a list of 2–5 non-empty strings. The `resolved_location` field must be a comma-separated structured address containing at least four non-empty address components. In addition to the format reward with a weight of 0.05, we introduce an output-type reward R_{type}, also weighted at 0.05, to check that the resolved location is a structured address rather than GPS coordinates or plain text. The auxiliary format and type reward for SU is therefore

$$R_{\mathrm{aux}}^{(\mathrm{SU})}(y) = 0.05\, R_{\mathrm{fmt}}^{(\mathrm{SU})}(y) + 0.05\, R_{\mathrm{type}}(y). \tag{8}$$

Table 11: **Key training hyperparameters of GeoExperts.** The base model is Qwen3.5-4B; regional SFT settings refer to the partition-scaling experiments. Effective batch sizes count examples for SFT and completions for GRPO. Token limits denote the full sequence length for SFT and separate prompt/completion limits for GRPO. A dash indicates that a setting does not apply.

Hyperparameter	SFT	GRPO
Global Backbone (Full-parameter SFT)		
Learning rate	1×10^{-4}	—
Effective batch size	1,024	—
Training epochs	1	—
LR schedule / warmup ratio	Cosine / 0.05	—
Regional Geocoding Experts		
Number of regional experts K	1, 5, 10, 20	10
LoRA rank / scaling parameter (r, α)	64 / 192	64 / 128
Learning rate	5×10^{-5}	1×10^{-5}
Effective batch size	128	64
Training epochs	4	1
KL penalty coefficient β	—	0.02
Address Classification Expert		
LoRA rank / scaling parameter (r, α)	128 / 256	128 / 256
Learning rate	1×10^{-5}	5×10^{-6}
Effective batch size	256	96
Training epochs	1	1
KL penalty coefficient β	—	0.05
Sampling temperature	—	1.0
Reward weights (format / exact IAR)	—	0.1 / 1.0
Semantic Understanding Expert		
LoRA rank / scaling parameter (r, α)	64 / 128	128 / 256
Learning rate	1×10^{-4}	2×10^{-6}
Effective batch size	128	112
Training budget	1 epoch	2,143 steps
LR schedule / warmup ratio	Cosine / 0.05	Cosine / 0.05
KL penalty coefficient β	—	0.05
Reward weights (format / output type / distance)	—	0.05 / 0.05 / 0.90
Shared Settings for Expert Adapters		
LoRA dropout	0.05	0.05
Completions per prompt G	—	8
Clipping thresholds $(\epsilon_{\mathrm{low}}, \epsilon_{\mathrm{high}})$	—	(0.20, 0.28)

C.2 HYPERPARAMETER SETTINGS

Key Training Hyperparameters. The key SFT and GRPO hyperparameters for the shared backbone and individual experts are listed in Tab. 11. These include learning rates, batch sizes, LoRA configurations, training budgets, and reward-related settings.

Geometric Filtering and Balanced Sampling. Target entities are restricted to named OSM nodes and compact Polygon or MultiPolygon objects. Point targets directly use their original OSM node coordinates. Area targets are retained only if they satisfy all of the following conditions: an area no greater than $250{,}000\,\mathrm{m}^2$, an iso-perimetric compactness of at least 0.15, a maximum distance of no more than $250\,\mathrm{m}$ from the interior representative point to any exterior vertex, and a conservatively estimated diameter of no more than $500\,\mathrm{m}$. We define isoperimetric compactness as $4\pi A/P^2$, where A and P denote the area and perimeter of the target, respectively. For each retained area target, we compute a deterministic spherical interior representative point as the coordinate ground truth, rather than using an unverified bounding-box center or vertex average.

Distance Metric and Invalid Outputs. We use the Haversine distance between the predicted coordinate and the frozen reference coordinate as the primary geocoding error. Given two points with latitudes and longitudes (ϕ_1, λ_1) and (ϕ_2, λ_2), respectively, we first compute

$$a = \sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right), \tag{9}$$

followed by

$$d = 2R\,\mathrm{atan2}\left(\sqrt{a}, \sqrt{1-a}\right), \tag{10}$$

where all angular values are converted to radians and the Earth radius is fixed to $R = 6371.0\,\mathrm{km}$. A prediction is considered valid only if it contains finite numeric coordinates with latitude in $[-90, 90]$ and longitude in $[-180, 180]$. Point targets use their OSM node coordinates as reference locations, whereas compact-area targets use deterministic interior representative points. For area targets, we additionally report whether the predicted point lies within the target geometry. Outputs that cannot be parsed, omit a required coordinate field, contain non-finite values, or fall outside the valid latitude or longitude ranges are marked as invalid. When computing aggregate distance metrics, each invalid output is assigned the largest error observed among the valid outputs within the same experimental group. An experimental group is defined by a fixed model, checkpoint, decoding configuration, and benchmark split.

