MapSearchBench: Evaluating and Enhancing LLMs for Intelligent Map Search with Real-World User Interactions

Anonymous ACL submission

Abstract

Online map search is moving beyond keyword retrieval toward conversational and agentic interaction, where users express spatial constraints, local-service needs, and multi-step planning intents in natural language. Existing geospatial and local-search benchmarks do not fully capture this pipeline, which involves query routing, intent understanding, evidence-grounded response generation, and end-to-end system behavior. We introduce **MapSearchBench**, a real-world benchmark built from anonymized user interactions collected from a large-scale commercial map application. It evaluates LLMs across staged map-search tasks, including scene routing, macro and micro intent recognition, controlled response generation, and end-to-end response quality. MapSearchBench combines accuracy-based metrics with five-dimensional response scoring, multi-judge evaluation, and controlled pairwise comparison. Experiments across 17 models and 4 complete systems show that domain adaptation substantially improves map-search understanding, while response generation and end-to-end performance depend on evidence synthesis, grounding, and service-aware system design. These findings suggest that intelligent map search requires diagnostic evaluation across pipeline stages rather than a single aggregate score.

1 Introduction

Online map search is becoming an increasingly important interface between users and the physical world. Users no longer interact with map services only through short keyword queries such as place names or addresses. Instead, they ask conversational and often underspecified questions, such as finding nearby services under time constraints, searching for places that satisfy multiple spatial and semantic conditions, or planning multi-step activities across locations. Large language models (LLMs) provide a promising foundation for such intelligent map search because they can interpret natural language, reason over constraints, and generate user-facing responses.

However, intelligent map search is not a standard question-answering task. A map-search system must first decide how a query should be processed, then identify the user's semantic and spatial intent, retrieve or use map-specific evidence, and finally generate a response that is accurate, useful, and compliant with map-service constraints. These stages require different capabilities. For example, routing a query to a lightweight retrieval engine is different from recognizing that a request requires agentic search; identifying whether a user is searching nearby, along a route, indoors, or at city scale is different from determining whether the user wants a place, a product, or geographic knowledge. A benchmark that only evaluates final answers cannot reveal where a model fails in this pipeline.

Existing benchmarks only partially cover this space. Geographic question answering and geospatial reasoning benchmarks study whether models understand geographic facts or spatial relations (Mai et al., 2021; Roberts et al., 2023; Dihan et al., 2024). Urban and multimodal benchmarks evaluate broader city-level or visual urban understanding (Feng et al., 2025b; Zhou et al., 2025). Recent local-service and agentic-search benchmarks move closer to real user needs (He et al., 2025; Lan et al., 2025), while general agent benchmarks examine tool use, browsing, or multi-step problem solving in open domains (Mialon et al., 2023; Yao et al., 2024; Wei et al., 2025; Du et al., 2025). These efforts are valuable, but they do not fully evaluate the decision chain of real map search. As a result, it remains unclear whether current LLMs fail because they misunderstand the query, choose the wrong processing path, misuse retrieved evidence, or lack sufficient evidence coverage.

To address this gap, we introduce **MapSearch-**

Bench, a benchmark for evaluating LLMs in realistic intelligent map search. MapSearchBench is constructed from anonymized real-world user interactions collected from a large-scale commercial map application and annotated by domain experts. It organizes evaluation into task-specific subsets with 400 labels for scene routing, 556 labels for macro intent recognition, 629 labels for micro intent recognition, 855 controlled response-generation cases, and 73 end-to-end response cases. The benchmark follows the structure of a practical map-search pipeline. It first evaluates whether a model can route a query to the appropriate processing tier, then tests whether it can infer both coarse-grained intent and fine-grained where/what attributes, and finally assesses response quality under controlled and end-to-end settings.

MapSearchBench further provides a multi-layered evaluation protocol. For query understanding, we use accuracy-based metrics at different granularities. For controlled response generation, we evaluate models using a five-dimensional rubric covering spatial-temporal integration, user demand satisfaction, platform compliance, retrieval/result processing, and output presentation. We use multiple independent LLM judges and blind pairwise comparison to reduce reliance on a single scoring scale. For end-to-end evaluation, we assess complete systems using absolute multi-dimensional scores, allowing us to examine how search, grounding, and generation interact in realistic usage.

The experimental results highlight why a staged benchmark is necessary. Domain-specific supervision yields large gains on query understanding, especially for routing decisions and fine-grained where/what intent prediction. At the same time, response generation under shared retrieved evidence depends more on a model's ability to filter, synthesize, and present map evidence, where strong general-purpose LLMs remain highly competitive with our response component. In the end-to-end setting, our production-style map-search pipeline achieves the second-best absolute score among four complete systems, suggesting that grounded retrieval and service-aware decision making remain valuable in realistic use. Taken together, these findings show that intelligent map search cannot be assessed by a single aggregate score. Routing, intent understanding, evidence-grounded response generation, and system-level search behavior need to be evaluated separately.

Our contributions are summarized as follows:

- We introduce **MapSearchBench**, a real-world benchmark for intelligent map search that covers query routing, macro and micro intent recognition, controlled response generation, and end-to-end response evaluation.
- We design a diagnostic evaluation protocol that combines task-specific accuracy metrics, five-dimensional response scoring, multi-judge evaluation, and end-to-end system assessment.
- We provide a broad empirical study across query-understanding models, response generators, and complete map-search systems, including a domain-adapted Hunyuan-7B model that demonstrates the effectiveness of MapSearchBench-style supervision for lightweight map-search understanding.

We will release the MapSearchBench data, labels, prompts, and evaluation scripts immediately upon paper acceptance.

2 Methods

2.1 Overview

MapSearchBench is designed to evaluate LLMs in realistic intelligent map search, where a model must not only understand geographic expressions but also decide how a query should be processed and how the final response should be presented. Unlike conventional geospatial QA benchmarks that mainly test factual knowledge or isolated spatial reasoning, map search is a pipeline task. A user query may require lightweight keyword retrieval, semantic intent parsing, or multi-step agentic search. To capture this structure, MapSearchBench separates evaluation into two complementary parts, query understanding and response quality assessment. The query understanding part evaluates whether a model can route a query to the appropriate processing tier and convert it into executable semantic and spatial intent labels. The response quality part evaluates whether a model or system can generate useful, grounded, and well-structured map-search answers.

