# EVGeoQA: Benchmarking LLMs on Dynamic, Multi-Objective Geo-Spatial Exploration

**Jianfei Wu[1], Zhichun Wang[1,2,3]*, Zhensheng Wang[1], Zhiyu He[4],**
[1]School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China
[2]Beijing Key Laboratory of Artificial Intelligence for Education, Beijing 100875, China
[3]Engineering Research Center of Intelligent Technology and Educational Application, Ministry of Education, Beijing 100875, China
[4]College of Computer Science and Technology, National University of Defense Technology, Changsha 410073, China
{jianfeiwu, jensenwang}@mail.bnu.edu.cn, zcwang@bnu.edu.cn,hezhiyu99@nudt.edu.cn

## Abstract

While Large Language Models (LLMs) demonstrate remarkable reasoning capabilities, their potential for purpose-driven exploration in dynamic geo-spatial environments remains under-investigated. Existing Geo-Spatial Question Answering (GSQA) benchmarks predominantly focus on static retrieval, failing to capture the complexity of real-world planning that involves dynamic user locations and compound constraints. To bridge this gap, we introduce EVGeoQA, a novel benchmark built upon Electric Vehicle (EV) charging scenarios that features a distinct location-anchored and dual-objective design. Specifically, each query in EVGeoQA is explicitly bound to a user's real-time coordinate and integrates the dual objectives of a charging necessity and a co-located activity preference. To systematically assess models in such complex settings, we further propose GeoRover, a general evaluation framework based on a tool-augmented agent architecture to evaluate the LLMs' capacity for dynamic, multi-objective exploration. Our experiments reveal that while LLMs successfully utilize tools to address sub-tasks, they struggle with long-range spatial exploration. Notably, we observe an emergent capability: LLMs can summarize historical exploration trajectories to enhance exploration efficiency. These findings establish EVGeoQA as a challenging testbed for future geo-spatial intelligence. The dataset and prompts are available at https://github.com/kg-bnu/EVGeoQA.

## 1 Introduction

The rapid evolution of LLMs has propelled the development of autonomous agents capable of intricate planning and tool utilization (Yao et al., 2022; Schick et al., 2023; Xi et al., 2025). While LLMs excel in processing textual knowledge, grounding them in dynamic geo-spatial environments presents a unique challenge, primarily due to the inherent complexity and immense diversity of realistic spatial scenarios (Mai et al., 2023). Recent efforts in GSQA have attempted to bridge this gap, yet most existing benchmarks remain limited to static retrieval (Feng et al., 2023; Li et al., 2025). For instance, a typical query might ask, *"What is the distance from the airport to the central railway station?"*. Such tasks rely solely on static spatial topology, neglecting the complexity of real-world mobility where decision-making is constrained by the user's dynamic location and composite demands (Zheng et al., 2014).


Figure 1: An illustrative EVGeoQA query. Identifying the optimal target requires combining semantic understanding with real-time location and POI information.

The EV charging domain exemplifies this complexity and serves as an ideal yet under-explored testbed. Due to the extended wait times associated with EV charging compared to traditional refueling (Philipsen et al., 2018), users are inclined to bundle this service with secondary activities to utilize the duration efficiently. As illustrated in Figure 1, a typical inquiry exemplifies this coupled requirement: *"I want to go charge my car and grab a coffee."* Consequently, the optimal solution depends not merely on the charging station itself, but on satis-

*Corresponding author.



*Findings of the Association for Computational Linguistics: ACL 2026*, pages 3809–3821

fying a compound objective involving the station's location relative to the user's real-time coordinates and the suitability of the surrounding Point of Interest (POI) context.

To systematically evaluate the geo-spatial reasoning capabilities of LLMs within this rigorous settings, we introduce EVGeoQA, a benchmark built upon the EV charging domain and designed for purpose-driven geo-spatial exploration. Unlike traditional GSQA datasets (Li et al., 2025; Kefalidis et al., 2023; Punjani et al., 2018), each query in EVGeoQA is explicitly bound to a user's real-time coordinate and integrates the dual objectives of a charging necessity and a co-located activity preference. Distinguished by this unique design, EVGeoQA shifts the focus from static fact-checking to dynamic planning.

Our dataset covers three representative Chinese cities—Hangzhou, Qingdao, and Linyi—spanning a hierarchical gradient from major metropolis to developing city. Furthermore, regarding user location generation, Instead of using traditional random coordinate sampling, we propose a synthesis strategy based on K-Means clustering (Ahmed et al., 2020) that integrates population density and road network data. By weighting these factors, we simulate user coordinates that statistically align with authentic spatial query distributions, thereby mitigating the spatial bias inherent in random sampling.

With this realistic testbed established, we further propose GeoRover, a general evaluation framework based on a tool-augmented agent architecture to investigate LLMs' geo-spatial exploration capabilities. Specifically, we design four interactive geo-spatial tools to enable the agent to iteratively explore the environment, synthesize historical exploration trajectories, and derive final answers.

Our experiments demonstrate that while accuracy remains relatively high when the answer lies within a short distance, it deteriorates significantly as the exploration distance increases. For instance, the average Hits@2 scores decline from $\sim$50% to $\sim$38% as the explore radius expands. This performance degradation underscores the critical limitations of LLMs in long-range spatial exploration. Intriguingly, we observe a spontaneous phenomenon: even in the absence of explicit instructions, LLMs actively summarize historical exploration trajectories to enhance exploration efficiency. These findings establish EVGeoQA as a challenging benchmark for future geo-spatial intelligence.

In summary, our contributions are as follows:

- We introduce EVGeoQA, the first GSQA benchmark designed for dynamic multi-objective geo-spatial exploration. It uniquely integrates dynamic user locations with dual-objective constraints to evaluate LLM performance in geo-spatial tasks.
- We propose GeoRover, a general evaluation framework utilizing a tool-augmented agent equipped with interactive geo-spatial tools to enable active, multi-step exploration, thereby assessing LLM performance in this pervasive yet previously overlooked domain of multi-objective geo-spatial reasoning.
- Our empirical evaluation reveals that while current LLMs struggle with long-range spatial reasoning, they exhibit a latent ability to summarize historical trajectories, positioning EVGeoQA as a rigorous benchmark for future research.