C.3 TRAINING AND EVALUATION PROMPTS

We provide the prompt templates used for training and evaluation to specify each module's task instructions and output requirements. The prompts for the regional geocoding and address classification experts are shown in Fig. 8. The prompts for the semantic understanding expert and the teacher model used to generate reasoning traces are shown in Fig. 9.

D CASE STUDY

Fig. 10 and Fig. 11 present representative success and failure cases of GeoExperts on Bench-A and Bench-N, respectively. The successful cases in Fig. 10 show that, even when the street field is missing or irrelevant business-form fields are added, GeoExperts can still localize the target by combining the country, locality, street context, and named-feature information. The failure cases, however, reveal that the model remains sensitive to certain input variations. Fig. 11 further illustrates the two-stage behavior of GeoExperts on geographic reasoning tasks. The model can identify the Embassy of Peru in South Korea from the combined clues that the entity is operated by Peru, is an embassy, and is located in South Korea. Nevertheless, the failure cases show that even when the model generates the correct structured address, substantial errors may still occur during the subsequent address-to-coordinate conversion stage.

E DISCUSSION

E.1 LIMITATIONS

GeoMosaic's geographic coverage and annotation quality are partly influenced by its open data sources, and the current evaluation could better represent real user queries and low-resource regions. GeoExperts may be affected by routing and semantic parsing errors when handling complex inputs, while adapting to changing geographic information requires ongoing updates to both data and models. Future work could further improve generalization and deployment efficiency by continually expanding the collection of real user queries, improving uncertainty estimation, and refining mechanisms for incremental regional updates.

E.2 BROADER IMPACTS

GeoExperts offers a viable approach to geocoding without online geographic retrieval, supporting offline map search, geographic data organization, and location services in resource-constrained environments. Its shared backbone and regional adapters help reuse general geographic knowledge and adapt to regional address conventions at a low cost. By jointly handling standard addresses, non-standard addresses, and natural-language location descriptions, GeoExperts improves support for diverse inputs. For applications that can tolerate some location error, it offers a low-cost alternative to commercial geocoding APIs and may help make geographic information services more accessible. In addition, local inference reduces the need to transmit address data to third-party services, helping protect user privacy. At the same time, this technology carries potential risks of misuse, such as inferring an individual's location from scattered textual clues or assisting unauthorized location inference and tracking. Location errors and performance differences across regions and languages may also affect delivery services, public service coverage, and resource allocation. These effects could disproportionately impact regions with limited geographic data coverage and the people living there. Deployment should therefore include appropriate restrictions on improper uses of location inference, ongoing evaluation of location errors across regions and languages, and mechanisms for human review and correction. For high-risk applications such as emergency response and healthcare, model outputs should be independently verified and should not serve as the sole basis for decisions.

Shared backbone | Coordinate SFT

INPUT

```
Albania, Korçë, Rruga Viktimat e 7 Shkurtit, Alumil
```

SUPERVISION

```
{"latitude": 40.6188505,
 "longitude": 20.7722096}
```

Regional expert | Coordinate SFT

INPUT

```
Iran, تهران, تقی زاده, پارچه مبلی رویال فایپکو
```

SUPERVISION

```
{"latitude": 35.6599493,
 "longitude": 51.3482184}
```

Address classification | SFT

INPUT

```
Senegal, Dakar, Amitie 3, Amitie 3, Matrix Digital
```

SUPERVISION

```
IAR-02
```

Address classification | GRPO

INPUT

```
Greece, Αχαρνές, Αγίας Τριάδος, Σπιτικό Φαγητό
```

REWARD REFERENCE

```
IAR-06
```

Regional expert | Coordinate GRPO

INPUT

```
Hungary, Magyarpolány, Iskola utca, Német Nemzetiségi Általános Iskola
```

REWARD REFERENCE

```
{"latitude": 47.1678704,
 "longitude": 17.5439186}
```

Semantic understanding | GRPO / N1

INPUT

```
Find the courthouse in Salzburg, Austria. One verified source states that the feature has the heritage designation Listed objects in Austria, and another states that its verified address places it on Rudolfsplatz. Check all clues before returning the unique coordinates.
```

REWARD REFERENCE

```
{
  "latitude": 47.7970116,
  "longitude": 13.052592,
  "resolved_location":
"Austria, Salzburg,
Rudolfsplatz, Justizgebäude
Salzburg",
  "task_type": "N1"
}
```

Semantic understanding | GRPO / N2

INPUT

```
Select the location named "Fire Station Staff Building" that serves as fire station; return its complete structured address.
```

REWARD REFERENCE

```
{
  "latitude": 36.094417,
  "longitude": 44.6660684,
  "resolved_location":
"Iraq, Koya, Haibat Sultan -
Koysinjaq Road, Fire Station
Staff Building",
  "task_type": "N2"
}
```

Figure 5: **More Cases of GeoMosaic.**

Semantic understanding | CoT-SFT / N1

INPUT

The form mentions the document processing office, which is unrelated to the destination. Find the school in Ямбол, Bulgaria that was named after Lyuben Karavelov. Check all clues before returning the unique coordinates.