Figure 1 summarizes the evaluation framework. For query understanding, we report accuracy-based metrics for scene routing, macro intent recognition, micro-level where/what intent recognition, and full exact-match prediction. For response quality, we use a five-dimensional rubric covering

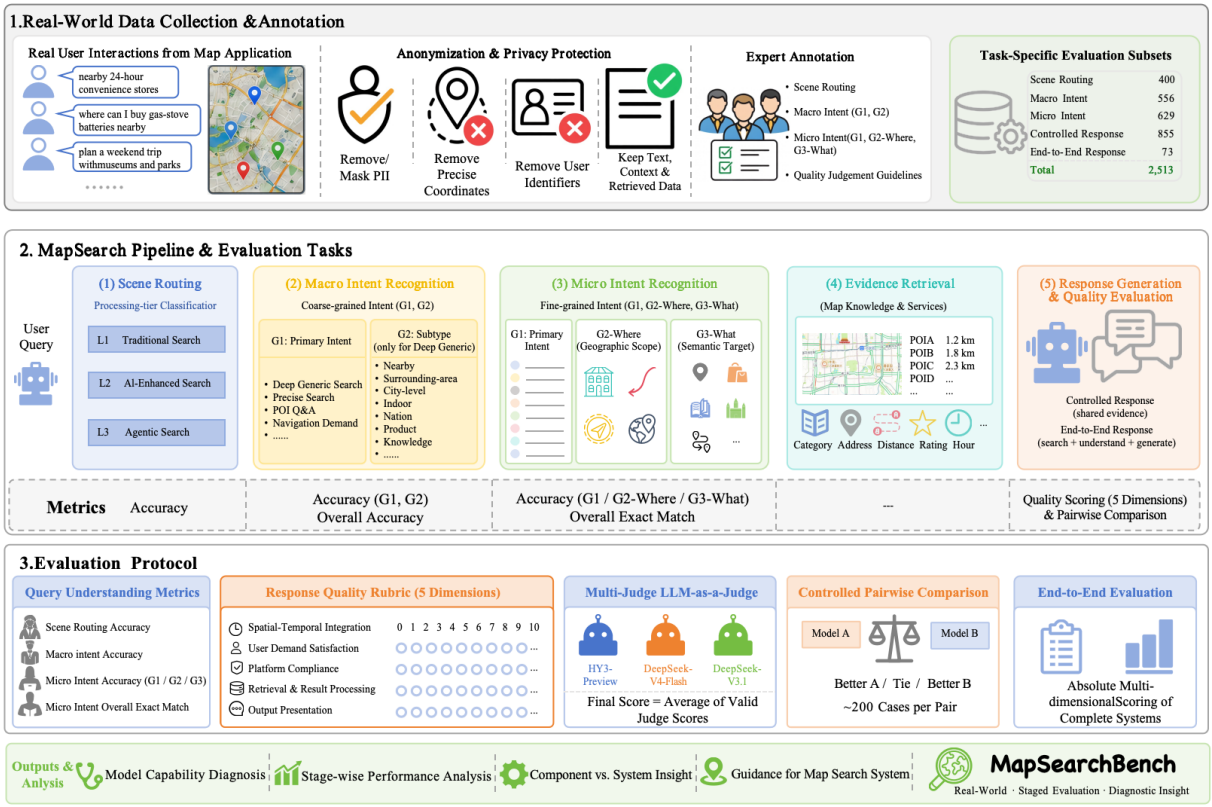

Figure 1: Overview of MapSearchBench. The benchmark evaluates intelligent map search from query understanding to response quality assessment. Query understanding is measured with routing and intent accuracy, while response quality is assessed through a five-dimensional rubric, multi-judge LLM scoring, controlled pairwise comparison, and end-to-end system evaluation.

spatial-temporal integration, user demand satisfaction, platform compliance, retrieval/result processing, and output presentation. To make response evaluation more reliable, we combine multi-judge LLM-as-a-Judge scoring with controlled pairwise comparison under shared retrieved evidence. We further include an end-to-end evaluation setting, where complete systems must perform search, interpretation, and response generation jointly. Thus, MapSearchBench can reveal not only which model performs better, but also why a model fails, whether due to incorrect routing, intent confusion, weak evidence grounding, or poor response structure.

2.2 Data Construction

We construct MapSearchBench from real-world user interactions collected from Tencent Map. The raw logs contain diverse map-search requests ranging from simple POI lookups to complex multi-constraint planning queries. To protect user privacy, we apply a strict anonymization pipeline before annotation. Personally identifiable information, precise coordinates, phone numbers, and user identifiers are removed or masked. Only the textual query, coarse-grained contextual information required for task labeling, and retrieved map evidence used for response evaluation are retained. This design allows the benchmark to preserve the linguistic and spatial complexity of real user requests while satisfying privacy and compliance requirements.

Because real map-search traffic is highly imbalanced, directly sampling from logs would over-represent frequent simple queries such as nearby POI lookup and under-represent long-tail cases involving ambiguous intent, multiple constraints, or cross-region reasoning. We therefore adopt a complexity-aware sampling strategy. Queries are first grouped according to preliminary scene types and intent patterns, and then sampled to ensure coverage of different difficulty levels, including traditional keyword retrieval, AI-enhanced semantic search, and agentic multi-step search. Table 1 summarizes the resulting task-specific evaluation subsets.

The annotation process is conducted by ten domain experts, each with experience in map-search product logic and geospatial query understanding. Each query is labeled under the task taxonomy used

Subset	Target	Size
Scene Routing	Processing-tier classification	400
Macro Intent	Coarse intent classification	556
Micro Intent	Fine-grained intent labeling	629
Controlled Response	Shared-evidence generation	855
End-to-End Response	Search-understand-generate	73

Table 1: Statistics of benchmark subsets.

by MapSearchBench, including its scene-routing label, macro intent, and micro-level spatial-semantic attributes. To improve annotation reliability, annotators are required to provide short rationales that justify their labels, especially for ambiguous or multi-intent queries. Samples with inconsistent annotations are reviewed through consensus discussion, and only labels that pass the final consistency check are included in the benchmark. This expert-guided process ensures that the dataset reflects not only surface-level query categories but also the underlying spatial and semantic reasoning required by real map-search systems. Appendix 5.4 provides representative task examples, and Appendix 5.7 describes the planned release format.

2.3 Task Formulation

Table 2 summarizes the label spaces and representative examples for the query-understanding tasks. We organize MapSearchBench into four evaluation tasks that correspond to the major decision points in an intelligent map-search pipeline. The first three tasks evaluate query understanding, including processing-tier selection and intent parsing, while the fourth task evaluates response generation quality. This formulation allows us to separately examine whether a model fails because it routes a query incorrectly, misunderstands the user's spatial intent, or generates an inadequate final response.

