

LocalSearchBench: Benchmarking Agentic Search in Real-World Local Life Services

Hang He^{1,2}, Chuhuai Yue^{1,3}, Chengqi Dong^{1,4}, Mingxue Tian^{1,5}, Hao Chen^{1,6}, Zhenfeng Liu¹, Jiajun Chai^{1*}, Xiaohan Wang¹, Yufei Zhang¹, Qun Liao¹, Guojun Yin^{1†}, Wei Lin¹, Chengcheng Wan^{2,7†}, Haiying Sun², Ting Su^{2†}

¹Meituan, ²East China Normal University, ³Beijing Institute of Technology

⁴University of Science and Technology of China, ⁵Shanghai Jiaotong University,

⁶North China University of Technology, ⁷Shanghai Innovation Institute

hang.he@stu.ecnu.edu.cn, {chajiajun, yinguojun02}@meituan.com, {ccwan, tsu}@sei.ecnu.edu.cn

Abstract

Recent advances in large reasoning models (*LRMs*) have enabled *agentic search* systems to perform complex multi-step reasoning across multiple sources. However, most studies focus on general information retrieval and rarely explores vertical domains with unique challenges. In this work, we focus on local life services and introduce LOCALSEARCHBENCH, which encompass diverse and complex business scenarios. Real-world queries in this domain are often ambiguous and require multi-hop reasoning across merchants and products, remaining challenging and not fully addressed. As the first comprehensive benchmark for *agentic search* in local life services, LOCALSEARCHBENCH comprises a database of over 1.3M merchant entries across 6 service categories and 9 major cities, and 900 multi-hop QA tasks from real user queries that require multi-step reasoning. We also developed LOCALPLAYGROUND, a unified environment integrating multiple tools for *LRMs* interaction. Experiments show that even state-of-the-art *LRMs* struggle on LOCALSEARCHBENCH: the best model (DeepSeek-V3.2) achieves only 35.60% correctness, and most models have issues with completeness (average 60.32%) and faithfulness (average 30.72%). This highlights the need for specialized benchmarks and domain-specific agent training in local life services. **Code, Benchmark, and Leaderboard** are available at localsearchbench.github.io.

CCS Concepts

- Computing methodologies → Natural language processing; Artificial Intelligence.

Keywords

Agentic Search, Large Reasoning Model, Local life services

1 Introduction

Recent advances in large reasoning models (*LRMs*) have significantly enhanced the capabilities of AI agents, enabling them to perform complex reasoning and planning tasks [1–7]. This progress has driven the development of *agentic search* systems, which can decompose complex queries, execute multi-step reasoning, and dynamically orchestrate multiple information sources to achieve autonomous cross-domain reasoning [8–16]. These systems also demonstrate the potential to expand from general domains to vertical application scenarios such as e-commerce [17, 18]. However,

*Project leader.

†Corresponding author.

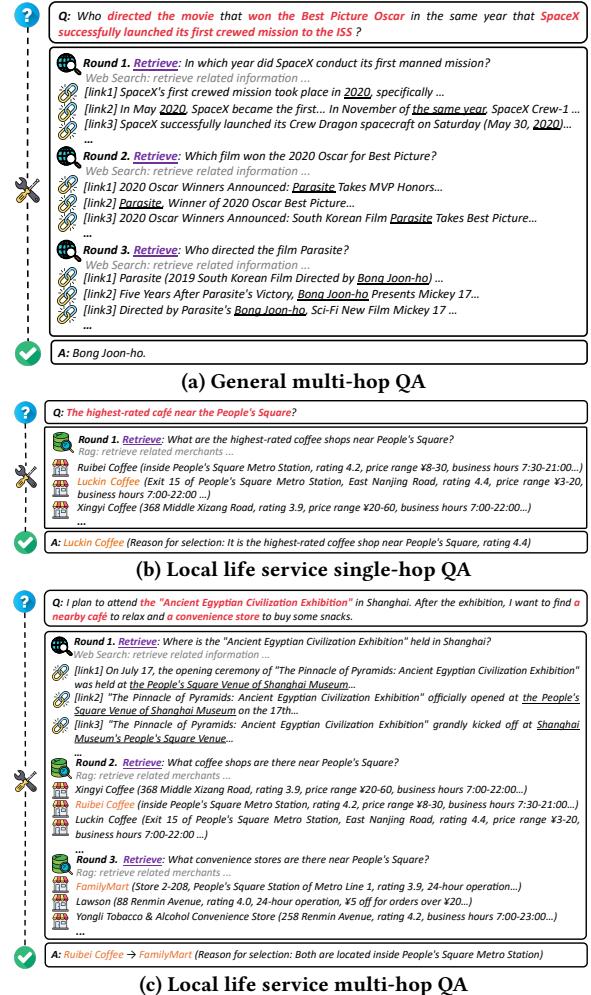


Figure 1: Illustration of different types of QA tasks

there has been little work applying *agentic search* to the local life service domain, and the lack of suitable evaluation benchmarks further constrains new research in this area.

Local life services are online-offline integrated services connecting local merchants with nearby users, including dining, lifestyle, shopping, accommodation, travel, healthcare and more to meet residents' daily needs [19, 20]. Leveraging location-based digital platforms, these services form complex networks of businesses,

user preferences, and location-dependent service offerings, creating a vast market [21–23]. As shown in Figure 1(c), real-world user queries often involve multi-constraint merchant recommendation, spatiotemporal service chain planning, and event-driven service bundling, which require complex multi-step processes to deliver the final response.

Lan et al. [19] target at this domain and propose the first systematic framework, but is limited to single-hop queries and cannot fully reflect the challenges of real business scenarios. Moreover, existing *agentic search* benchmarks (Table 1) focus on general domains and miss critical local-life challenges such as geographic constraints and multi-service integration. Therefore, there is an urgent need for a multi-hop benchmark tailored to local life scenarios, accurately reflecting the complexity of real-world search tasks and fully challenging existing *agentic search* methods.

To bridge these gaps, we first introduce **LOCALSEARCHBENCH**, a comprehensive benchmark comprising over 1,300,000 data entries across 6 primary service categories, spanning 9 major cities in China. These entries are transformed from raw data into high-quality augmented and anonymized database. Based on these data, we propose a high-quality instruction synthesis method to design 900 multi-hop QA tasks across two intelligence levels. These multi-hop QA tasks are constructed through a multi-agent data generation process, requiring agents to perform complex reasoning across multiple information sources and reasoning steps.

Building upon LOCALSEARCHBENCH, we further develop **LOCALPLAYGROUND**, an *agentic search* framework to evaluate LLM agents for local life service. Given the absence of appropriate benchmarks and the critical need for comprehensive evaluation systems, our approach addresses these challenges through comprehensive dataset construction, systematic evaluation framework and targeted model evaluation. We build a comprehensive evaluation environment with LOCALRAG based on merchant database and web search tools to simulate real-world local life service search scenarios where models need to retrieve and synthesize information from multiple sources, and conduct evaluations of multiple *LRMs*.

In summary, our contributions are threefold.

- To the best of our knowledge, we are the first to build a comprehensive offline high-quality real-world *agentic search* benchmark for local life service.
- We propose a systematic *agentic search* framework LOCALPLAYGROUND and tool environment for evaluating and applying *LRMs* in local life services.
- We evaluate existing models and find that untrained models struggle to complete these tasks, while trained models perform poorly on domain-specific professional requirements.