## 2 Related Work

### 2.1 Benchmarks for GSQA

The landscape of GSQA has been shaped by foundational benchmarks such as GeoQA201 (Punjani et al., 2018), GeoQA1809 (Kefalidis et al., 2023), and subsequent works like MapQA (Li et al., 2025) and GeoQAMap (Feng et al., 2023), which predominantly focus on static retrieval over offline databases, neglecting the complexity of real-world environments. Driven by the rapid advancements in Embodied AI, benchmarks such as OpenEQA (Majumdar et al., 2024), SQA3D (Ma et al., 2022), and ScanQA (Azuma et al., 2022) have also catalyzed the development of the GSQA domain. However, they are primarily confined to small-scale indoor scenarios with static scene representations. To effectively address geo-spatial exploration at large scale, LLMs are required to possess synergistic capabilities in planning (Xie et al., 2024; Song et al., 2023), active exploration (Zhou et al., 2024), and information summarization (Chen et al., 2023; Liang et al., 2023). This multi-faceted requirement significantly escalates the complexity and challenges of the task.

### 2.2 Applications of LLMs in the GSQA Domain

Driven by exceptional inductive reasoning and information synthesis capabilities, LLMs have been broadly applied to a wide range of real-world tasks,

including financial analysis (Singh et al., 2024; Wang et al., 2025a), data annotation (Wu et al., 2025; Wang et al., 2024), and intelligent education (Sun et al., 2024; Lu et al., 2026). This proficiency in managing complex logical tasks is naturally extending into the geo-spatial domain to address the challenges of interpreting geographical information. To handle the unique spatial constraints and heterogeneous data inherent to this field, researchers have integrated LLMs with established autonomous agent frameworks such as ReAct (Yao et al., 2022), Toolformer (Schick et al., 2023), and ToolLLM (Qin et al., 2023). These integrations empower LLMs to function as autonomous agents capable of independent decision-making and tool-based interaction within geographic environments. Within this context, several specialized works have recently emerged to enhance geo-spatial reasoning capabilities. For instance, Spatial-RAG (Yu et al., 2025) introduces a spatial retrieval-augmented generation framework that utilizes a dual-retrieval strategy to resolve real-world spatial reasoning questions. CityGPT (Feng et al., 2025) focuses on empowering urban-scale spatial cognition by injecting structural knowledge of street networks and urban functional zones into model parameters through specialized instruction tuning. Our benchmark is fundamentally built upon these significant advancements and provides a specialized adaptation for GSQA scenarios.

# 3 The EVGeoQA Dataset

| City | Hangzhou | Qingdao | Linyi |
|---|---|---|---|
| **Stations** | 258 | 165 | 157 |
| **Locations** | 997 | 995 | 997 |
| **POIs** | 25 | 23 | 21 |
| **QA Pairs** | 19940 | 14162 | 14416 |

Table 1: Statistics of the EVGeoQA Dataset.

## 3.1 Problem Formulation

The core philosophy of this dynamic exploration task is distinct from traditional GSQA benchmarks: it embodies the behavioral pattern of "going to one place to do two things". Formally, unlike traditional GSQA queries that rely solely on geo-spatial interaction, a query $Q$ in EVGeoQA is explicitly anchored to a user's real-time coordinate $L_u$. The goal is to find a target charging station $S$ that simultaneously satisfies two constraints:

- Charging Necessity: The primary task, where the user explicitly requires charging services for their EV.
- Co-located Activity: The station must be within a walkable distance to a POI $P$ that fulfills the user's secondary intent.

## 3.2 Data Acquisition and Pre-processing

To ensure diversity across different urban scales, we selected three representative cities in China: Hangzhou (Provincial Capital), Qingdao (Regional Economic Hub), and Linyi (Prefecture-level City). To construct the geo-spatial foundation for the QA pairs within these regions, we integrate charging station records from the State Grid Corporation of China[1] with POI data within a 1km radius of each station retrieved via the Gaode API[2]. Figure 2(b) presents the categorical distribution of these contextual POIs.

## 3.3 User Location Synthesis via Multi-Source Fusion

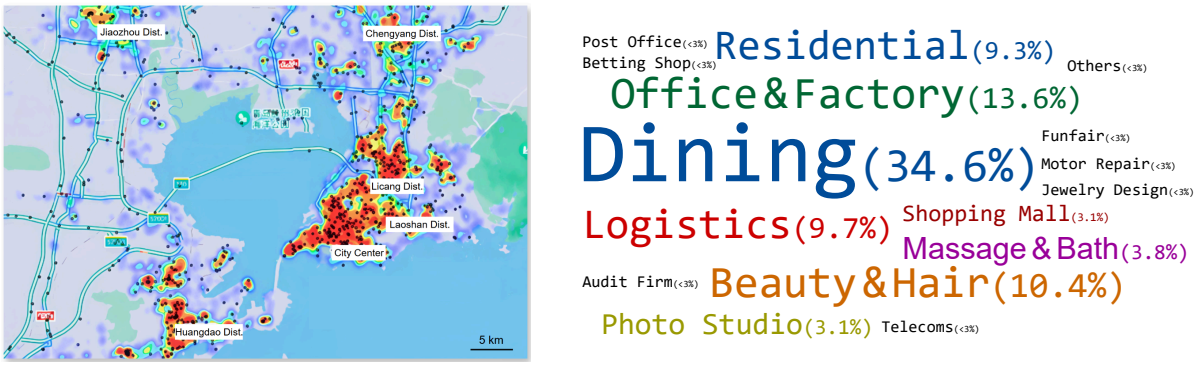
To generate realistic user locations, we synthesize coordinates by leveraging population flow and road network information derived from Baidu heatmap[3], as illustrated in Figure 2(a). Formally, we treat the raw heatmap image as a pixel set $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$, where each $x_i \in \mathbb{R}^3$ represents the RGB vector of the $i$-th pixel. We employ the K-Means algorithm (Ahmed et al., 2020) to partition these pixels into $K$ semantic clusters $\mathcal{C} = \{C_1, \ldots, C_K\}$, representing varying population density tiers and road contours, by minimizing the within-cluster sum of squares:

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \quad (1)$$

where $\mu_k$ denotes the centroid of cluster $C_k$. Subsequently, to simulate realistic human distribution, we assign a density score $w_k$ to each cluster based on its semantic importance. The user coordinates are then sampled from these clusters, where the probability $P(k)$ of sampling a location from cluster $C_k$ is determined by a Softmax function.

$$P(k) = \frac{\exp(w_k)}{\sum_{j=1}^{K} \exp(w_j)} \quad (2)$$

[1] http://www.evs.sgcc.com.cn/
[2] https://lbs.amap.com/
[3] https://map.baidu.com/




(a) (b)

Figure 2: (a) Distribution of query-anchored locations in Qingdao, reflecting the concentration within densely populated regions and along road networks. (b) Distribution of contextual POIs surrounding charging stations in Qingdao, showcasing categorical diversity and non-uniform spatial coverage.