Scene Routing. Scene Routing evaluates whether a model can determine the appropriate processing tier for a user query before expensive retrieval or generation is invoked. This task is necessary because real map-search queries vary substantially in complexity. Some can be answered by direct keyword retrieval, while others require semantic expansion or multi-step agentic reasoning. We formulate Scene Routing as a three-way classification task. Given a query q, the model predicts one of three labels. L_1 for traditional keyword-based retrieval, L_2 for AI-enhanced semantic search, and L_3 for agentic search that requires multi-step reasoning or tool use. This task measures whether a model can make cost-aware routing decisions, which is crucial for balancing latency and intelligence in practical map-search systems. For example, "Starbucks nearby" is typically routed to L_1, while "find a quiet place nearby with few people and preferably a mountain-top view" requires a higher-level semantic or agentic search path.

Macro Intent Recognition. Macro Intent Recognition evaluates whether a model can identify the high-level map-search intent of a query. We define G_1 as the primary intent, including Deep Generic Search, Precise Search, POI Question Answering, and Navigation Demand. Since Deep Generic Search covers a broad range of exploratory map-search behaviors, we further annotate a conditional G_2 label for this category, such as nearby search, surrounding-area search, city-level search, indoor search, national search, product search, and knowledge-oriented search. This macro taxonomy provides an operational view of how a query should be handled by downstream map-search modules.

Micro Intent Recognition. Although Macro Intent Recognition provides executable intent categories, its G_2 labels mix two different factors. The spatial scope of the search and the semantic target of the request. For example, a query such as "where buy gas-stove batteries nearby" is both a nearby search and a product search. To avoid forcing such queries into a single mixed label, we introduce Micro Intent Recognition, which decomposes Deep Generic Search into two orthogonal dimensions. G_2-Where describes the geographic scope, such as basic local search, surrounding-area search, indoor search, city-level search, along-route search, or national search. G_3-What describes the semantic target, such as place search, product search, knowledge search, or route search. Given a query q, the model predicts (g_1, g_2, g_3), separating where the user wants to search from what the user wants to find.

Response Quality Evaluation. While the previous tasks evaluate query understanding, map search ultimately requires a model to produce useful responses for users. We therefore include Response Quality Evaluation to assess the final output quality. We consider two complementary settings. In the controlled response setting, all models are given the same user query and retrieved map evidence, and are asked to generate an answer. This setting isolates response generation ability under shared evidence. In the end-to-end setting, models receive

Task	Label / Dimension	Label Space	Example
Scene Routing	Processing tier	Traditional Search; AI-Enhanced Search; Agentic Search	"Starbucks nearby" → Traditional; "plan a weekend trip" → Agentic
Macro Intent	G_1: Primary intent	Deep Generic Search; Precise Search; POI Q&A; Navigation Demand	"Where is Baicheng Beach?" → Precise Search
	G_2: Conditional subtype for Deep Generic Search	Nearby; Surrounding-area; City-level; Indoor; National; Product; Knowledge	"nearby 24-hour convenience stores" → Nearby; "where to buy gas-stove batteries" → Product
Micro Intent	G_1: Primary intent	Deep Generic Search; Precise Search; POI Q&A; Navigation Demand; Map Demand; Dialogue; Travel Guide	"show real-time buses nearby" → Map Demand
	G_2: Where	Basic local; Surrounding-area; City-level; Indoor; Along-route; National	"charging stations along the Shanghai-Chongqing Expressway" → Along-route
	G_3: What	Place; Product; Knowledge; Route	"where can I buy watch straps nearby" → Product

Table 2: Task taxonomy in MapSearchBench. Macro Intent provides an operational coarse-grained taxonomy, while Micro Intent disentangles the geographic scope (G_2-Where) from the semantic target (G_3-What).

only the user query and must complete the search-understanding-generation process by themselves. This setting evaluates the full system-level capability using absolute multi-dimensional scoring.

2.4 Evaluation Protocol

MapSearchBench uses task-specific evaluation protocols rather than a single aggregate metric. Query understanding tasks are evaluated with objective accuracy-based metrics, while response generation tasks are evaluated with multi-dimensional quality scoring and pairwise comparison. This design reflects the heterogeneous nature of map search,in that routing and intent parsing require exact label prediction, whereas response generation requires assessing usefulness, grounding, compliance, and presentation quality.

Accuracy-based metrics for query understanding. For Scene Routing and Intent Recognition, we report accuracy as the primary metric. For Scene Routing, a prediction is correct if the model assigns the query to the same processing tier as the expert label. For Macro Intent Recognition, we compute overall accuracy over coarse-grained intent categories. For Micro Intent Recognition, we report accuracy at multiple granularities, including the primary intent dimension G_1, the geographic-scope dimension G_2, the semantic-attribute dimension G_3, and the overall exact-match accuracy across the full tuple. These metrics allow us to distinguish whether a model fails at high-level intent recognition or at more fine-grained spatial-semantic decomposition.

Multi-dimensional response scoring. For response generation, exact-match metrics are unsuitable because multiple valid responses may satisfy the same user request. We therefore score each response with a five-dimensional rubric: spatial-temporal integration, user demand satisfaction, platform compliance, retrieval/result processing, and output presentation. Each dimension is rated on a 10-point scale, and the absolute score is the average of valid dimension scores. Controlled response generation is scored by three LLM judges, while end-to-end response quality is scored by two judges using the same rubric. Table 3 summarizes the rubric, and Appendix 5.2 provides simplified prompt templates for response generation and scoring.

Shared retrieved evidence. In the controlled response setting, all models receive the same user query and the same retrieved map evidence. Each evidence set is produced by a fixed map-retrieval pipeline and contains structured POI fields such as name, category, address, distance when available, rating, business hours, tags, and short descriptive snippets. Models are not allowed to access additional search results during generation. This design isolates response generation from retrieval variance and tests how models filter, organize, and present the same evidence. We retain cases with sparse or partially irrelevant retrieval results, since such imperfections are common in real map search. Appendix 5.3 illustrates the shared-evidence format.

Multi-judge evaluation. To reduce the bias of a single evaluator, we adopt a multi-judge LLM-

Dimension	Description
Spatial-Temporal Integration	Whether the response correctly handles location, distance, direction, time, and other spatial-temporal constraints.
User Demand Satisfaction	Whether the response addresses the user's explicit and implicit needs with relevant and useful recommendations.
Platform Compliance	Whether the response follows map-service constraints, avoids unsafe or unsupported claims, and respects service boundaries.
Retrieval & Result Processing	Whether the response accurately uses retrieved evidence, filters irrelevant results, and synthesizes results logically.
Output Presentation	Whether the response is clear, concise, well-structured, and easy for users to follow.

Table 3: Five-dimensional rubric for response quality.