2 Related Work

2.1 Agentic Search

Gou et al. [31] define *agentic search* as systems capable of integrating multiple auxiliary tools to autonomously perform interactive generation through search tools (e.g., web search APIs or external databases) for handling complex search tasks. Leveraging the powerful language understanding capabilities of LLMs, these search agents can proactively address ambiguous or underspecified queries

through clarification or reasonable assumptions [32]. They decompose the initial search task into manageable sub-queries, perform dynamic reasoning and planning based on accumulated contextual information, adjust search strategies in real time, and synthesize information from multiple sources [33].

Recent advances in the reasoning abilities of *LRMs* further foster the development of *LRM*-based agentic search systems. The most representative products include OpenAI Deep Research [34], Gemini Deep Research [35], and Tongyi Deep Research [36]. In addition to accessing real-time information through search APIs, these systems leverage various auxiliary tools to enable deeper and broader research on highly complex problems. In the open-source community, Li et al. [13] first proposed integrating the agentic search workflow into the o1-style reasoning process of *LRMs* to achieve autonomous knowledge supplementation. Li et al. [4], Jin et al. [37], and Song et al. [38] further enhanced the autonomous search capabilities of LLMs by applying reinforcement learning, enabling the models to generate queries and invoke external retrieval systems during reasoning.

These methods have made significant progress in general information retrieval and question answering. However, there has been little advancement in vertical domains, especially in areas related to local lifestyle services, due to the challenges associated with training domain-specific agents and the lack of high-quality evaluation benchmarks. In this paper, we propose LOCALSEARCHBENCH, a comprehensive benchmark specifically designed for evaluating *LRMs* in local life service scenarios.

2.2 Benchmarking Agentic capabilities

Existing benchmarks for agentic search systems in general domain can be broadly categorized into closed-ended and open-ended QA tasks. Closed-ended QA benchmarks include multi-hop reasoning datasets that require synthesizing information from multiple sources [24, 25, 39, 40], challenging QA benchmarks designed for long-horizon questions with long-tail knowledge [26, 41–43], and fact-checking tasks that evaluate claim verification abilities [44–47]. Open-ended QA benchmarks focus on deep information seeking tasks, targeting multi-perspective queries [29] and expert-level research tasks [27, 30, 48], with some incorporating multi-modal capabilities [49–51].

Despite covering various reasoning abilities, existing benchmarks differ from real user queries in local life services and lack assessment of multi-domain coordination. Users often face complex tasks requiring simultaneous service bookings and constraint satisfaction, which current datasets do not adequately evaluate.

However, although these benchmarks cover a range of reasoning capabilities, they have significantly different distribution from real user queries in local life service scenarios, lacking assessment of the multi-domain service coordination required in real-world applications. In reality, users often need to accomplish complex tasks that require coordinating multiple services, such as simultaneously booking restaurants, arranging transportation, and planning entertainment, all while satisfying geographic and temporal constraints. Existing specialized datasets are insufficient for evaluating challenging scenarios like event planning that require multiple

Table 1: Comparison of existing user interaction benchmarks across key evaluation dimensions. “✓” indicates fully addressed, “✗” indicates partially addressed, and “✗” indicates not addressed.

Benchmark	Reasoning Complexity			Domain Specificity		Task Complexity			Scale (samples)
	Multi-hop Reasoning	Tool Integration	Context Dependency	Domain Knowledge	Geographic Constraints	Multi-service Integration	Real-world Scenarios		
HotpotQA [24]	✓	✗	✓	✗	✗	✗	✗	✗	113K
MuSiQue [25]	✓	✗	✓	✗	✗	✗	✗	✗	25K
BrowseComp [26]	✓	✓	✗	✗	✗	✗	✗	✗	1K
DeepResearch [27]	✓	✓	✓	✗	✗	✗	✓	✓	500
DeepWideSearch [28]	✓	✓	✓	✗	✗	✗	✓	✓	220
ResearchY [29]	✓	✓	✓	✗	✗	✗	✓	✓	1.5K
ProxyQA [30]	✓	✗	✓	✗	✗	✗	✗	✗	1550
LOCALSEARCHBENCH (ours)	✓	✓	✓	✓	✓	✓	✓	✓	1.35M + 900

API calls across different service domains and real-time constraint satisfaction.

We propose LOCALSEARCHBENCH to assess multi-hop reasoning and coordinated retrieval for complex local life service tasks, and LOCALPLAYGROUND as evaluation frameworks that comprehensively measure performance in these specialized scenarios.

3 Construction of LOCALSEARCHBENCH

In this section, we introduce LOCALSEARCHBENCH, a novel benchmark for *agentic search* in local life services. LOCALSEARCHBENCH consists of two main components: Local Merchant Database (Section 3.1) and a multi-hop QA benchmark (Section 3.3).

3.1 Local Merchant Database Construction

In this subsection, we present the construction of the local merchant database for LOCALSEARCHBENCH, which serves as the foundation for multi-hop QA tasks. As shown in Figure 2, our database construction consists of three stages: *Merchant Seed Data Collection*, *Data Augmentation* and *Synthesis*, and *Data Validation*.

3.1.1 Merchant Seed Data Collection. To ensure LOCALSEARCHBENCH reflects real-world local life service complexities, we built our local merchant database based on raw data from a leading local life service platform M. We sampled over 1,600,000 real merchant records spanning from January to December 2025, covering diverse service categories and landmarks across 9 major Chinese cities, involving over 9,900,000 retrievals from raw data, with the generation of these seed data requiring 1,418 hours of computation time.

Multi-Scenario Coverage. Based on platform M’s core business verticals, we identified six primary service categories reflecting real-world user demand: **Dining (35%), Lifestyle (25%), Shopping (20%), Accommodation (10%), Healthcare (5%), and Tourism (5%)**, as shown in Figure 4(a). We curated a keyword pool of 927 unique terms across these categories to reflect authentic user search behaviors and regional preferences.

Multi-City Coverage. The collection spans 9 major Chinese cities with comprehensive coverage of benchmarks (Table 2). The location database includes 10,741 landmarks across diverse categories: commercial centers, transportation hubs, cultural landmarks, educational institutions, medical facilities, and residential areas. Figure 3 shows word clouds of landmark distributions for three of the nine cities.

Table 2: Geographic coverage across 9 major Chinese cities

City	Landmarks	Districts	Retrievals
Shanghai	1556	16	1,442,412
Beijing	1451	16	1,345,077
Guangzhou	1026	11	951,102
Shenzhen	1039	9	963,153
Hangzhou	1129	13	1,046,583
Chengdu	1073	20	994,671
Chongqing	1008	38	934,416
Wuhan	1040	13	964,080
Suzhou	1149	9	1,065,123
Total	10,741	145	9,956,907

Merchant Seed Data Retrieval Query Generation. We construct scenario-specific queries by randomly combining 1–2 keywords from the 927-term pool to simulate natural search patterns. Applied across 145 administrative divisions, this multi-city keyword strategy generates diverse query variations and comprehensive coverage of local search intents and regional preferences, forming the foundation for LOCALSEARCHBENCH’s merchant database synthesis.

3.1.2 Data Augmentation and Synthesis.

Data Augmentation. The original merchant seed data contains only 12 basic fields with many missing or unstandardized attributes, which is inadequate for *agentic search* tasks that require dense, high-coverage merchant profiles. We employ a *Data Augmentation Agent* to expand these 12-field records into 29-field profiles using LLMs across six dimensions (see Figure 2 and Figure 14). The agent enriches each merchant’s basic, locational, operational, service, contact, and promotional information through 17 newly added fields, as illustrated in Figure 2.