The specific weight assignments $w_k$ are detailed in Appendix A.1.

### 3.4 Dual-Objective Query Generation

Inspired from prior work in template based QA generation (Johnson et al., 2017; Li et al., 2025; Wang et al., 2025b; Pampari et al., 2018), we develop a template-based pipeline to generate natural language queries that reflect the dual-objective nature of the task.

We first generate structured seed queries by instantiating predefined templates with a refined subset of semantically significant POI categories defined in Section 3.2. We provide more details about this process in Appendix A.2.

However, raw template-generated seed queries often lack linguistic diversity and purpose-driven context. To address this, we employ a powerful LLM (qwen2.5-72B (QwenTeam, 2024)) equipped with Few-Shot (Brown et al., 2020) and Chain-of-Thought (CoT) (Wei et al., 2022) prompting techniques to paraphrase these seeds. Crucially, this phase involves a functional mapping from static POI categories to realistic intents. For instance, a template slot containing "Stadium" is mapped to activities such as "running" or "exercising". We provide more details about this functional mapping in Appendix A.5.

### 3.5 Answer Generation and Quality Control

To establish ground truth answers, we perform an exhaustive match across all charging station to identify all possible candidates. We verify secondary intent alignment by computing cosine similarity between the query's POI slot and station POIs using the CoNAN embedding model[4], enforcing a stringent threshold of 0.85 to minimize false positives.

Moreover, acknowledging that real-world spatial planning often yields multiple valid optimal solutions, We rank all semantically valid stations by their vehicle driving distance to the user's query location and retain up to five distinct stations as the ground truth set.

Finally, we conducted a manual verification of approximately 1000 QA pairs sampled across all POI categories to guarantee the linguistic naturalness and logical correctness of the dataset.

### 3.6 Scalability and Representativeness

Our proposed QA generation pipeline exhibits high extensibility, allowing for seamless adaptation to broader geo-spatial exploration domains beyond the EV charging context. Furthermore, as visualized in Appendix A.3, the spatial distribution of charging stations spans a wide density spectrum, facilitating the evaluation of both fine-grained discrimination and long-range exploration, thereby ensuring strong universality.

## 4 GeoRover Evaluation Framework

While EVGeoQA benchmark lays the foundation for multi-objective geo-spatial exploration, effectively evaluating LLMs within this domain necessitates a specialized framework. In this section, we introduce GeoRover, a general evaluation framework based on a tool-augmented agent architecture to systematically investigate the geo-spatial exploration capabilities of LLMs.

[4] https://huggingface.co/TencentBAC/Conan-embedding-v2

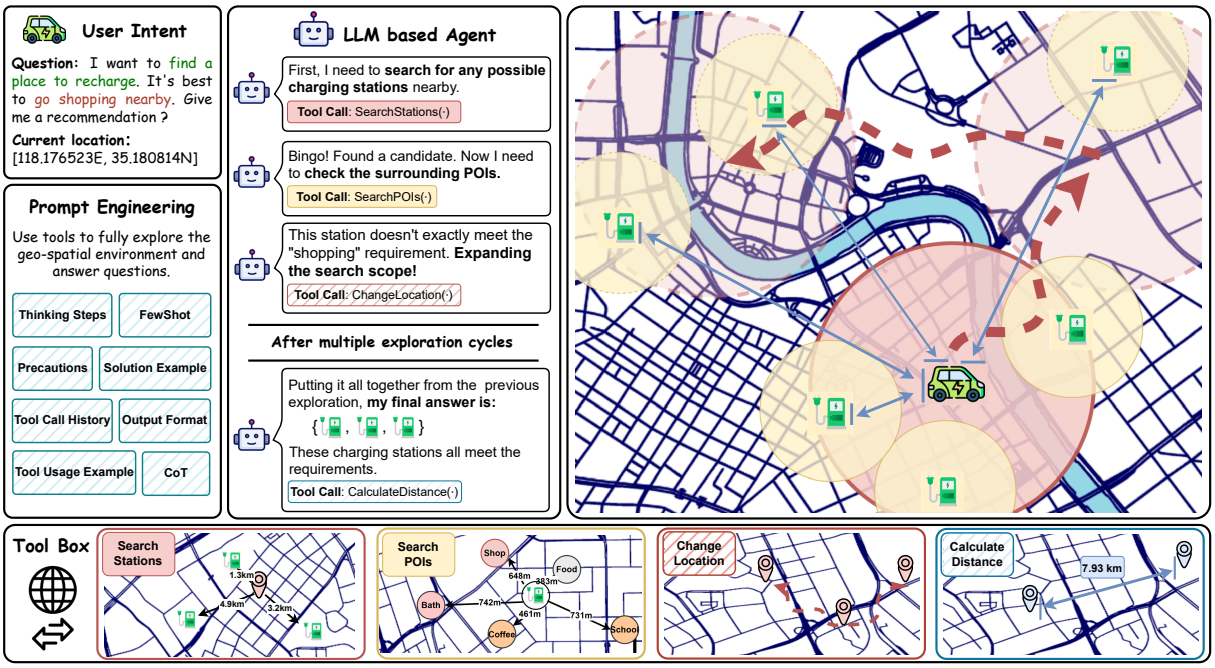



Figure 3: GeoRover Framework Overview. The agent leverages interactive tools to explore the geo-spatial environment, synthesizing the tool invocation trajectory to locate a station that satisfies both charging and activity demands.