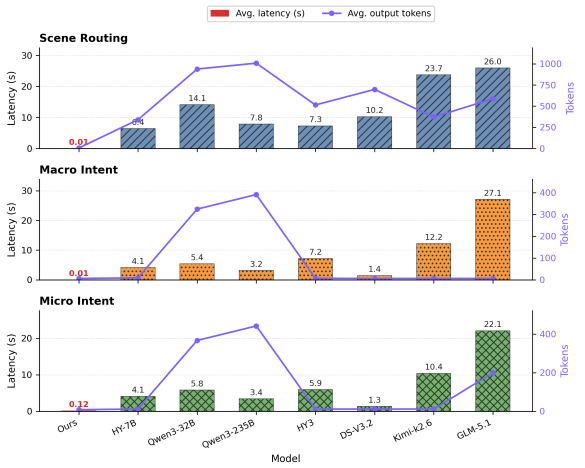

Figure 2: Latency and output-token distribution on the query understanding tasks for representative models.

as-a-Judge protocol. In the controlled response setting, responses are scored by three independent judges including Hunyuan-3-Preview, DeepSeek-V4-Flash, and DeepSeek-V3.1. We report both individual-judge scores and cross-judge averages. This protocol enables us to examine whether model rankings are stable across different evaluators and prevents conclusions from relying on a single judge's scoring preference. Appendices 5.5 and 5.6 report judge-agreement statistics and discuss judge overlap.

Pairwise comparison. In addition to absolute scoring, we conduct blind pairwise comparison in the controlled response setting to obtain relative preferences between model outputs. For each query, two model responses are randomly ordered and presented as anonymous answers to an LLM judge, which selects the better response or declares a tie. Pairwise comparison complements absolute scoring by reducing sensitivity to judge-specific score calibration and by revealing head-to-head capability differences that may be obscured by aggregate scores.

3 Experiments

3.1 Experimental Setup

Models and systems. We evaluate 17 model families or response components for query understanding and controlled response generation, covering Qwen, DeepSeek, GLM, Kimi, MiniMax, GPT-OSS, MiMo, Hunyuan, and OURS. Throughout the experiments, OURS denotes the corresponding component of our map-search system including the domain-adapted query-understanding model for classification tasks, our response generation component for controlled response evaluation, and the complete map-search pipeline for end-to-end evaluation. The query-understanding component is initialized from Hunyuan-7B and trained with full-parameter SFT on 2,899, 3,005, and 3,012 de-duplicated examples for scene routing, macro intent recognition, and micro intent recognition, respectively. Appendix 5.1 provides additional training details. For end-to-end evaluation, we compare four complete systems, including OURS, DeepSeek-V3.1, DeepSeek-V4-Flash, and Hunyuan-3-Preview.

Evaluation. We evaluate Scene Routing, Macro Intent Recognition, and Micro Intent Recognition on 400, 556, and 629 labeled instances, respectively. Controlled response generation uses 855 cases with shared retrieved map evidence, while end-to-end response evaluation uses 73 complex queries where each system must search, interpret, and generate jointly. Query-understanding tasks are evaluated with accuracy and exact-match metrics. Response quality is scored with the five-dimensional rubric in Table 3. Controlled response quality uses three independent LLM judges, and end-to-end response quality uses two judges. Additional controlled pairwise diagnostics are reported in Appendix 5.10.

3.2 Query Understanding Results

Overall performance. Table 4 shows that our domain-adapted model substantially outperforms general-purpose LLMs on all query understanding tasks. On Scene Routing, our model achieves an accuracy of 0.8600, more than doubling the strongest baseline, MiniMax-M2.7 (0.4375). On Macro Intent Recognition, our model reaches

Table 4: Performance comparison across query understanding and controlled response quality. **ALL** denotes overall accuracy. ST, UD, PC, RP, and OP denote spatial-temporal integration, user demand satisfaction, platform compliance, retrieval/result processing, and output presentation, respectively. For OURS, query-understanding columns report the domain-adapted understanding model, while controlled-response columns report our response generation component. The **best** results are highlighted in bold, and the second-best results are underlined.

	Scene	Macro Intent			Micro Intent				Controlled Response					
Model	**Acc.**	**ALL**	**G1**	**G2**	**ALL**	**G1**	**G2**	**G3**	**Avg.**	**ST**	**UD**	**PC**	**RP**	**OP**
OURS	**0.8600**	**0.8417**	**0.9874**	**0.7507**	**0.8203**	**0.9173**	0.8405	0.8348	**8.69**	8.56	8.65	**7.53**	**9.18**	**9.54**
MiniMax-M2.7	0.4375	0.5879	0.9561	0.5336	0.4642	0.7663	0.9288	0.8291	6.36	6.14	5.86	6.02	6.34	7.38
HY3-Preview	0.4325	0.7751	0.9518	0.6644	0.4706	0.7377	0.9088	0.8604	7.43	7.45	6.97	6.82	7.77	8.33
Qwen3-32B	0.4200	0.6385	0.9712	0.6103	0.4738	0.7758	0.9145	0.8718	7.00	6.99	6.39	7.19	6.71	7.82
Qwen3-235B-A22B	0.4125	0.5090	0.9694	0.4986	0.4738	0.7727	0.9316	0.8575	7.33	7.40	6.86	7.05	7.34	8.01
GLM-5.1	0.4100	0.7058	0.9774	0.6021	0.4817	0.7742	0.9345	0.8632	6.46	6.14	6.05	5.35	6.43	7.77
MiMo-V2-Flash	0.4075	0.4856	0.8687	0.5501	0.4308	0.6677	0.8604	0.8661	5.19	5.56	4.58	5.43	5.02	5.23
DS-V3.1	0.4050	0.7446	0.9676	0.6304	0.4610	0.6979	0.9150	**0.8886**	8.49	8.44	8.31	7.05	8.86	9.49
DS-V3.2	0.4025	0.7050	0.9550	0.5759	0.4690	0.7186	0.9117	0.8803	8.68	**8.67**	**8.74**	6.99	9.13	9.42
DS-V4-Flash	0.3925	0.7266	0.9353	0.6476	0.4610	0.7440	0.9060	0.8519	8.30	8.29	8.54	6.87	8.64	9.16
Qwen3-8B	0.3925	0.4299	0.8993	0.5330	0.4626	0.7647	0.9306	0.8497	6.42	6.54	5.81	6.73	6.16	6.94
DS-R1-70B	0.3875	0.5755	0.9406	0.5549	0.4499	0.7679	0.8889	0.8376	5.54	6.18	5.32	6.38	5.36	4.73
Kimi-K2.6	0.3800	0.6653	0.9743	0.5355	0.4913	0.7822	**0.9459**	0.8832	7.95	8.05	7.43	6.68	7.92	8.29
DS-R1-8B	0.3700	0.3525	0.7734	0.3754	0.2655	0.5787	0.5641	0.5470	4.34	5.02	3.69	5.71	3.67	3.79
Hunyuan-7B	0.3675	0.3381	0.7824	0.4441	0.4022	0.7281	0.8262	0.7607	6.10	5.97	5.41	5.62	6.12	7.03
GPT-OSS-120B	0.3625	0.6151	0.8723	0.5530	0.4897	0.7822	**0.9459**	0.8746	6.56	5.63	6.11	4.68	6.33	8.90
GPT-OSS-20B	0.2929	0.4227	0.7914	0.3506	0.0276	0.1721	0.3697	0.1758	2.86	3.03	2.20	2.95	2.18	3.33

0.8417 accuracy, outperforming the best general-purpose baseline, Hunyuan-3-Preview (0.7751), by 6.66 percentage points. The advantage is even more pronounced on Micro Intent Recognition, where our model obtains 0.8203 accuracy, compared with 0.4913 from the strongest baseline, Kimi-K2.6. These results indicate that query understanding in map search remains challenging for general-purpose LLMs, especially when fine-grained spatial-semantic labels are required.