Privacy Rewriting. To comply with data protection regulations and enable safe public release, we employ a *Privacy Rewriting Agent* on the augmented data (see Figure 15). Real-world merchant data contains personally identifiable information and business-sensitive details that must be anonymized [52, 53]. The agent rewrites seven privacy-related fields to anonymize merchant identity, location, and contact details while keeping the remaining business information usable, as illustrated in Figure 2.

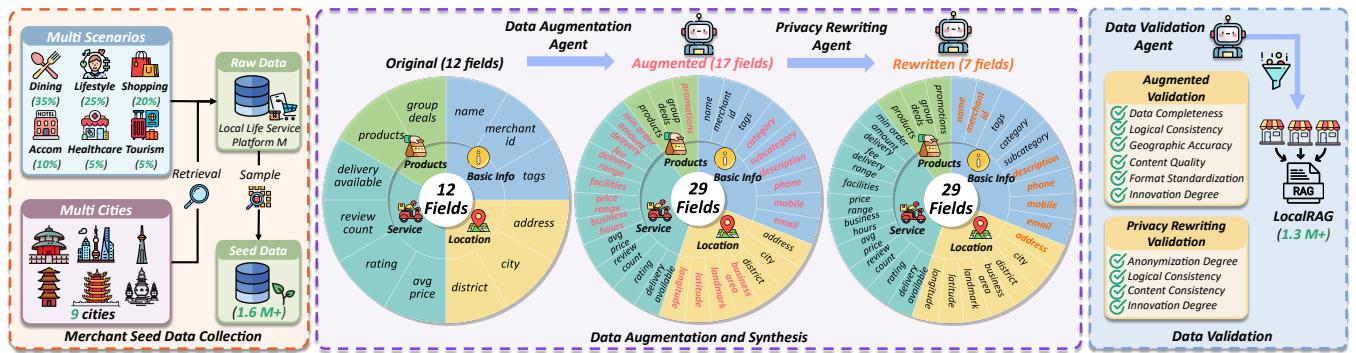


Figure 2: Workflow of Local Merchant Database Construction.

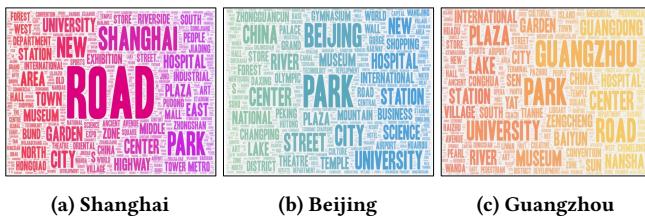


Figure 3: Word clouds showing the landmark locations across 3 major Chinese cities.

3.1.3 Data Validation. A *Validation Agent*, implemented as an LLM-as-a-Judge framework, conducts a comprehensive, two-stage quality assessment to ensure both data quality and privacy compliance in the generated database. Detailed instructions have been carefully designed for augmented data validation and privacy rewriting validation, as illustrated in Figure 14.

Augmented Data Validation. It evaluates augmented data quality against the original seed data along six weighted dimensions: Data Completeness (20%), Logical Consistency (30%), Geographic Accuracy (10%), Content Quality (20%), Format Standardization (10%), and Innovation Degree (10%), as summarized in Table 2(a).

Validation Thresholds: Thresholds are calibrated from an empirical study on 1,000 samples (overall score 0.86), benchmark dataset standards (minimum 0.80), and alignment with human ratings. A merchant record passes if the Overall Score ≥ 0.80 , all six dimensions ≥ 0.70 , key dimensions (Logical Consistency, Data Completeness, Content Quality) meet their target thresholds, and there are no critical failures in essential fields (name format, category alignment, address consistency).

Privacy Rewriting Data Validation. It evaluates the quality of privacy rewriting against the augmented data along four dimensions: Anonymization Degree (50%), Logical Consistency (20%), Content Consistency (20%), and Innovation Degree (10%), as detailed in Table 2(b).

Validation Thresholds: Due to the higher privacy risk, this stage adopts stricter criteria informed by GDPR requirements and an empirical study on 1,000 samples (overall score 0.92, Anonymization Degree 0.95) [54]. A record passes if the Overall Score ≥ 0.85 , Anonymization Degree ≥ 0.90 , Content Consistency ≥ 0.90 , all

Table 3: Evaluation Frameworks for Data Quality Assessment

(a) Augmented Data Quality

Evaluation Dimension	Weight	Evaluation Fields
Data Completeness	20%	All
Logical Consistency	30%	<i>name, category, subcategory, price_range, business_hours, facilities</i>
Geographic Accuracy	10%	<i>city, district, landmark, address</i>
Content Quality	20%	<i>description, rating, review_count, tags, promotions, products, group_deals</i>
Format Standardization	10%	<i>phone, mobile, email</i>
Innovation Degree	10%	<i>name</i>

(b) Privacy Writing Data Quality

Evaluation Dimension	Weight	Evaluation Fields
Anonymization Degree	50%	<i>name, address, email</i>
Logical Consistency	20%	<i>description, phone, mobile</i>
Content Consistency	20%	<i>category, subcategory, city, district, landmark, price_range, business_hours, rating, review_count, delivery_available, delivery_fee, delivery_range, min_order_amount</i>
Innovation Degree	10%	<i>name</i>

Table 4: Validation Thresholds Summary

Validation Stage	Overall Score	Critical Dimensions	All Dimensions
Augmented Data	≥ 0.80	Logical Consistency ≥ 0.85 Data Completeness ≥ 0.80 Content Quality ≥ 0.80	≥ 0.70
Privacy Rewriting	≥ 0.85	Anonymization Degree ≥ 0.90 Content Consistency > 0.90	≥ 0.75

dimensions ≥ 0.75 , and there are no privacy leaks in sensitive fields (name, address, email).

3.1.4 Local Merchant Database Statistics. Overall, we constructed 1,354,185 representative samples from a pool of over 1,600,000 original seed records through data augmentation, privacy rewriting, and validation, which constitute our LOCALSEARCHBENCH database.

The proportions of each category, the number of data for each city are presented in Figure 4(c)–4(k). The distribution closely mirrors that of the raw data, ensuring our benchmark maintains the same categorical balance as real-world local life service platforms. The geographical distribution of merchant data for each city was derived through geospatial inference to generate latitude-longitude heatmaps, as shown in Figure 5. The distribution patterns align well with urban clustering phenomena and regional economic hierarchies, demonstrating the dataset’s realistic geographical representation.