SUPERVISION

reasoning

1. The geographic scope is the city of Ямбол in Bulgaria, and the requested destination is a school.
2. The reference to the document processing office is explicitly excluded, so it does not constrain the destination.
3. The identifying clue is that the school was named after Lyuben Karavelov; in Ямбол this selects Основно училище „Любен Каравелов".
4. The resolved destination is Bulgaria, Ямбол Province, Ямбол, ул. „Любен Каравелов" № 31, Основно училище „Любен Каравелов".

resolved_location

Bulgaria, Ямбол, Надежда, ОУ Любен Каравелов

Semantic understanding | CoT-SFT / N2

INPUT

Identify "Sea Turtle Hatchery & Rescue Center" as a park location and return its complete structured address.

SUPERVISION

reasoning

1. The request seeks a park-type destination named "Sea Turtle Hatchery & Rescue Center."
2. That exact named attraction is located along Galle Road in Hikkaduwa, within Galle District in Sri Lanka's Southern Province.
3. The destination therefore resolves to the Hikkaduwa park and rescue facility, with the named feature retained as the final address component.

resolved_location

Sri Lanka, Hikkaduwa, Colombo-Galle Road, Sea Turtle Hatchery & Rescue Center

Figure 6: **More Cases of GeoMosaic.**

Bench-C | Complete address

INPUT

```
Mozambique
Maputo
Avenida 25 de Setembro
Palácio da Justiça da Cidade de Maputo
```

EVALUATION GROUND TRUTH

```
{"latitude": -25.9749361,
 "longitude": 32.5753578}
```

Bench-A / A1 | Missing street

INPUT

```
Country: Sierra Leone
Administrative area: Kailahun
Locality: Buedu
Named feature: Abdul Ahmid's Tailor Shop
```

EVALUATION GROUND TRUTH

```
{"latitude": 8.2804697,
 "longitude": -10.3709642}
```

Bench-A / A2 | Multilingual name

INPUT

```
Country: Lithuania
Locality: Vilnius
Street: Maironio g.
Named feature: Sint-Annakerk
```

EVALUATION GROUND TRUTH

```
{"latitude": 54.6831545,
 "longitude": 25.2932292}
```

Bench-A / A3 | Source-backed alias

INPUT

```
Country: Finland
Locality: Helsinki
Street: Aleksanterinkatu/Mannerheimintie
Named feature: Kolmen sepän patsas
```

EVALUATION GROUND TRUTH

```
{"latitude": 60.1686688,
 "longitude": 24.9409553}
```

Bench-A / A4 | Character and spacing noise

INPUT

```
Country: Pakistan
Locality: Sibi
Street:  کوئٹہ- جیکب آباد شاہراہ
Named feature: Baba  Hotel n2d portion
```

EVALUATION GROUND TRUTH

```
{"latitude": 29.4906121,
 "longitude": 67.8464728}
```

Bench-A / A5 | Irrelevant form fields

INPUT

```
Country: Vietnam
Administrative area: Lâm Đồng
Locality: Thành phố Đà Lạt
Street: Đường Đống Đa Phường 3
Named feature: New Century Dalat Hotel
delivery_note: ZR-115
department: ZX-831
floor: NQ-204
phone: 000-000-0000
recipient: NXQ-731
room: XM-508
```

EVALUATION GROUND TRUTH

```
{"latitude": 11.9237406,
 "longitude": 108.4432594}
```

Bench-N / N1 | Factual clues

INPUT

```
Identify the entity that satisfies the
following clues and return its
coordinates: 1) The entity is an instance
of tavern; 2) It was founded or built in
1884.
```

EVALUATION GROUND TRUTH

```
{"latitude": -36.8226711,
 "longitude": 174.748312}
```

Bench-N / N2 | Name disambiguation

INPUT

```
There are multiple locations corresponding
to "Sociedade são Vicente de Paula". I
want to visit the one used as place of
worship. Return the coordinates of the
location that best matches this request.
```

EVALUATION GROUND TRUTH

```
{"latitude": -19.9308278,
 "longitude": -44.106345}
```