Fine-grained spatial semantics are the main bottleneck. The performance gap is particularly large in Scene Routing and Micro Intent Recognition. Scene Routing requires models to decide whether a query should be handled by traditional retrieval, AI-enhanced semantic search, or agentic search. General-purpose LLMs often lack this system-level notion of map-search processing tiers, leading to unstable routing decisions. Micro Intent Recognition further requires separating what the user wants from where the user wants it, which involves implicit geographic scope, route-based constraints, indoor search, and national-scale reasoning. The large gap on this task suggests that general LLMs can recognize many surface-level intents but struggle to produce executable spatial-semantic signals for map systems.

Efficiency. Figure 2 reports the latency and output-token distribution for a representative subset of models on the three query understanding tasks; the full-model version is provided in Appendix 5.9. All systems are evaluated on H20 GPUs under the same environment. Our domain-adapted understanding model is designed for real-time map-search routing and intent parsing, where the output is constrained to short labels or label tuples rather than open-ended text. As a result, it achieves 0.01 seconds per query for both Scene Routing and Macro Intent Recognition, and 0.12 seconds for Micro Intent Recognition. In contrast, representative general-purpose models often require several seconds to tens of seconds, especially on Scene Routing and Micro Intent Recognition. This efficiency is important because query understanding serves as the first-stage gateway of the map-search pipeline and must respond before downstream retrieval or generation is invoked.

3.3 Controlled Response Quality

Overall ranking. We next evaluate response generation in the controlled setting, where all models are given the same user query and the same retrieved map evidence. This setting isolates the ability to organize retrieved evidence into a useful map-search answer. As shown in the Controlled Response Quality columns of Table 4, OURS obtains the highest cross-judge score of 8.69, followed closely by DeepSeek-V3.2 (8.68), DeepSeek-V3.1 (8.49), and DeepSeek-V4-Flash (8.30). This re-

System	Avg.	HY3 Judge	V4 Judge
Hunyuan-3-Preview	**7.88**	**8.17**	**7.60**
Ours	7.18	6.97	7.39
DeepSeek-V3.1	6.83	7.21	6.45
DeepSeek-V4-Flash	6.74	7.26	6.22

Table 5: End-to-end response quality of complete systems. Avg. is the average score over two LLM judges.

sult shows that our response component remains competitive when all systems are evaluated under the same retrieved evidence. At the same time, the small gap between OURS and strong general-purpose LLMs suggests that response generation depends not only on domain adaptation but also on general evidence synthesis and presentation ability.

Dimension-level analysis. The dimension-level scores further explain where high-scoring models gain their advantage. OURS achieves the best scores on retrieval/result processing (9.18), output presentation (9.54), and platform compliance (7.53), indicating that its response component is well aligned with map-specific evidence use and service constraints. DeepSeek-V3.2 obtains the best scores on spatial-temporal integration (8.67) and user demand satisfaction (8.74), showing strong general response synthesis ability. These results suggest that controlled response quality depends on both map-specific grounding and general answer organization. In contrast, platform compliance remains a relatively weaker dimension for most general-purpose models, suggesting that map-specific service constraints are not fully captured by general instruction-following ability.

Additional controlled pairwise diagnostics are provided in Appendix 5.10.

3.4 End-to-End Response Quality

Overall performance. We finally evaluate complete systems in an end-to-end setting, where each system receives only the user query and must perform search, interpretation, and response generation jointly. This setting is closer to real map-search usage than the controlled response setting, because the quality of the final answer depends not only on language generation but also on evidence acquisition and system-level decision making. As shown in Table 5, Hunyuan-3-Preview achieves the highest average score of 7.88. Our system ranks second with an average score of 7.18, outperforming DeepSeek-V3.1 (6.83) and DeepSeek-V4-Flash (6.74). These results show that a domain-adapted map-search pipeline remains competitive in realistic end-to-end scenarios. Additional discussion is provided in Appendix 5.8.

4 Related Work

LLM and search-agent benchmarks. Recent LLM benchmarks have moved from static knowledge tests toward realistic task execution, including general assistant evaluation (Mialon et al., 2023), tool-user interaction (Yao et al., 2024), software engineering (Jimenez et al., 2023), web browsing (Wei et al., 2025), and deep research workflows (Du et al., 2025; Sharma et al., 2025). These benchmarks motivate our diagnostic evaluation design, but they do not target the map-search pipeline, where routing, spatial intent parsing, evidence grounding, and user-facing response generation must be evaluated separately.

Geospatial and map-search evaluation. A growing body of work evaluates geographic knowledge, spatial reasoning, and urban understanding in foundation models (Mai et al., 2021; Roberts et al., 2023; Bhandari et al., 2023; Gurnee and Tegmark, 2023; Feng et al., 2025b; Zhou et al., 2025; Pyo et al., 2025). Closer to map services, MapEval, MapBench, LocalEval, and LocalSearchBench study map-based reasoning, geolocalization, local lifestyle services, and agentic local search (Dihan et al., 2024; Ji et al., 2026; Lan et al., 2025; He et al., 2025). MapSearchBench differs by using real map-search interactions to jointly evaluate processing-tier routing, macro and micro intent recognition, controlled response generation under shared retrieved evidence, and end-to-end system behavior.

Adapting LLMs to geospatial and search tasks. Prior work enhances LLMs with geospatial knowledge through domain pretraining, knowledge injection, or synthetic spatial data (Manvi et al., 2023; Deng et al., 2024; Balsebre et al., 2024; Feng et al., 2025a). Another line of work augments LLMs with retrieval or tool use instead of relying on memorized geographic knowledge (Yu et al., 2025; Jin et al., 2025; Team et al., 2025; Hu et al., 2025), following broader advances in reasoning-oriented LLMs (Guo et al., 2025). Our work is complementary: rather than proposing only a new model or agent, we provide a benchmark and analysis framework for measuring which stage of intelligent map search benefits from domain adaptation and which stage still relies on general evidence synthesis.