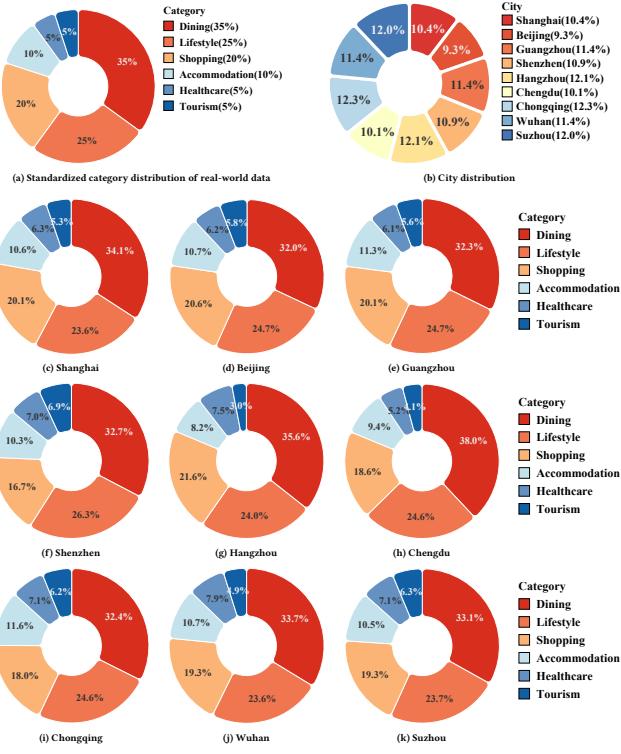


Figure 4: City distribution and category distribution of 1,354,185 merchant data across 9 major cities in China.

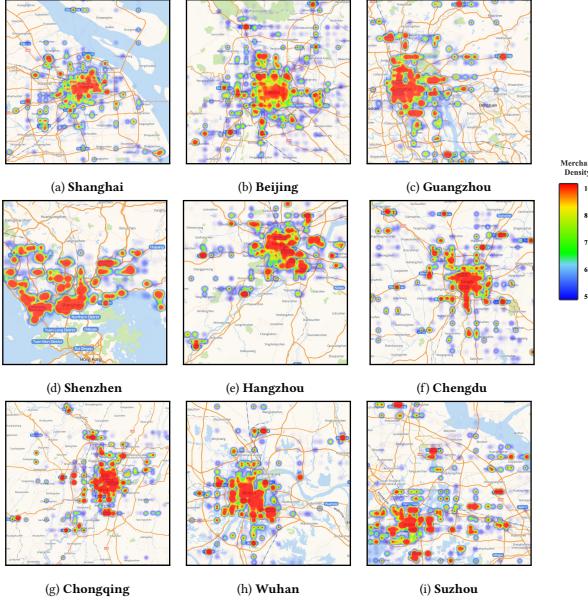


Figure 5: Geographical distribution heatmaps of 1,354,185 merchant data across 9 major cities in China. The color scale represents merchant density from 50 (blue) to 100 (red).

3.2 LOCALRAG Construction

To support efficient and accurate merchant retrieval for LRM_s, we build **LOCALRAG**, a Retrieval-Augmented Generation system on

Table 5: Intelligence Level Classification for Seed Questions.

Intelligence Level	Grade	Requirement Understanding	Planning-Search-Reflection Loop
L1	Basic	Standard expression of single precise requirement	Supply matching search with ≤ 2 conditions
L2	Advanced	Ambiguous expression of single precise requirement; Standard expression of single ambiguous requirement	Supply matching search with ≥ 3 conditions; Semantic retrieval search supply; Multi-source information integration supply
L3	Intelligent	Composite requirements; Multi-turn conversational requirements	Reasoning with certain complexity; Semantic retrieval with external information dependency; Procedural planning requirements; Comparison and contrast scenarios
L4	Expert	Personalized requirements	External hard-to-access dependency; Open-ended planning requirements
L5	AGI	Cross-platform coordination; Complex decision-making; Exception handling and reflection	Cross-platform dependency; Task execution; Adaptive reflection

our merchant database (Figure 8). LOCALRAG uses a embedding model to convert each merchant profile into a high-dimensional vector. Structured fields are directly encoded into vectors and indexed in a city-partitioned vector database. During query processing, the target city is first extracted from the query text, then the corresponding city-partitioned vector database is selected. The query is embedded with the LOCALRAG embedding model and matched via cosine similarity against merchant vectors from that city, selecting the top- N semantically related candidates through semantic retrieval. LOCALRAG then applies a reranking model to the top- N candidates, scoring query-merchant relevance and returning the top- K merchants as the final retrieval results for LLM agents.

3.3 Multi-hop QA Construction

In this subsection, we present the multi-hop QA construction of LOCALSEARCHBENCH. Building upon the data construction pipeline described in the previous subsections, our multi-hop QA construction pipeline consists of four main stages: seed question collection, question instantiation, answer collection, and QA validation. Each of these stages is detailed in the following subsubsections.

3.3.1 Seed Question Collection. To ensure LOCALSEARCHBENCH reflects real-world local life service complexities, we conducted a statistical analysis to understand user search behaviors and service interactions. Based on requirement complexity and reasoning demands, we classify questions into five intelligence levels (Table 5): **L1 (Basic)** for simple factual queries; **L2 (Advanced)** for ambiguous or multi-condition searches; **L3 (Intelligent)** for composite requirements with multi-turn reasoning; **L4 (Expert)** for personalized planning; **L5 (AGI)** for cross-platform coordination with adaptive reflection.

We collect 1,200 single-hop questions from local life service platform M (90%) and enriched sources (10%) across 6 service categories (in § 3.1.1). The platform data captures authentic user queries, while enriched data provides coverage for complex scenarios underrepresented in natural logs. Following the intelligence level taxonomy in Table 5, we focus on **L3 (Intelligent)** and **L4 (Expert)** questions, as they require multi-step reasoning within a static environment, while **L5 (AGI)** questions involve dynamic cross-platform interactions that are difficult to simulate offline.

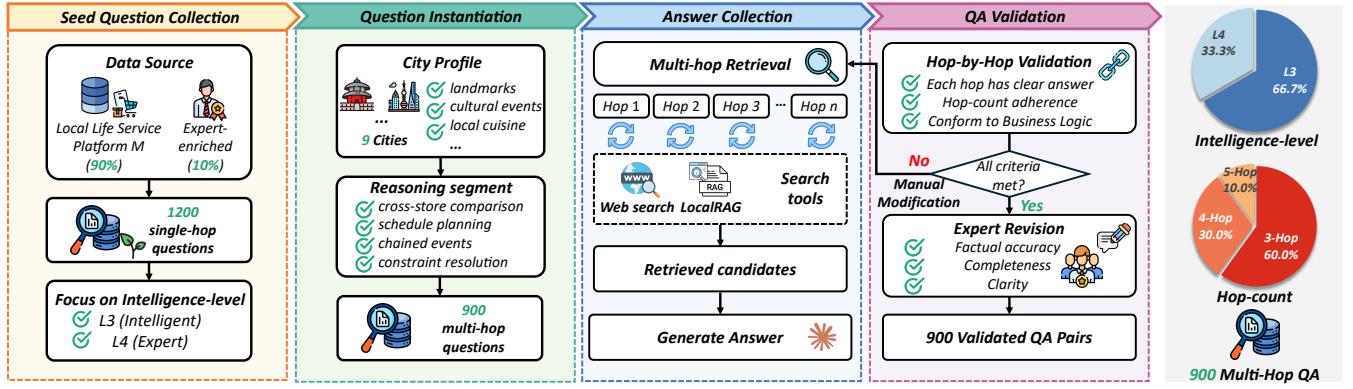


Figure 6: How Multi-hop QA pairs in LOCALSEARCHBENCH are constructed.