### 4.1 Geo-Spatial Toolset Definitions

We prioritize the evaluation of the LLM's geo-spatial exploration capabilities rather than simple information extraction (Lewis et al., 2020; Fan et al., 2024). Consequently, to enforce a realistic setting of partial observability, we design four atomic tools that restrict the agent to obtaining only local information per interaction. This configuration compels the agent to perform iterative reasoning to resolve the query. The specific definitions for these tools are as follows:

**SearchStations Tool** Addressing the user's EV charging demand serves as the foundational step of the resolution process. Consequently, the agent is required to explore the environment to identify charging stations near the user's coordinate. To enable this retrieval capability, we design the `SearchStations` tool which allows the agent to perceive the distribution of charging stations within a localized 5km radius centered on a specific coordinate.

**SearchPOIs Tool** To verify secondary activity constraints, the agent is required to examine the local context of candidate stations to verify their alignment with specific user demands. The `SearchPOIs` Tool empowers the agent to inspect the POI context surrounding a specific coordinate, specifically retrieving POIs within a walkable 1km distance.

**ChangeLocation Tool** The `ChangeLocation` tool functions as the core mechanism of our framework, enabling the agent to actively explore the environment to acquire geographical information from a wider range. Specifically, this tool allows the agent to shift its position from the current coordinate in one of the four cardinal directions by an arbitrary distance. Upon execution, it returns the updated coordinate, which serves as the basis for the agent's subsequent decision-making. By utilizing this new position to re-invoke the `SearchStations` and `SearchPOIs` tools, the agent can effectively expand its perception scope, achieving the goal of autonomous spatial exploration.

**CalculateDistance Tool** To facilitate quantitative spatial reasoning, we equip the agent with the `CalculateDistance` tool. This tool computes the precise vehicle driving distance between coordinates and helps the agent better evaluate the cost-efficiency of candidate targets.

As illustrated in Figure 3, it is crucial to emphasize that the invocation sequence and frequency of these four tools are not pre-defined but are dynamically determined by the agent itself. The agent autonomously directs the exploration, synthesiz-

| City | Distance | <10km | | | <20km | | | No Limit | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Model** | **Hits@1** | **Hits@2** | **Hits@3** | **Hits@1** | **Hits@2** | **Hits@3** | **Hits@1** | **Hits@2** | **Hits@3** |
| **Hangzhou** | Qwen3-8B | 0.3663 | 0.4916 | 0.5381 | 0.2469 | 0.3421 | 0.3867 | 0.2061 | 0.2889 | 0.3295 |
| | Qwen3-8B* | 0.3965 | 0.4909 | 0.5011 | 0.3076 | 0.3972 | 0.4513 | 0.2637 | 0.3452 | 0.4375 |
| | Qwen3-30B-a3b | 0.3445 | 0.5251 | 0.5810 | 0.2718 | 0.3717 | 0.4012 | 0.2339 | 0.2942 | 0.3133 |
| | Qwen3-30B-a3b* | 0.3906 | 0.5396 | 0.5667 | 0.2804 | 0.3979 | 0.4285 | 0.2283 | 0.3215 | 0.3418 |
| | Qwen25-72b | 0.4833 | **0.6379** | 0.7073 | 0.3827 | **0.5479** | 0.6037 | 0.3333 | **0.4878** | 0.5614 |
| | GPT-OSS-20B* | 0.3906 | 0.5322 | 0.6379 | 0.3085 | 0.4302 | 0.5121 | 0.2612 | 0.3762 | 0.4479 |
| | GPT-OSS-120B* | 0.3417 | 0.4484 | 0.4828 | 0.3075 | 0.3432 | 0.3973 | 0.1842 | 0.2187 | 0.2415 |
| | Gemini-2.5-Flash | 0.4648 | 0.5214 | 0.6890 | **0.4121** | 0.4713 | 0.6147 | 0.3513 | 0.4114 | 0.5110 |
| | Gemini-2.5-Pro* | **0.4871** | 0.5398 | **0.7173** | 0.4034 | 0.4812 | **0.6373** | **0.3616** | 0.4304 | **0.5921** |
| | **Average** | 0.4073 | 0.5252 | 0.6024 | 0.3245 | 0.4203 | 0.4925 | 0.2693 | 0.3527 | 0.4196 |
| **Qingdao** | Qwen3-8B | 0.2687 | 0.4250 | 0.4875 | 0.2461 | 0.3423 | 0.3807 | 0.2132 | 0.2773 | 0.3161 |
| | Qwen3-8B* | 0.2361 | 0.3770 | 0.5245 | 0.2056 | 0.3295 | 0.4318 | 0.1401 | 0.2424 | 0.3129 |
| | Qwen3-30B-a3b | 0.2705 | 0.4176 | 0.4664 | 0.1653 | 0.2615 | 0.3012 | 0.1038 | 0.1872 | 0.2755 |
| | Qwen3-30B-a3b* | 0.2236 | 0.3802 | 0.4705 | 0.1891 | 0.3076 | 0.3663 | 0.1235 | 0.2125 | 0.2565 |
| | Qwen25-72b | 0.2638 | **0.5625** | **0.6206** | 0.2286 | **0.4555** | **0.4875** | 0.1678 | **0.3385** | **0.3805** |
| | GPT-OSS-20B* | 0.2712 | 0.4133 | 0.5894 | 0.2124 | 0.2996 | 0.3723 | 0.1513 | 0.2012 | 0.2819 |
| | GPT-OSS-120B* | 0.2254 | 0.3521 | 0.4703 | 0.2014 | 0.2477 | 0.3934 | 0.1433 | 0.2310 | 0.3173 |
| | Gemini-2.5-Flash | 0.3073 | 0.3974 | 0.4673 | 0.2857 | 0.3911 | 0.4198 | **0.2673** | 0.3350 | 0.3789 |
| | Gemini-2.5-Pro* | **0.3590** | 0.4218 | 0.5383 | **0.3106** | 0.3583 | 0.4467 | 0.2588 | 0.2738 | 0.3191 |
| | **Average** | 0.2695 | 0.4163 | 0.5150 | 0.2272 | 0.3326 | 0.4000 | 0.1743 | 0.2554 | 0.3154 |
| **Linyi** | Qwen3-8B | 0.3043 | 0.4637 | 0.5362 | 0.2376 | 0.3564 | 0.4257 | 0.1846 | 0.2923 | 0.3461 |
| | Qwen3-8B* | 0.4216 | **0.5929** | 0.6866 | 0.3325 | 0.4828 | 0.5674 | 0.3157 | **0.4416** | **0.5434** |
| | Qwen3-30B-a3b | 0.2899 | 0.4927 | 0.5507 | 0.2277 | 0.3861 | 0.4455 | 0.1769 | 0.3153 | 0.3615 |
| | Qwen3-30B-a3b* | 0.4512 | 0.5487 | 0.6671 | 0.3703 | 0.4537 | 0.5660 | 0.3253 | 0.4047 | 0.5161 |
| | Qwen25-72b | 0.4578 | 0.5122 | 0.6582 | 0.3669 | 0.4205 | 0.5773 | 0.3253 | 0.3821 | 0.5357 |
| | GPT-OSS-20B* | 0.3557 | 0.4496 | 0.4931 | 0.2810 | 0.3987 | 0.4407 | 0.2365 | 0.3494 | 0.3837 |
| | GPT-OSS-120B* | 0.3653 | 0.4722 | 0.4763 | 0.3121 | 0.4114 | 0.4218 | 0.2411 | 0.3414 | 0.3619 |
| | Gemini-2.5-Flash | 0.3993 | 0.5164 | 0.5778 | 0.3333 | 0.4173 | 0.4667 | 0.2962 | 0.3713 | 0.4225 |
| | Gemini-2.5-Pro* | **0.4887** | 0.5867 | **0.6989** | **0.3975** | **0.4898** | **0.6113** | **0.3827** | 0.4404 | 0.4985 |
| | **Average** | 0.3926 | 0.5150 | 0.5939 | 0.3177 | 0.4241 | 0.5025 | 0.2760 | 0.3709 | 0.4410 |