Figure 7: **More Cases of GeoMosaic-Bench.**

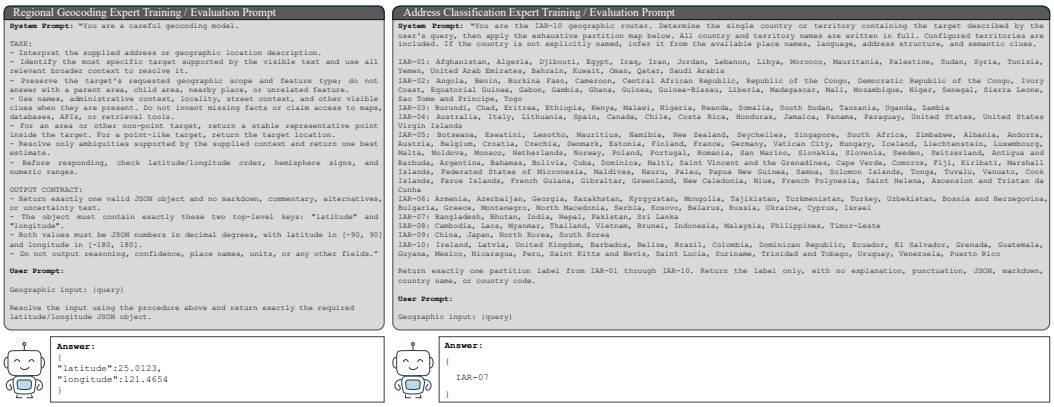


Figure 8: Prompts for the RG and AC experts.

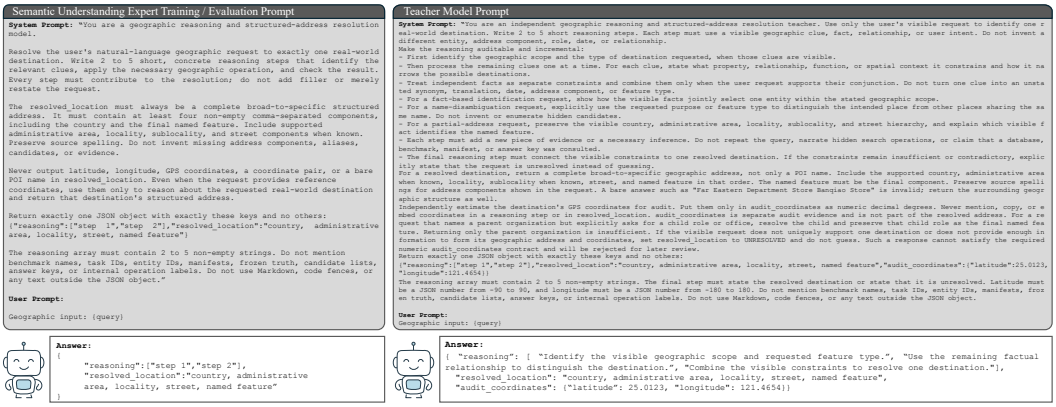


Figure 9: Prompts for the SU expert and the teacher model.



Figure 10: **Typical and failure cases of Bench-A.**

Success | N1 / Embassy of Peru in South Korea

USER

Identify the entity that satisfies the following clues and return its coordinates: 1) Its operator is Peru; 2) The entity is an instance of embassy; 3) The entity is located in South Korea.



GEOEXPERTS

Semantic understanding

reasoning

1. The operator clue identifies Peru as the country associated with the entity.
2. The entity is explicitly an embassy, so the destination is Peru's embassy rather than a general Peruvian organization.
3. The embassy is located in South Korea, specifically in Seoul's Jung-gu district on Sejong-daero.
4. The resolved destination is the Embassy of Peru in Seoul, South Korea.

resolved_location

South Korea, 중구, 서울특별시, 세종대로, 주한페루대사관

Coordinate prediction

{"latitude":37.562001,"longitude":126.977611}

GROUND TRUTH (37.5617746, 126.9750459) **0.23 km**

Failure | N1 / Embassy of Canada in Mali

USER

Identify the entity that satisfies the following clues and return its coordinates: 1) Its operator is Canada; 2) The entity is an instance of embassy; 3) The entity is located in Mali.



GEOEXPERTS

Semantic understanding

reasoning

1. The operator clue identifies Canada as the country associated with the entity.
2. The entity is explicitly an embassy, so the destination is Canada's embassy rather than a general Canadian organization.
3. The embassy is located in Bamako, Mali, on Avenue de la République.
4. The resolved destination is the Embassy of Canada in Bamako.

resolved_location

Mali, Bamako, Avenue de la République, Ambassade du Canada

Coordinate prediction

{"latitude":17.400001,"longitude":-149.556101}

GROUND TRUTH (12.6594984, -7.9630597) **14634.26 km**

Figure 11: **Typical and failure cases of Bench-N.**