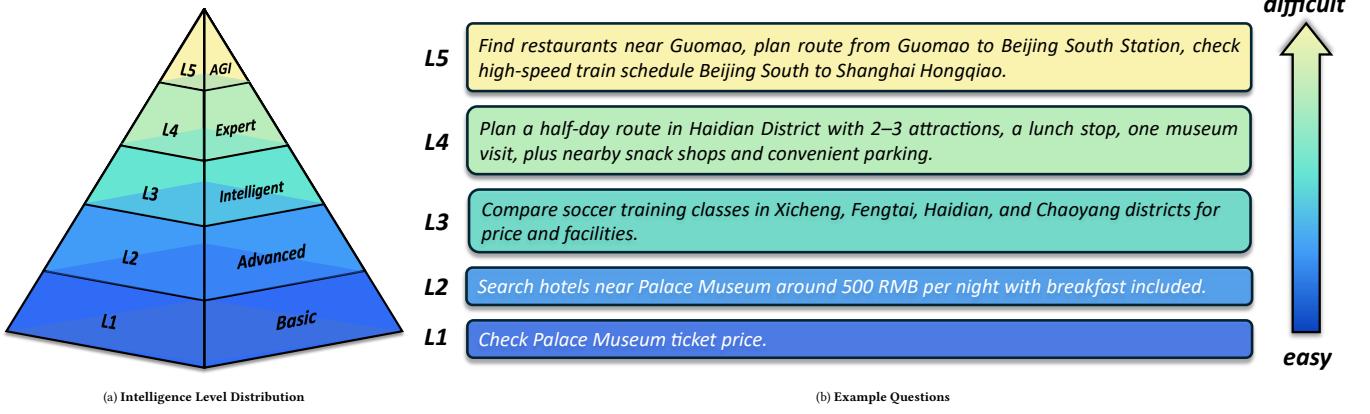


Figure 7: Intelligence Level Analysis for Collected Questions

3.3.2 Question Instantiation. We instantiate the collected questions with city-specific information through a two-phase process. First, we construct city profiles for each of the 9 cities (§ 3.1.1), collecting structured information from public resources including landmarks, cultural events, local cuisine, price bands, etc. To facilitate consistent evaluation, temporal elements are anchored to static timestamps, creating a fixed reference point for all queries. The collected city data is preprocessed using automated scripts to remove duplicates, normalize formats, and correct inconsistencies, then validated by annotators to ensure factual accuracy and cultural authenticity. Second, annotators from platform M manually instantiate questions into concrete multi-hop queries. For each city, approximately 100 questions are selected from the collected 1,200 questions and grounded with city-specific information. To transform single-hop questions into multi-hop queries, annotators enrich questions with 2-4 reasoning segments based on business relevance and logical coherence. These segments include: cross-merchant comparison, schedule planning, chained events, and constraint resolution. We maintain target intelligence-level and hop-count distributions (Figure 6), prioritizing underrepresented categories to ensure balanced coverage across service categories and cities in § 3.1.1.

3.3.3 Answer Collection. For each instantiated question, we generate reference answers through a multi-step process using LOCALRAG (§ 3.2), web search, and LLM-based generation. We first check whether the question needs real-time information (e.g., current events, weather, news) or can be answered with merchant data alone. If real-time signals are present, we invoke web search; otherwise, the query relies on the city-partitioned merchant vector database. For all questions, LOCALRAG retrieves merchants scoped to the target city. For multi-hop questions, the hop count, retrieval target, and required search tool for each hop are predetermined during question construction. Search tools then execute iterative retrieval per hop, gathering merchants across categories/locations and combining them with any web-search snippets to form the full reasoning context. Claude-4.5-opus [55] then generates preliminary answers with expert oversight. The LLM reranks each hop's retrieved candidates and selects the top- k passages ($k=5$) as evidence, then synthesizes a final answer grounded in these snippets while following the fixed hop plan and the original query. The prompt enforces: (1) grounding in retrieved evidence, (2) alignment with the fixed hop plan, (3) explicit citation of merchant and web evidence, and (4) avoidance of unsupported content. The resulting question-answer pairs serve as input for validation.

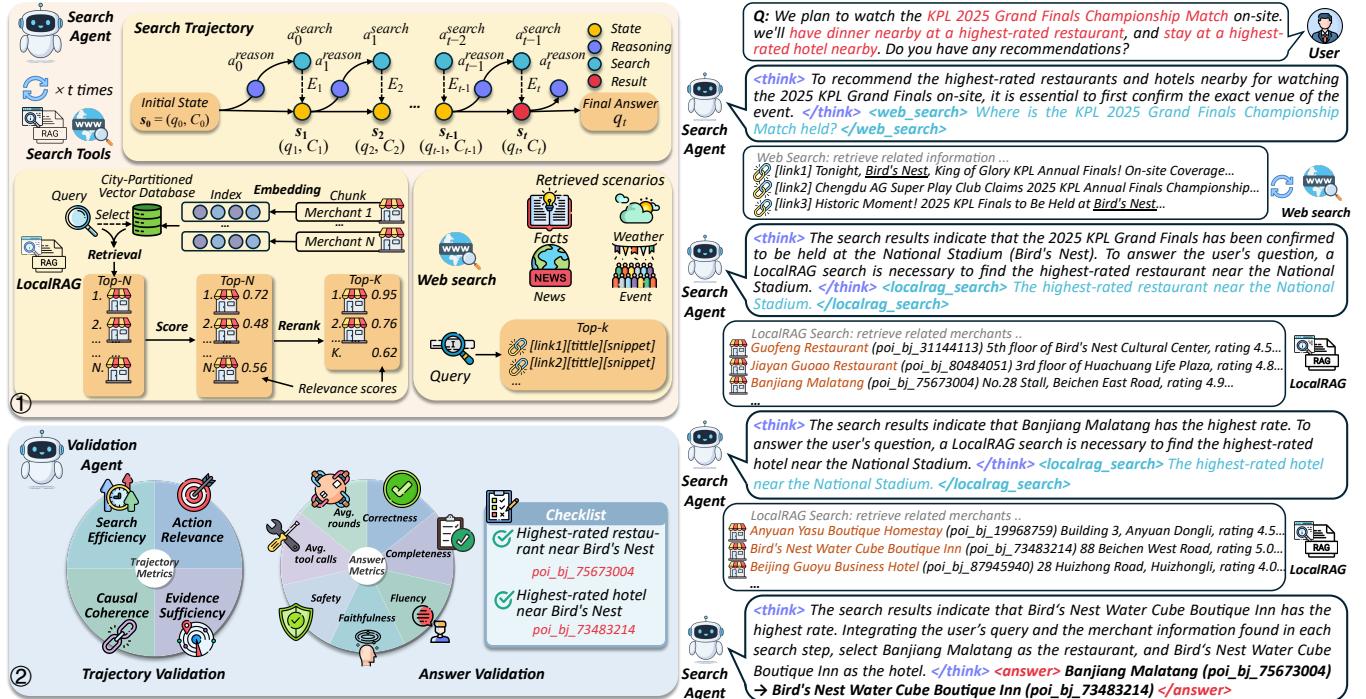


Figure 8: Overview of LOCALPLAYGROUND.

3.3.4 QA Validation. All preliminary QA pairs undergo a two-stage validation process (Figure 6). First, hop-by-hop validation ensures: (1) every hop has a clear answer, meaning retrieval targets must be found through LOCALRAG or web search with sufficient evidence; (2) strict adherence to the target hop-count chain; (3) conformity to local life service business logic. Questions failing these criteria are manually modified and re-submitted to answer collection (§ 3.3.3). Second, three domain experts independently review answers for factual accuracy, completeness, and clarity, resolving disagreements through consensus. Answers are deterministic with balanced distribution across service categories and cities (§ 3.1.1), yielding 900 high-quality QA pairs for the final benchmark (Figure 6).