Table 2: Experimental results on the EVGeoQA dataset, with the best results in **bold**. LLMs employing the *Thinking* mechanism are marked with *.

ing historical observations to assess information sufficiency and determine when to terminate.

To further enhance the agent's task comprehension and reasoning stability, we incorporate the Few-Shot (Brown et al., 2020) and Chain-of-Thought (CoT) (Wei et al., 2022) prompting techniques. From an implementation perspective, all APIs are developed upon the Gaode platform [5]. To ensure experimental rigor, we apply strict filtering mechanisms to the raw API returns. This process eliminates irrelevant noise and aligns the retrieved data with the ground truth distributions of the EVGeoQA, thereby maximizing the accuracy of the evaluation results.

# 5 Experiment

## 5.1 LLMs Selection

To comprehensively evaluate the geo-spatial exploration capabilities of LLMs utilizing the EVGeoQA benchmark, we select a diverse set of LLMs representing different parameter scales and reasoning paradigms. Specifically, we include the Qwen series (Qwen3-8B, Qwen3-30B-a3b, and Qwen2.5-72B) (QwenTeam, 2024, 2025), the GPT-OSS series (20B and 120B) (OpenAI, 2025), and the Gemini-2.5 family (Flash and Pro) (Comanici et al., 2025). Furthermore, to investigate the specific impact of explicit reasoning on this task, we evaluate the "Thinking" variants for a subset of these LLMs, including Qwen3-8B, Qwen3-30B-a3b, GPT-OSS-20B, GPT-OSS-120B, and Gemini-2.5-Pro.

## 5.2 Evaluation Metrics

Given the multi-solution nature of real-world spatial planning, we employ Hits@$K$ ($K = 1, 2, 3$) as the primary metric to assess the accuracy of the recommended charging stations. Specifically, a prediction is considered a valid 'hit' if it matches any station within the ground truth set. To systematically analyze performance across different spatial scales, we split the dataset by geodesic distance between the user's location and the optimal target station. The samples are divided into three difficulty tiers:

- $<$ 10km (twice the search radius of the *SearchStations* Tool): Scenarios where the

[5] https://lbs.amap.com/

target is within a short driving radius.

- $<$ 20km (four times the search radius of the *SearchStations* Tool): Scenarios requiring medium-range planning.
- No Limit: The most challenging setting with no distance constraints.

### 5.3 Main Results and Analysis

As illustrated in Table 2, while large-scale models perform reasonably well in short-range scenarios, their performance drops significantly as the exploration distance increases. Evidently, there is a clear gap between current model performance and the requirements for reliable geo-spatial agents. We summarize three major findings below.

**LLM "laziness" induces insufficient exploration and performance degradation.** There is a consistent and pronounced performance degradation across all LLMs as the exploration radius expands. For instance, in Hangzhou, the average Hits@2 score drops from 0.5252 to 0.3527 as the evaluation shifts from the local scope to the No Limit setting. We attribute this decline primarily to a pervasive 'laziness' phenomenon: when faced with long-range exploration demands, LLMs often prematurely terminate the exploration process. Instead of conducting a sufficient exploration, they tend to fabricate seemingly plausible answers based on limited information retrieved from the previous steps.

**"Thinking" mechanisms enhance exploration depth via retrospective reflection.** We find that LLMs equipped with explicit "Thinking" modes consistently outperform their standard counterparts. For instance, in the Hangzhou No Limit setting, Qwen3-8B-thinking* achieves a Hits@2 score of 0.3452, showing a distinct advantage over the standard Qwen3-8B (0.2889). We attribute this efficacy to the model's capacity to reflect on historical search trajectories. Unlike standard models that are prone to premature termination, "Thinking" models actively evaluate the sufficiency of retrieved information against the dual-objective constraints and conduct additional exploration steps when information is deemed insufficient. We quantitatively analyze this phenomenon in Section 5.4.

**The Scaling Law persists in Geo-Spatial Exploration Task.** Large-scale foundation models, such as Qwen2.5-72B and Gemini-2.5-Pro, consistently dominate across all metrics. In contrast, smaller models (e.g., Qwen3-8B) exhibit a steeper performance decline as task difficulty increases, suggesting that limited parameter counts restrict the capacity to process the high-load spatial contexts inherent to dense urban environments.