Limitations

MapSearchBench has several limitations. First, it is constructed from anonymized interactions of a single commercial map application, and may therefore reflect platform-specific user behavior and regional usage patterns. Second, the current benchmark mainly reflects the language and map-search habits observed in the collected data; extending it to multilingual and cross-regional settings is important future work. Third, although we use multiple independent LLM judges and controlled pairwise comparison, LLM-as-a-Judge may still introduce preferences for certain writing styles or response formats. Larger-scale human evaluation would further strengthen the reliability of response-quality assessment. Finally, our end-to-end evaluation contains 73 complex queries due to the cost of running complete search-understanding-generation systems. Expanding this subset and including more system architectures would provide a more comprehensive view of real-world map-search agents.

References

Pasquale Balsebre, Weiming Huang, and Gao Cong. 2024. Lamp: A language model on the map. *arXiv preprint arXiv:2403.09059*.

Prabin Bhandari, Antonios Anastasopoulos, and Dieter Pfoser. 2023. Are large language models geospatially knowledgeable? In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pages 1–4.

Cheng Deng, Tianhang Zhang, Zhongmou He, Qiyuan Chen, Yuanyuan Shi, Yi Xu, Luoyi Fu, Weinan Zhang, Xinbing Wang, Chenghu Zhou, and 1 others. 2024. K2: A foundation language model for geoscience knowledge understanding and utilization. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 161–170.

Mahir Labib Dihan, Md Tanvir Hassan, Md Tanvir Parvez, Md Hasebul Hasan, Md Almash Alam, Muhammad Aamir Cheema, Mohammed Eunus Ali, and Md Rizwan Parvez. 2024. Mapeval: A map-based evaluation of geo-spatial reasoning in foundation models. *arXiv preprint arXiv:2501.00316*.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2025. Deepresearch bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*.

Jie Feng, Tianhui Liu, Yuwei Du, Siqi Guo, Yuming Lin, and Yong Li. 2025a. 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.

Jie Feng, Jun Zhang, Tianhui Liu, Xin Zhang, Tianjian Ouyang, Junbo Yan, Yuwei Du, Siqi Guo, and Yong Li. 2025b. Citybench: Evaluating the capabilities of large language models for urban tasks. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5413–5424.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

Hang He, Chuhuai Yue, Chengqi Dong, Mingxue Tian, Hao Chen, Zhenfeng Liu, Jiajun Chai, Xiaohan Wang, Yufei Zhang, Qun Liao, and 1 others. 2025. Localsearchbench: Benchmarking agentic search in real-world local life services. *arXiv preprint arXiv:2512.07436*.

Yulan Hu, Xiangwen Zhang, Sheng Ouyang, Hao Yi, Lu Xu, Qinglin Lang, Lide Tan, Xiang Cheng, Tianchen Ye, Zhicong Li, and 1 others. 2025. Amap agentic planning technical report. *arXiv preprint arXiv:2512.24957*.

Yuxiang Ji, Yong Wang, Ziyu Ma, Yiming Hu, Hailang Huang, Xuecai Hu, Guanhua Chen, Liaoni Wu, and Xiangxiang Chu. 2026. Thinking with map: Reinforced parallel map-augmented agent for geolocalization. *arXiv preprint arXiv:2601.05432*.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2023. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Xiaochong Lan, Jie Feng, Jiahuan Lei, Xinlei Shi, and Yong Li. 2025. Benchmarking and advancing large language models for local life services. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 4566–4577.

Gengchen Mai, Krzysztof Janowicz, Rui Zhu, Ling Cai, and Ni Lao. 2021. Geographic question answering: challenges, uniqueness, classification, and future directions. *AGILE: GIScience series*, 2:8.

Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. 2023. Geollm: Extracting geospatial knowledge from large language models. *arXiv preprint arXiv:2310.06213*.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*.

Jiyoon Pyo, Yuankun Jiao, Dongwon Jung, Zekun Li, Leeje Jang, Sofia Kirsanova, Jina Kim, Yijun Lin, Qin Liu, Junyi Xie, and 1 others. 2025. Frieda: Benchmarking multi-step cartographic reasoning in vision-language models. *arXiv preprint arXiv:2512.08016*.

Jonathan Roberts, Timo Lüddecke, Sowmen Das, Kai Han, and Samuel Albanie. 2023. Gpt4geo: How a language model sees the world's geography. *arXiv preprint arXiv:2306.00020*.

Manasi Sharma, Chen Bo Calvin Zhang, Chaithanya Bandi, Clinton Wang, Ankit Aich, Huy Nghiem, Tahseen Rabbani, Ye Htet, Brian Jang, Sumana Basu, and 1 others. 2025. Researchrubrics: A benchmark of prompts and rubrics for evaluating deep research agents. *arXiv preprint arXiv:2511.07685*.

Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, and 1 others. 2025. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*.

Dazhou Yu, Riyang Bao, Gengchen Mai, and Liang Zhao. 2025. Spatial-rag: Spatial retrieval augmented generation for real-world spatial reasoning questions. *arXiv preprint arXiv:2502.18470*.

Baichuan Zhou, Haote Yang, Dairong Chen, Junyan Ye, Tianyi Bai, Jinhua Yu, Songyang Zhang, Dahua Lin, Conghui He, and Weijia Li. 2025. Urbench: A comprehensive benchmark for evaluating large multimodal models in multi-view urban scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10707–10715.

Appendix

5 Additional Details for MapSearchBench

This appendix provides additional details for data examples, model adaptation, prompt templates, shared retrieved evidence, judge reliability, and the planned release of MapSearchBench.

5.1 Domain-adapted Model Training

We train the domain-adapted query-understanding component of OURS using full-parameter supervised fine-tuning based on Hunyuan-7B. The training data is derived from real-world online queries, LLM-assisted annotation, and manual correction by domain experts. We remove duplicates between the training data and all evaluation subsets. We train three separate models for the three query-understanding tasks instead of using a single multi-task model, because each task requires a different output schema and decision granularity. Table 6 summarizes the training data sizes.

Task	Training Size	Output Format
Scene Routing	2,899	Single routing label
Macro Int. Rec	3,005	Intent label or intent-subtype tuple
Micro Int. Rec	3,012	G_1, G_2-Where, G_3-What tuple

Table 6: Training data sizes for the domain-adapted query-understanding models.

The fine-tuning hyperparameters are kept similar across the three tasks. Table 7 reports the configuration used for scene routing; macro and micro intent models use similar settings. Scene routing training converges at around 256 steps and takes approximately 3 hours and 40 minutes. All training runs are conducted on H20 GPUs.