4 LOCALPLAYGROUND

We design LOCALPLAYGROUND for automated testing with LOCALSEARCHBENCH. It contains a *Search Agent* for multi-hop retrieval and answering, and a *Validation Agent* for quality examination of answer and reasoning trajectory (Figure 8).

Search Agent. The Search Agent uses two tools: LOCALRAG for vector-DB merchant retrieval and web search for real-time information (facts, weather, news, events) via Baidu APIs [56]. It performs iterative search up to N rounds, calling each tool at most once per round to refine answers. For each query q_0 , it records the trajectory $\tau = \{s_0, a_0, s_1, a_1, \dots, s_t\}$, with states $s_i = (q_i, C_i)$ storing the current sub-question and accumulated evidence. At step i , it selects a reasoning action a_i^{reason} and, if needed, a search action a_i^{search} to obtain evidence E_{i+1} , update $C_{i+1} = C_i \cup E_{i+1}$, and generate q_{i+1} from (q_i, C_{i+1}) . A final reasoning step synthesizes q_t , and $s_{i+1} = F(s_i, a_i)$ forms a directed chain over sub-questions and evidence for reasoning.

Validation Agent. We employ an LLM-as-judge (Claude-4.5-opus) to evaluate each agent interaction along two dimensions: *answer quality* and *trajectory quality*. *Answer quality* is assessed on 7 metrics: correctness (factual accuracy), completeness (covers the query), fluency (readability), faithfulness (no hallucination), safety (no harmful content), avg. tool calls and avg. rounds. Each query is associated with a multi-attribute constraint *checklist* defining all necessary and sufficient conditions for a valid answer. During evaluation, the *Validation Agent* verifies whether the agent’s retrieved merchant satisfies every constraint on the checklist. If multiple merchants meet all listed conditions, each is considered a correct answer. *Trajectory quality* is scored on 4 dimensions: action relevance (step aligns with intent), evidence sufficiency (evidence is enough), causal coherence (later steps correctly reference earlier evidence), and efficiency (avoids redundant/ineffective calls).

Each instance is evaluated five times per dimension and averaged. Systems are anonymized, question order is randomized, and the judge model is distinct from all evaluated models.

5 Evaluation

5.1 Evaluated Models

We evaluate 16 state-of-the-art thinking and non-thinking models for agents: GPT series (GPT-4.1, o3 (high)) [57], Gemini series (Gemini-2.5-Flash, Gemini-2.5-Pro) [35], LongCat series (LongCat-Flash-Chat, LongCat-Flash-Thinking) [58], Deepseek-V3.2 [59], GLM-4.6 [60], and Qwen3 series (Qwen3-14B, Qwen3-32B, Qwen3-235B-A22B) [61]. Unless specified, LOCALPLAYGROUND uses default

Table 6: Performance comparison of different models on LOCALSEARCHBENCH (Max Round $N = 5$). The top and worst performing results are highlighted in red (1st) and blue (bottom) backgrounds, respectively.

Models	Answer Evaluation Metrics							Trajectory Evaluation Metrics			
	Avg. tool calls	Avg. rounds	Correctness(%)	Completeness(%)	Fluency (%)	Faithfulness(%)	Safety (%)	Action Relevance (%)	Evidence Sufficiency (%)	Causal Coherence(%)	Search Efficiency (%)
Qwen3-235B-A22B (w/o thinking)	2.00	2.93	21.20	50.94	69.16	25.28	79.72	75.22	43.61	50.68	45.99
Qwen3-235B-A22B (w/ thinking)	2.31	3.17	30.20	71.20	71.58	26.90	81.76	80.42	45.75	52.04	48.63
Qwen3-32B (w/o thinking)	2.78	3.11	19.80	40.96	68.50	21.38	80.76	74.67	46.52	48.82	49.84
Qwen3-32B (w/ thinking)	2.80	3.12	25.60	40.66	68.44	22.40	79.54	75.06	46.99	48.87	49.13
Qwen3-14B (w/o thinking)	2.53	2.07	24.20	40.60	69.62	27.44	80.78	80.46	46.44	50.96	50.02
Qwen3-14B (w/ thinking)	2.57	2.12	25.20	40.98	69.32	28.40	80.44	81.19	47.24	52.52	51.65
GPT-4.1	1.73	2.42	18.50	45.37	65.93	28.76	77.38	68.47	38.62	45.83	42.29
o3(high)	2.91	3.38	31.50	69.64	70.72	33.89	81.87	75.93	44.71	51.78	42.96
Gemini-2.5-Flash	1.84	2.51	21.00	58.44	68.04	35.72	79.70	70.58	41.61	48.94	46.91
Gemini-2.5-Pro	2.75	3.10	32.30	71.03	71.12	34.93	82.32	77.31	45.73	52.87	41.78
LongCat-Flash-Chat	2.34	3.07	25.30	52.98	69.45	27.49	83.61	77.78	47.33	50.86	52.29
LongCat-Flash-Thinking	3.04	3.20	30.70	68.83	69.07	31.47	80.10	78.50	47.37	53.18	53.27
GLM-4.6 (w/o thinking)	2.86	3.86	29.00	76.45	70.37	35.40	81.40	74.28	48.44	50.79	52.76
GLM-4.6 (w/ thinking)	3.08	4.06	32.80	76.83	70.27	37.48	81.30	77.66	48.90	52.67	54.43
Deepseek-V3.2 (w/o thinking)	3.19	4.18	31.70	77.08	70.12	35.96	81.52	75.51	48.49	52.23	54.33
Deepseek-V3.2 (w/ thinking)	3.21	4.20	35.60	77.56	70.92	39.78	81.13	75.58	48.86	52.62	54.83

Table 7: Ablation study of tool integration on thinking models using LOCALSEARCHBENCH (Max Round $N = 5$). The top and worst performing results are highlighted in red (1st) and blue (bottom) backgrounds, respectively.

Methods	Answer Evaluation Metrics							Trajectory Evaluation Metrics			
	Avg. tool calls	Avg. rounds	Correctness(%)	Completeness(%)	Fluency (%)	Faithfulness(%)	Safety (%)	Action Relevance (%)	Evidence Sufficiency (%)	Causal Coherence(%)	Search Efficiency (%)
LOCALRAG workflow	1.00	1.00	15.18	58.30	70.80	66.10	85.40	37.70	23.53	32.77	39.35
Gemini-2.5-Pro (w/o tools)	0.00	1.00	0.00*	38.42	91.65	9.82	99.89	25.21	8.72	21.70	13.62
+ LOCALRAG	1.86	2.83	23.11	63.92	82.32	38.65	84.95	68.10	37.37	46.93	39.96
+ LOCALRAG & web search	2.75	3.10	32.30	71.03	71.12	34.93	82.32	77.31	45.73	52.87	41.78
GLM-4.6 (w/o tools)	0.00	1.00	0.00*	36.06	91.92	12.12	99.67	46.88	7.92	28.85	17.08
+ LOCALRAG	2.33	3.32	23.91	69.21	83.75	41.45	84.72	63.33	36.21	48.02	52.05
+ LOCALRAG & web search	3.08	4.06	32.80	76.83	70.27	37.48	81.30	77.66	48.90	52.67	54.43
Deepseek-V3.2 (w/o tools)	0.00	1.00	0.00*	39.30	90.20	18.56	99.55	52.65	10.80	32.10	41.15
+ LOCALRAG	2.88	3.87	27.04	70.05	82.98	40.59	83.85	65.55	42.78	47.85	54.07
+ LOCALRAG & web search	3.21	4.20	35.60	77.56	70.92	39.78	81.13	75.58	48.86	52.62	54.83

* The retrieved merchant must have a poi_id from LOCALRAG.