### 5.4 Analysis of ChangeLocation Tool Usage

| City | Distance | <10km | <20km | No Limit |
|---|---|---|---|---|
| Linyi | Qwen3-8b | 0.35 | 1.36 | 2.12 |
| | Qwen3-8b* | 0.79 | 2.11 | 4.03 |
| | Qwen3-30b-a3b | 0.37 | 1.13 | 3.26 |
| | Qwen3-30b-a3b* | 0.42 | 1.72 | 3.74 |
| | Gemini-2.5-Flash | 0.40 | 2.77 | 3.91 |
| | Gemini-2.5-Pro* | 1.12 | 3.31 | 5.62 |
| | Average | 0.58 | 2.07 | 3.78 |

Table 3: Average tool invocation frequency of the `ChangeLocation` Tool across different distance tiers in Linyi.

As discussed in Section 4, the `ChangeLocation` tool is the core mechanism for expanding exploration scope. To quantitatively assess how LLMs utilize this capability, we record its average invocation frequency across different difficulty tiers in Linyi. Specifically, this metric is defined as the mean number of times the `ChangeLocation` tool is invoked by the agent within a single exploration episode.

The results in Table 3 reveal that the invocation frequency is much lower than anticipated, especially on long-range tasks. We attribute this to two primary factors. First, consistent with the "laziness" bottleneck, agents often terminate exploration prematurely without exhaustively exploring the environment. Second, we observe an emergent capability in advanced models to synthesize spatial contexts from interaction history and infer new coordinates without explicitly invoking the tool.

Despite these nuances, a distinct positive correlation exists between the frequency of tool usage and the agent's performance in complex scenarios. This increased active exploration directly aligns with the superior accuracy reported in our main results in Table 2, confirming that active exploration is a determinant factor for success in large-scale geo-spatial planning tasks.

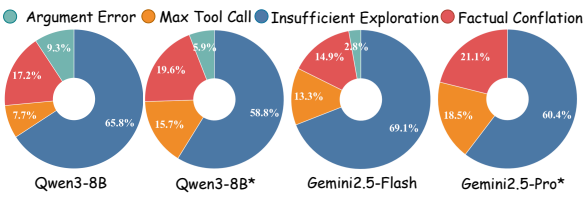



Figure 4: Distribution of Error Causes in Linyi


Figure 5: Case Study: Visualization of a multi-step exploration trajectory by Gemini-2.5-Pro* in Qingdao.

### 5.5 Error Analysis

As illustrated in Figure 4, a significant portion of failures stems from the insufficient exploration across the environment. All LLMs-based agents exhibit a substantial degree of "laziness" when facing complex planning tasks. This suggests that current LLMs lack the capacity to maintain a long-horizon search strategy without explicit guidance or reinforcement.

Furthermore, we observe that the integration of heterogeneous data, such as exploration trajectories, charging stations and POI details, triggers the "Lost in the Middle" phenomenon (Liu et al., 2024; Li et al., 2024). Agents frequently conflate attributes in these complex long contexts, producing responses that are linguistically fluent but factually erroneous.

Finally, Except for Gemini2.5-Pro, all evaluated LLMs encounter argument-related errors to varying degrees, indicating that the effective invocation of geo-spatial tools remains a significant challenge.

The detailed definitions and settings for these error categories are provided in Appendix A.4.

### 5.6 Case Study

We also qualitatively examine how agents navigate and reason within these complex geo-spatial environments. As shown in Figure 5, high-performing LLMs can actively summarize historical exploration trajectories to optimize future search steps.

Specifically, the agent initially executes a localized exploration surrounding the user's query coordinate (Steps 1–4). Upon determining that the nearby area lacks charging stations that satisfy the dual constraints, the agent autonomously performs long-range spatial transfers (Step 5 and 9) to its exploration scope. Although our prompt design provides no predefined exploration rules, the agent appears to synthesize its historical trajectory, selecting new anchor points that maximize spatial coverage and avoid redundant search efforts. Finally, the agent leverages the `CalculateDistance` tool to acquire precise distance metrics, synthesizing all accumulated observations to formulate the final recommendation.

This emergent behavior highlights the potential of LLMs to comprehend spatial layouts and conduct purpose-driven exploration. However, the agent's exploration process exhibits distinct limitations. As observed in the left quadrant of Figure 5, a qualified charging station located closer to the initial position is overlooked during the exploration process. This omission aligns with the "laziness" phenomenon discussed in Section 5.3. This suggests that while LLMs exhibit potential spatial reasoning, their ability to guarantee global optimality in geo-spatial exploration remains a critical bottleneck requiring future optimization.

## 6 Conclusion and Discussion

In this paper, we introduced EVGeoQA, the first benchmark designed to evaluate the purpose-driven exploration capabilities of LLMs within dynamic geo-spatial environments. To facilitate systematic assessment, We further propose GeoRover, a general evaluation framework adopting a tool-augmented agent architecture to investigate LLMs' geo-spatial exploration capabilities.

Our experimental results reveal that while LLMs perform effectively in localized, short-range scenarios, they suffer from pronounced performance degradation in long-range tasks. Although a clear gap exists between current model performance and the requirements for reliable geo-spatial agents , the latent ability of trajectory summarization highlights a significant potential for LLM-based complex geo-spatial reasoning.

By exposing key bottlenecks such as insufficient exploration and attribute conflation, EVGeoQA serves as a rigorous testbed to guide the development of more robust and strategically-aware geo-spatial agents for open-world applications.

## Limitations

The EVGeoQA benchmark is constructed based on urban data from three distinct Chinese cities of varying scales, with QA pairs predominantly in Chinese. This linguistic specificity may introduce inherent biases. Our future work aims to mitigate this by incorporating multi-lingual data and a broader range of global cities.

Furthermore, our current evaluation primarily assesses the inherent reasoning capabilities of LLMs. We acknowledge that advanced techniques, such as domain-specific fine-tuning (SFT) (Gururangan et al., 2020; Zheng et al., 2024; Hu et al., 2022), hold significant potential for enhancing model performance in specialized spatial tasks, and investigating the feasibility and efficacy of such optimization strategies constitutes a key direction for our subsequent research. Moving forward, we remain dedicated to the field of geo-spatial exploration, striving to refine our benchmarks and frameworks to further advance embodied spatial intelligence.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62276026) and the Fundamental Research Funds for the Central Universities (No. 2253500001)

## References

Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam. 2020. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8):1295.