Hyperparameter	Value
Base model	Hunyuan-7B
Fine-tuning method	Full-parameter SFT
Sequence length	2048
Training epochs	3
Global batch size	32
Micro batch size	1
Learning rate	1×10^{-4}
Minimum learning rate	1×10^{-5}
Evaluation interval	50 steps
Checkpoint interval	50 steps
Number of checkpoints	2
Warmup token ratio	0
LR decay token ratio	1.0
Model-parallel size	8
Training time	$\sim$3h40m

Table 7: Training configuration for the scene-routing model. Macro and micro intent models follow similar settings.

At inference time, the query-understanding models are deployed on a single H20 GPU. Since the outputs are constrained to short structured labels,

decoding is much faster than open-ended response generation.

5.2 Simplified Prompt Templates

The full prompts used in our experiments are long because they contain taxonomy definitions, formatting constraints, and examples. We will release the full prompts together with the benchmark code after the institutional open-source review process. Here we provide simplified versions to illustrate the input-output format.

Scene Routing. The model is asked to assign each query to one processing tier.

> Given a user query, decide which processing tier should handle it. Labels: Traditional Search, AI-Enhanced Search, Agentic Search. Output only the final label.

Macro Intent Recognition. The model predicts the primary map-search intent and, for deep generic search, a conditional subtype.

> Given a user query, identify its primary map-search intent. Primary intents include Deep Generic Search, Precise Search, POI Question Answering, and Navigation Demand. If the primary intent is Deep Generic Search, also output the subtype. Output format: `G1` or `G1, G2`.

Micro Intent Recognition. The model predicts the primary intent, geographic scope, and semantic target.

> Given a user query, output three labels: G_1: primary intent; G_2-Where: geographic scope; G_3-What: semantic target. Output format: `G1, G2-Where, G3-What`.

Controlled Response Generation. The model receives a user query and shared retrieved map evidence.

> Given a user query and retrieved map candidates, generate a concise map-search answer. Use only the provided evidence if the query has explicit distance or location constraints. Do not fabricate distance, address, business hours, or other POI attributes. Organize the answer in a user-friendly format.

Response Quality Scoring. The judge evaluates a model response using the five-dimensional rubric.

> Given a user query, retrieved evidence, and model response, score the response from 0 to 10 on: spatial-temporal integration, user demand satisfaction, platform compliance, retrieval/result processing, and output presentation. Provide dimension scores and a brief rationale.

Controlled Pairwise Comparison. The judge compares two anonymized responses under the same query and evidence.

> Given one user query and two anonymous model responses, decide which response better satisfies the user's map-search need. Consider relevance, evidence grounding, spatial constraints, usefulness, and presentation. Output: A wins, B wins, or Tie.

5.3 Shared Retrieved Evidence

For controlled response generation, all models receive the same retrieved map evidence. Each evidence set is produced by the same map-retrieval pipeline and contains a fixed list of candidate POIs and structured metadata. This setting isolates response generation from retrieval variance. Table 8 illustrates the evidence format.

We retain cases with sparse, noisy, or partially irrelevant retrieved evidence because such imperfections are common in real map search. These cases test whether a model can avoid unsupported claims and organize imperfect evidence responsibly.

5.4 Representative Task Examples

Tables 9, 10, and 11 provide representative examples for the query-understanding tasks. We use translated or anonymized queries where necessary to reduce privacy risk.

5.5 Judge Agreement

We analyze agreement among the three LLM judges used for controlled response scoring. Table 12 reports both response-level and model-level correlations. Response-level correlations measure agreement on individual model responses, while model-level correlations measure agreement over the average scores of the evaluated models.

The response-level correlations show moderate to strong agreement, while the model-level correlations are high. This indicates that individual response scores may vary across judges, but the overall model rankings are stable under the multi-judge protocol.

5.6 Judge Overlap

Two judge families also appear among the evaluated systems. We therefore report individual judge scores, cross-judge averages, and the agreement statistics in Table 12. The high model-level rank correlations suggest that the main rankings are not driven by a single evaluator. Nevertheless, a fully

Field	Description / Example
User query	"nearby places suitable for couples to watch movies"
Context	Current time, city, and coarse user location. Precise coordinates are removed or masked in the released version.
Retrieved candidate fields	POI name, category, address, distance when available, rating, popularity, price, business hours, matched tags, display tags, and short descriptive snippets.
Candidate example	A cinema candidate may include category "cinema", rating 4.7, distance 1.7 km, business hours 10:00–22:00, and tags such as Dolby cinema, good viewing experience, and convenient location.
Expected response behavior	Select relevant cinema candidates, filter irrelevant venues, respect distance constraints, and generate concise recommendations grounded in the provided evidence.

Table 8: Example of shared retrieved evidence used in the controlled response setting. Sensitive identifiers, exact coordinates, and user-specific information are removed or masked.

Query	Scene Label
"Starbucks nearby"	Traditional Search
"Where is Baicheng Beach?"	Traditional Search
"nearby 24-hour convenience stores"	AI-Enhanced Search
"highly rated Thai restaurants nearby with private rooms"	AI-Enhanced Search
"plan a weekend trip with museums and parks"	Agentic Search
"find a quiet place nearby with few people and a mountain-top view"	Agentic Search

Table 9: Representative examples for Scene Routing.

disjoint judge panel and larger-scale human calibration remain important future directions.

5.7 Data Release Plan

MapSearchBench is built from anonymized real-world user interactions collected from a commercial map application. We have submitted the benchmark for institutional open-source review. Upon paper acceptance, we plan to release a public version of the benchmark, including representative anonymized examples for each task, task labels, prompt templates, scoring rubrics, and evaluation scripts. Sensitive location expressions, exact coordinates, user identifiers, phone numbers, and other personally identifiable information will be removed or masked before release.

For data fields that cannot be released due to privacy or platform constraints, we will provide a redacted version and an evaluation harness that allows external models to be evaluated under the same protocol. The appendix examples illustrate the released data format and annotation schema.

5.8 End-to-End Discussion

Controlled versus end-to-end evaluation. The end-to-end results complement the controlled response evaluation. In the controlled setting, all models receive the same retrieved evidence, so the evaluation mainly reflects response organization ability. In contrast, the end-to-end setting additionally tests whether a system can obtain useful evidence, interpret the query under map-specific constraints, and decide how to answer when evidence is incomplete or uncertain. The ranking difference between the two settings suggests that controlled response quality alone is insufficient for assessing practical map-search systems. A model that writes strong answers given evidence may still face difficulties when it must acquire and ground that evidence by itself.

Implications. The competitive end-to-end score of our system suggests that domain-adapted map-search pipelines can provide reliable user-facing responses when grounded map evidence is available. At the same time, the remaining gap to Hunyuan-3-Preview indicates that end-to-end map-search agents still benefit from broader evidence coverage and more proactive answer synthesis. This observation points to a practical direction for future systems: combining the factual grounding and service-boundary control of production map search with the richer evidence aggregation and response organization abilities of online LLMs.