Table 8: Sensitivity analysis of Max Round N on model performance. The top and worst performing results are highlighted in red (1st) and blue (bottom) backgrounds, respectively.

Max Round	Answer Evaluation Metrics							Trajectory Evaluation Metrics			
	Avg. tool calls	Avg. rounds	Correctness(%)	Completeness(%)	Fluency (%)	Faithfulness(%)	Safety (%)	Action Relevance (%)	Evidence Sufficiency (%)	Causal Coherence(%)	Search Efficiency (%)
$N = 2$	2.46	2.82	30.50	70.61	76.44	54.65	89.94	59.94	31.17	44.82	40.77
$N = 3$	2.78	3.30	33.50	74.79	73.57	43.50	84.58	61.68	42.96	47.42	48.50
$N = 4$	3.03	3.95	34.20	76.34	72.27	41.26	82.53	72.97	49.28	51.32	54.13
$N = 5$	3.21	4.20	35.60	77.56	70.92	39.78	81.13	75.58	48.86	52.62	54.83
$N = 6$	3.76	4.75	34.90	74.52	70.36	38.56	80.91	68.63	49.28	51.65	47.10
$N = 7$	3.87	4.85	33.70	73.60	70.19	37.36	80.84	69.50	49.37	48.20	44.99
$N = 8$	4.02	5.01	32.40	73.15	70.24	37.35	80.39	66.89	49.39	47.94	43.58

settings: maximum tool call rounds $N = 5$; LOCALRAG with Qwen3-Embedding-8B and Qwen3-Reranker-8B [62] (top-100 dense retrieval, top-20 reranking); web search top- $k = 20$; and LLM temperature = 0. Experiments run on Intel Xeon Gold 5218 @ 2.30GHz with 1 × H20-141G GPU.

5.2 Data and LOCALRAG Quality

We evaluate LOCALSEARCHBENCH database with the two-stage framework in §3.1.3, sampling 1,000 merchants. As shown in Table 18, the augmented and privacy-rewritten data achieve high overall quality scores of 0.86 and 0.92 (out of 1).

We quantify human-machine agreement on the 20 fields by computing weighted Cohen's κ [63] per field against each of three evaluators and averaging over evaluators and fields. The same 3 evaluators run blind 5-point Likert checks on consistency, completeness, and anonymization (4~5 = satisfied, 1~3 = not satisfied). The model reaches $\bar{\kappa} = 0.74$ (95% CI: [0.71, 0.77]), corresponding to 88.34% raw agreement.

Since retrieval quality directly affects downstream QA performance, we evaluate LOCALRAG on 27,360 hop-level queries using NDCG@10/20 and MRR@10/20 over the top-20 results of the full retrieval pipeline (§3.2). Our best configuration, Qwen3-Embedding-8B, reaches NDCG@10/20 of 0.84/0.87 and MRR@10/20 of 0.82/0.83 (Table 9). These hop-level queries come from the multi-hop QA construction in §3.3 with ground-truth evidence for each hop; metric definitions and full retrieval results are detailed in Appendix A.1.

5.3 Overall Performance

The choice of underlying *LRM* significantly influences performance across all metrics. Deepseek-V3.2 achieves the highest Correctness (35.60%), followed by GLM-4.6 (32.80%) and Gemini-2.5-Pro (32.36%). It also leads in Completeness (77.56%) and Faithfulness (39.78%). Thinking models outperform non-thinking ones in answer quality metrics (Correctness: 30.49% vs. 21.34%, Completeness: 64.59% vs. 49.10%) and most trajectory evaluation metrics, with similar Search Efficiency. Despite these variations, overall performance remains limited, with the best model achieving only 35.60% Correctness, underscoring the substantial challenge of multi-hop local search. Across all 16 models, correctness further drops on L4 tasks (23.51%) compared to L3 (28.66%), reflecting the greater challenge of L4 (see Appendix A.6).

Human evaluation. While LLM-as-judge provides scalable assessment, it may introduce biases. Our human evaluation validates its reliability, showing high consistency with human judgment in relative model ranking. We conduct two studies: (1) 3 annotators evaluate 900 outputs from Deepseek-V3.2, and (2) 100 randomly sampled outputs each from GLM-4.6, Gemini-2.5-Pro, LongCat-Flash-Thinking, and Qwen3-235B-A22B. Human scores demonstrate substantial inter-rater reliability (Cohen's $\kappa = 0.79$, 95% CI: [0.75, 0.83], 87.91% raw agreement). Detailed results are in Appendix A.2.

5.4 Ablation Study

We conduct ablation study on tool integration—a core design element for *agentic search* in local life services. Specifically, we evaluate three configurations across top-performing *LRMs* (Gemini-2.5-Pro, GLM-4.6, Deepseek-V3.2): (1) *LRM*-only (no tools), (2) *LRM* + LOCALRAG (merchant retrieval only), and (3) *LRM* + LOCALRAG + *web search* (full toolchain). The goal is to quantify how each tool impacts answer and trajectory quality.

As shown in Table 7, incorporating web search into LOCALPLAY-GROUND improves answer quality but reduces faithfulness. On average, it boosts Correctness by 8.90 pp and Completeness by 7.41 pp, while decreasing Faithfulness by 2.83 pp. This trade-off is most evident in Gemini-2.5-Pro, which gains 9.25 pp in Correctness but loses 3.72 pp in Faithfulness. These results highlight web search's

dual role in enabling critical information retrieval and introducing noise in multi-hop reasoning.

5.5 Sensitivity Study

To investigate the impact of maximum conversation rounds N on tool usage and model performance, we conduct a sensitivity study using Deepseek-V3.2, the best-performing model, with varying N from 2 to 8. As shown in Table 8, the optimal configuration is $N = 5$, achieving the highest Correctness (35.60%). When $N < 5$, insufficient rounds limit information gathering, resulting in lower correctness (30.50%–34.20%). When $N > 5$, despite increased tool usage, correctness declines (34.90%–32.48%), suggesting excessive rounds introduce noise that hinders answer quality. In terms of computational cost, tool calls and conversation length increase with N , peaking at $N = 8$ (4.02 tool calls, 5.01 rounds). Therefore, $N = 5$ represents the optimal balance of performance and efficiency.

5.6 In-Depth Analysis

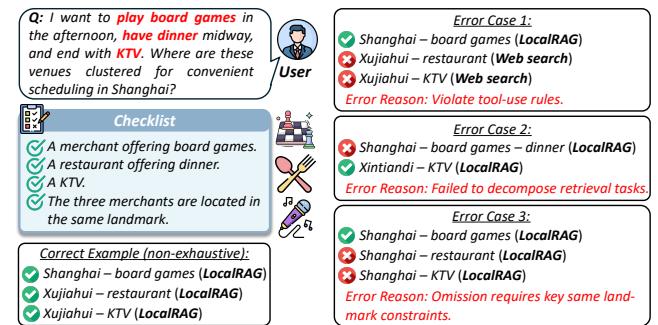


Figure 9: Failure Case Studies. A typical chain event: the 2nd and 3rd hops depend on the 1st hop.