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. 2022. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *NIPS*, volume 33, pages 1877–1901.

Howard Chen, Ramakanth Pasunuru, Jason Weston, and Asli Celikyilmaz. 2023. Walking down the memory maze: Beyond context limit through interactive reading. *arXiv preprint arXiv:2310.05029*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6491–6501.

Jie Feng, Tianhui Liu, Yuwei Du, Siqi Guo, Yuming Lin, and Yong Li. 2025. Citygpt: Empowering urban spatial cognition of large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 591–602.

Yu Feng, Linfang Ding, and Guohui Xiao. 2023. Geoqamap-geographic question answering with maps leveraging llm and open knowledge base. In *12th International Conference on Geographic Information Science (GIScience 2023)*, volume 277, page 28.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Sergios-Anestis Kefalidis, Dharmen Punjani, Konstantinos Plasa Eleni Tsalapati, Mariangela Pollali, Myrto Tsokanaridou Michail Mitsios, Manolis Koubarakis, and Pierre Maret. 2023. Benchmarking geospatial question answering engines using the dataset geoquestions1089. In *The Semantic Web - ISWC 2023 - 22nd International Semantic Web Conference, Athens, Greece, November 6-10, 2023, Proceedings, Part II*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024. Long-context llms struggle with long in-context learning. *Transactions on Machine Learning Research*.

Zekun Li, Malcolm Grossman, Mihir Kulkarni, Muhao Chen, Yao-Yi Chiang, and 1 others. 2025. Mapqa: Open-domain geospatial question answering on map data. *arXiv preprint arXiv:2503.07871*.

Xinnian Liang, Bing Wang, Hui Huang, Shuangzhi Wu, Peihao Wu, Lu Lu, Zejun Ma, and Zhoujun Li. 2023. Unleashing infinite-length input capacity for large-scale language models with self-controlled memory system. *arXiv preprint arXiv:2304.13343*, 10.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Tong Lu, Zhichun Wang, Yaoyu Zhou, Yiming Guan, Zhiyong Bai, and Junsheng Du. 2026. Scimkg: A multimodal knowledge graph for science education with text, image, video and audio. 40(18):15466–15474.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. 2022. Sqa3d: Situated question answering in 3d scenes. In *The Eleventh International Conference on Learning Representations*.

Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, Chris Cundy, Ziyuan Li, Rui Zhu, and Ni Lao. 2023. On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv preprint arXiv:2304.06798*, abs/2304.06798.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, and 1 others. 2024. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16488–16498.

OpenAI. 2025. gpt-oss-120b and gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368.

Ralf Philipsen, Teresa Brell, Waldemar Brost, Teresa Eickels, and Martina Ziefle. 2018. Running on empty – users' charging behavior of electric vehicles versus traditional refueling. *Transportation Research Part F: Traffic Psychology and Behaviour*.

Dharmen Punjani, K. Singh, Andreas Both, Manolis Koubarakis, I. Angelidis, Konstantina Bereta, Themis Beris, Dimitris Bilidas, Theofilos Ioannidis, Nikolaos Karalis, Christoph Lange, D. Pantazi, Christos Papaloukas, and Georgios I. Stamoulis. 2018. Template-based question answering over linked geospatial data. *Proceedings of the 12th Workshop on Geographic Information Retrieval*.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, and 1 others. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *The Twelfth International Conference on Learning Representations*.

QwenTeam. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

QwenTeam. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.

Kuldeep Singh, Simerjot Kaur, and Charese Smiley. 2024. Finqapt: Empowering financial decisions with end-to-end llm-driven question answering pipeline. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 266–273.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2023. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2998–3009.

Jianing Sun, Zhichao Zhang, and Xueli He. 2024. Llm4edukg: Llm for automatic construction of educational knowledge graph. In *2024 International Conference on Networking and Network Applications (NaNA)*, pages 269–275. IEEE.

Jingru Wang, Wen Ding, and Xiaotong Zhu. 2025a. Financial analysis: Intelligent financial data analysis system based on llm-rag. *arXiv preprint arXiv:2504.06279*.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI conference on human factors in computing systems*, pages 1–21.

Zhensheng Wang, Wenmian Yang, Kun Zhou, Yiquan Zhang, and Weijia Jia. 2025b. Retqa: A large-scale open-domain tabular question answering dataset for real estate sector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25452–25460.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jianfei Wu, Xubin Wang, and Weijia Jia. 2025. Enhancing text annotation through rationale-driven collaborative few-shot prompting. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, and 1 others. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101.

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.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Dazhou Yu, Riyang Bao, Gengchen Mai, and Liang Zhao. 2025. Spatial-rag: Spatial retrieval augmented generation for real-world spatial reasoning questions. *arXiv e-prints*, pages arXiv–2502.

Jiawei Zheng, Hanghai Hong, Xiaoli Wang, Jingsong Su, Yonggui Liang, and Shikai Wu. 2024. Fine-tuning large language models for domain-specific machine translation. *CoRR*.

Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. 2014. Urban computing: Concepts, methodologies, and applications. *ACM Trans. Intell. Syst. Technol.*, 5:38:1–38:55.

Gengze Zhou, Yicong Hong, and Qi Wu. 2024. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7641–7649.

# A Appendix

## A.1 User Location Synthesis via Multi-Source Fusion

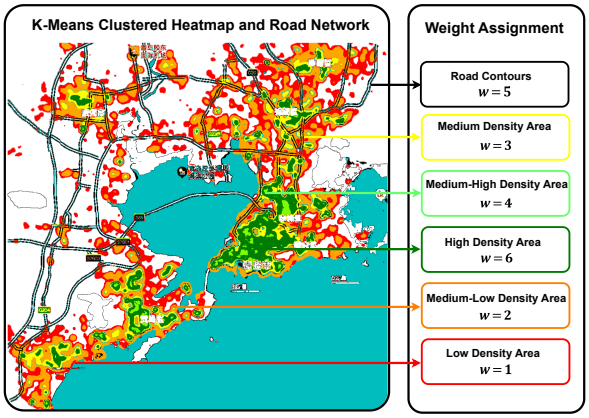



Figure 6: Illustration of the Multi-Source Fusion strategy for user location synthesis.