5.9 Full Efficiency Results

5.10 Controlled Pairwise Diagnostics

This appendix reports the controlled pairwise results in Figure 4 and a representative grounding case in Table 13.

Pairwise validation. The absolute response scores suggest that some smaller or medium-sized

Query	G1	G2
"How much does bicycle rental cost at Xi'an City Wall?"	POI Q&A	–
"How far is it from Tsinghua University to the Palace Museum?"	Navigation Demand	–
"Where is Baicheng Beach in Xiamen?"	Precise Search	–
"nearby 24-hour convenience stores"	Deep Generic Search	Nearby
"where can I buy gas-stove batteries nearby"	Deep Generic Search	Product
"popular hotpot restaurants in Chengdu"	Deep Generic Search	City-level

Table 10: Representative examples for Macro Intent Recognition. G_2 is defined for Deep Generic Search.

Query	G1	G2-Where	G3-What
"nearby 24-hour convenience stores"	Deep Generic	Basic local	Place
"restaurants around the railway station"	Deep Generic	Surrounding-area	Place
"charging stations along the highway"	Deep Generic	Along-route	Place
"where can I buy watch straps nearby"	Deep Generic	Basic local	Product
"ancient towns worth visiting in Sichuan"	Deep Generic	City-level	Knowledge
"what are the Three Mountains and Five Sacred Peaks?"	Deep Generic	National	Knowledge

Table 11: Representative examples for Micro Intent Recognition. G_2-Where captures spatial scope, while G_3-What captures semantic target.

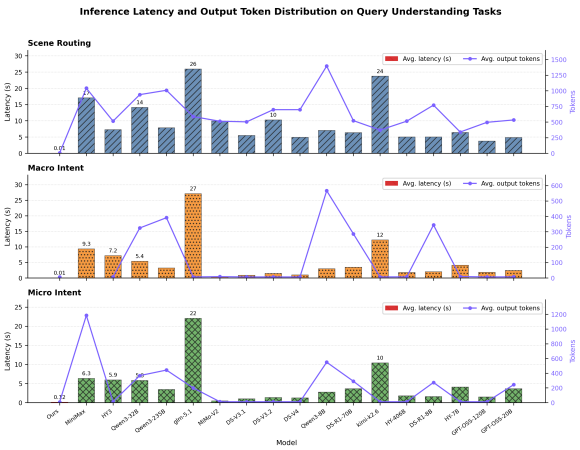

Figure 3: Inference latency and output-token distribution on query understanding tasks for all evaluated models.

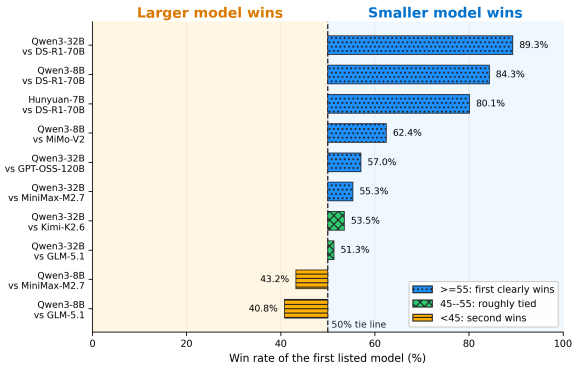

Figure 4: Pairwise win rates for selected small-over-large model comparisons under shared retrieved evidence.

models can outperform substantially larger models in the controlled setting. To examine whether this pattern is merely an artifact of independent score calibration, we conduct controlled pairwise comparisons on selected small-over-large model pairs. For each comparison, two model responses to the same query and the same retrieved evidence are anonymized, randomly ordered, and judged side by side. As shown in Figure 4, the small-over-large pattern remains in several pairs: Qwen3-8B wins against DeepSeek-R1-Distill-Llama-70B in 84.3% of comparisons, and Hunyuan-7B wins in 80.1% against the same baseline. Qwen3-32B also remains competitive against larger or stronger systems, achieving 57.0% against GPT-OSS-120B and 53.5% against Kimi-K2.6. These results suggest that the pattern is not solely caused by independent scoring, but also appears when judges directly compare two answers under identical evidence.

Qualitative analysis. We manually inspect pairwise cases where a smaller model is preferred over a larger model. The main reason is not that smaller models generate more detailed responses, but that they more consistently respect the retrieved evidence and query constraints. In several cases, larger models produce fluent answers but introduce unsupported venues, ignore distance or category constraints, or fail to flag that the retrieved results do not directly match the user intent. Smaller models can therefore win when they are more conservative and better grounded in the provided evidence.

Table 13 shows a representative case where Qwen3-32B is preferred over GPT-OSS-120B. The

Judge Pair	Response-level			Model-level		
	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
HY3 vs. V4	0.753	0.729	0.554	0.982	0.970	0.908
HY3 vs. DS-V3.1	0.714	0.681	0.515	0.908	0.906	0.775
V4 vs. DS-V3.1	0.681	0.644	0.481	0.933	0.897	0.800

Table 12: Agreement among LLM judges for controlled response evaluation.

Item	Content
Pair	Qwen3-32B vs. GPT-OSS-120B
User query	"Where can I buy music discs?"
Retrieved evidence	The retrieved map results do not contain a direct CD, DVD, or audio-disc store. The available results mainly include a music tutoring center, a piano and instrument store, a music bar, a music training center, and a music-themed restaurant.
Qwen3-32B response	The model first states that the current map results do not include a store that directly sells music discs. It then recommends related alternatives from the retrieved evidence, such as instrument stores and music-themed venues, while making clear that these places are alternatives rather than direct matches.
GPT-OSS-120B response	The model presents several plausible-looking audio or CD stores, such as Huaxing Audio-Video Mall, Chikan Xinhua Bookstore, Starlight Audio-Video Store, and Ledong Audio Specialty Store. These venues are not included in the retrieved evidence.
Pairwise judgment	Both judges select Qwen3-32B as the better answer. The main reason is that Qwen3-32B remains grounded in the provided evidence, whereas GPT-OSS-120B introduces unsupported venues.

Table 13: A representative controlled pairwise case where Qwen3-32B is preferred over GPT-OSS-120B because it better respects the retrieved evidence.

user asks for places to buy music discs. The retrieved map evidence contains music-related venues, but does not contain a direct CD/DVD or audio-disc store. Qwen3-32B explicitly states this evidence gap and recommends related alternatives based on the retrieved results. In contrast, GPT-OSS-120B generates several plausible-looking audio stores that are not present in the retrieved evidence. Both judges prefer Qwen3-32B because it is more faithful to the evidence and avoids unsupported venue hallucination.