The left panel displays the spatial segmentation of density tiers and road contours derived via K-Means clustering (exemplified in Qingdao). The right panel details the semantic weight assignment ($w$), where higher weights are explicitly allocated to road networks and high-density regions to prioritize their selection. These weights drive a non-uniform sampling mechanism to generate realistic user coordinates, which subsequently serve as spatial anchors for question generation.

## A.2 Template Based Dual-Objective Query Generation

In this section, we present the templates used for generating QA pairs for the EV charging scenario. The placeholder `{position}` denotes the Point of Interest (POI) category in the vicinity of the charging station (e.g., "shopping mall", "park". The templates are presented in their **original Chinese format** as follows:

- 帮我找一个在{position} 附近的充电桩.
- 我想去{position} ,并顺便给我的电动汽车充电. 你能不能给我推荐一个充电桩?

- 我要去{position} 办点事，顺便想给我的电动车充电。哪里比较合适?
- 有在{position} 附近的充电桩吗?
- 你能把{position} 区域的充电桩位置给我展示一下吗?
- 在{position} 附近，有推荐的充电桩吗?
- 帮我推荐一个距离{position} 最近的充电桩?

We also provide the **English translations** of these templates for reference:

- Help me find a charging station near {position}.
- I plan to go to {position} and charge my EV. Can you recommend a charging station?
- I am going to {position} to run some errands and want to charge my electric vehicle. Where would be a suitable place?
- Are there any charging stations near {position}?
- Can you show me the locations of charging stations in the {position} area?
- Do you have any recommended charging stations near {position}?
- Please recommend the charging station closest to {position}.

Although these seed templates appear relatively simple and exhibit a degree of linguistic redundancy, their combination with the advanced LLM-based paraphrasing strategy (as detailed in Section 3.4) and a diverse spectrum of POI categories ensures a comprehensive coverage of user query variations within the EV charging domain.

### A.3 Spatial Distribution of Charging Stations

To demonstrate the topological diversity of the EV-GeoQA benchmark, we visualize the spatial density of charging stations across the three selected cities: Hangzhou, Qingdao, and Linyi. As illustrated in Figure 7, the distribution of charging infrastructure exhibits a strong correlation with regional economic development and population density.

Specifically, the distribution patterns reveal two distinct spatial characteristics that correspond to different challenges in geo-spatial reasoning:

- **High-Density Urban Cores:** In economically developed city centers and densely populated districts, charging stations are clustered with high density. In these scenarios, the agent must filter through numerous candidate stations in close proximity to identify the specific one that optimally satisfies the user's secondary activity constraint (e.g., determining which of the five nearby stations is closest to a specific type of restaurant).
- **Sparse Peripheral Regions:** Conversely, in suburban areas and developing districts, the distribution of stations becomes significantly sparser. These scenarios simulate long-range exploration tasks, where the agent cannot rely on immediate availability. Instead, it must engage in multi-step planning and active map traversal to locate feasible resources.

### A.4 Definitions and Settings for Error Analysis

We categorize the observed errors into four distinct types. Detailed definitions and their causes are provided below:

- **Argument Error**: This error occurs when an agent generates a response in an incorrect format that violates the tool's API format.
- **Max Tool Call**: This category refers to exploration process that terminate because the agent reaches the maximum limit of 20 tool invocations. This phenomenon typically arises from two scenarios: (1) the model falls into a logical loop by repeatedly invoking the same tool with identical parameters, or (2) the exploration strategy is inefficient, causing the agent to exhaust its call budget before identifying a valid target.
- **Insufficient Exploration**: This failure is characterized by premature termination of the search process. Despite failing to locate a charging station that satisfies all constraints during the exploration, the agent ceases its exploration and provides a final incorrect response. This aligns with the "laziness" bottleneck discussed in our primary analysis.
- **Factual Conflation**: This error is primarily driven by hallucinations or the "Lost in the Middle" effect. In these cases, although all necessary information to resolve the query has

Hangzhou
(Provincial Capital)

Qingdao
(Regional Economic Hub)

Linyi
(Prefecture-level City)

Figure 7: Spatial distribution of charging stations across the three representative cities. Visualizations are zoomed into city centers for clarity. The full dataset is available in our code repository.

been retrieved during the exploration steps, the agent fails to correctly synthesize the context, leading to a factually erroneous output despite the availability of valid evidence.

### A.5 LLM-based Query Paraphrasing and Functional Mapping

In real-world geo-spatial QA scenarios, user queries typically exhibit high linguistic variability and implicit intent. Unlike structured seed queries, users rarely specify static POI categories (e.g., *"Find a charging station near a restaurant"*); instead, they tend to express functional needs or activities (e.g., *"I need to charge my car while having dinner"*).

To bridge the semantic gap between template-generated seeds queries and authentic user inquiries, we implement a functional mapping strategy. Specifically, we employ Qwen2.5-72B, a large language model with strong instruction-following capabilities, to paraphrase the initial queries. The rewriting process is governed by three key principles:

1. **Intent Transformation:** Converting explicit location types into functional descriptions (e.g., mapping "Gym" to "working out").

2. **Contextual Logic:** Ensuring the activity is logically compatible with the charging duration and location type.

3. **Linguistic Diversity:** Varying sentence structures and tones to mimic casual, spoken language.

To vividly illustrate this transformation, we alse provide three representative examples comparing the raw template-generated seeds queries with their LLM-polished counterparts:

- **Case 1 (Dining):**
  *Template:* "Help me find a charging station near {Dinning/Restaurant}"
  *Polished:* "I'm starving and my car is running low. Is there a place where I can eat and charge simultaneously?"

- **Case 2 (Exercise):**
  *Template:* "I plan to go to {Stadium/Sports Venue} and charge my EV. Can you recommend a charging station? "
  *Polished:* "I want to get a workout in while waiting for my EV to charge. Can you recommend a spot?"

- **Case 3 (Accommodation):**
  *Template:* "Do you have any recommended charging stations near {Hotel/Accommodation}?
  *Polished:* "I'm looking for a place to rest for the night, and my car needs power too. Can you help me find a suitable location?"