We study factors affecting *LRM* performance on *agentic search* tasks in local life services. Figure 9 shows concrete examples.

Unstable tool call strategy (38.6%). Lack of a unified strategy for tool call timing and frequency leads to insufficient tool calls or wasted budget. *LRMs* may also violate tool-use rules, such as using web search for merchant queries or making redundant tool calls within a single dialogue turn.

Missing explicit multi-hop planning (30.9%). *LRMs* lacking hop sequences fail to decompose tasks for retrieval or tool use. Without explicit hop targets, sub-goals remain unordered, causing empty matches and higher coverage failures in multi-hop tasks. Retrieval failures cascade through steps due to the absence of validation or backtracking.

Query generation dilutes critical constraints (11.8%). Naively generated long-form queries often omit key constraints, weakening retrieval signals and reducing recall of relevant candidates.

Long-context noise (18.7%). Lengthy search results and information sent to *LRMs* bring irrelevant data, reduce the ability to distinguish relevant results, and can exceed context window limits.

6 Conclusion

We propose LOCALSEARCHBENCH, a benchmark for *agentic search* in local life services evaluating multi-hop reasoning. The benchmark contains 1,354,185 database and 900 multi-hop QA tasks across 9 Chinese cities and 6 service categories. Experiments on 16 leading LRM_s using LOCALPLAYGROUND show current models struggle, with DeepSeek-V3.2 achieving only 35.60% correctness. We open-source LOCALSEARCHBENCH and LOCALPLAYGROUND to advance domain-specific agentic search.

Limitations

We highlight two main limitations.

Dataset construction and agent workflows. Our dataset construction and agent workflows rely on large, often closed-source LLM_s and human-in-the-loop verification. This dependence improves generation quality but raises reproducibility and cost concerns, consuming approximately 200M tokens and costing approximately \$1,200 for API usage, plus over 220 hours of human labor for verification; open-source substitutes require substantially more compute and operational effort to match performance.

Language and Locale Specificity. The benchmark is monolingual (Chinese) and geographically scoped to mainland China. This design prioritizes domain depth, authenticity, and precise geospatial grounding, which are critical for evaluating complex, real-world planning. However, it inherently limits direct applicability to other languages and regions, posing a challenge for models without relevant language or domain knowledge. The core framework LOCALPLAYGROUND, including its task formulation, evaluation suite, and tool-integrated environment, remains fully generalizable. Therefore, this work provides a necessary, localized instantiation and a clear methodological template for the community to build and compare culturally adapted benchmarks.

References

- [1] Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao, and Yong Li. 2025. Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models. *arXiv:2501.09686 [cs.AI]* <https://arxiv.org/abs/2501.09686>
- [2] Mohamed Amine Ferrag, Norbert Tihanyi, and Merouane Debbah. 2025. From LLM Reasoning to Autonomous AI Agents: A Comprehensive Review. *arXiv:2504.19678 [cs.AI]* <https://arxiv.org/abs/2504.19678>
- [3] Junde Wu, Jiayuan Zhu, Yuyuan Liu, Min Xu, and Yueming Jin. 2025. Agentic Reasoning: A Streamlined Framework for Enhancing LLM Reasoning with Agentic Tools. *arXiv:2502.04644 [cs.AI]* <https://arxiv.org/abs/2502.04644>
- [4] Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. 2025. Webthinker: Empowering large reasoning models with deep research capability. *arXiv preprint arXiv:2504.21776* (2025).
- [5] Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, Peifeng Wang, Silvio Savarese, et al. 2025. A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems. *arXiv preprint arXiv:2504.09037* (2025).
- [6] Jiajun Chai, Guojun Yin, Zekun Xu, Chuhuai Yue, Yi Jia, Siyu Xia, Xiaohan Wang, Jiwen Jiang, Xiaoguang Li, Chengqi Dong, Hang He, and Wei Lin. 2025. RLFFactory: A Plug-and-Play Reinforcement Learning Post-Training Framework for LLM Multi-Turn Tool-Use. *arXiv:2509.06980 [cs.LG]* <https://arxiv.org/abs/2509.06980>
- [7] Hao Chen, Zhexin Hu, Jiajun Chai, Haocheng Yang, Hang He, Xiaohan Wang, Wei Lin, Luhang Wang, Guojun Yin, and Zhuofeng Zhao. 2025. ToolForge: A Data Synthesis Pipeline for Multi-Hop Search without Real-World APIs. *arXiv:2512.16149 [cs.AI]* <https://arxiv.org/abs/2512.16149>
- [8] Shuang Sun, Huatong Song, Yuhao Wang, Ruiyang Ren, Jinhao Jiang, Junjie Zhang, Fei Bai, Jia Deng, Wayne Xin Zhao, Zheng Liu, Lei Fang, Zhongyuan Wang, and Ji-Rong Wen. 2025. SimpleDeepSearcher: Deep Information Seeking via Web-Powered Reasoning Trajectory Synthesis. *arXiv:2505.16834 [cs.CL]* <https://arxiv.org/abs/2505.16834>
- [9] Dingchu Zhang, Yida Zhao, Jialong Wu, Baixuan Li, Wenbiao Yin, Liwen Zhang, Yong Jiang, Yufeng Li, Kewei Tu, Pengjun Xie, and Fei Huang. 2025. EvolveSearch: An Iterative Self-Evolving Search Agent. *arXiv:2505.22501 [cs.CL]* <https://arxiv.org/abs/2505.22501>
- [10] Wenzuan Shi, Haochen Tan, Chuqiao Kuang, Xiaoguang Li, Xiaozhe Ren, Chen Zhang, Hanting Chen, Yasheng Wang, Lifeng Shang, Fisher Yu, and Yunhe Wang. 2025. Pangu DeepDiver: Adaptive Search Intensity Scaling via Open-Web Reinforcement Learning. *arXiv:2505.24332 [cs.CL]* <https://arxiv.org/abs/2505.24332>
- [11] Zhengliang Shi, Shen Gao, Lingyong Yan, Yue Feng, Xiuyi Chen, Zhumin Chen, Dawei Yin, Suzan Verberne, and Zhaochun Ren. 2025. Tool Learning in the Wild: Empowering Language Models as Automatic Tool Agents. *arXiv:2405.16533 [cs.CL]* <https://arxiv.org/abs/2405.16533>
- [12] Weizhi Zhang, Yangning Li, Yuanchen Bei, Junyu Luo, Guanchen Wan, Liangwei Yang, Chenxuan Xie, Yuyao Yang, Wei-Chieh Huang, Chunyu Miao, et al. 2025. From Web Search towards Agentic Deep Research: Incentivizing Search with Reasoning Agents. *arXiv preprint arXiv:2506.18959* (2025).
- [13] Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366* (2025).
- [14] Fengran Mo, Chuan Meng, Mohammad Aliannejadi, and Jian-Yun Nie. 2025. Conversational search: From fundamentals to frontiers in the LLM era. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 4094–4097.
- [15] Chuhuai Yue, Chengqi Dong, Yinan Gao, Hang He, Jiajun Chai, Guojun Yin, and Wei Lin. 2025. Promoting Efficient Reasoning with Verifiable Stepwise Reward. *arXiv:2508.10293 [cs.AI]* <https://arxiv.org/abs/2508.10293>
- [16] Chuhuai Yue, Jiajun Chai, Yufei Zhang, Zixiang Ding, Xihao Liang, Peixin Wang, Shihai Chen, Wang Yixuan, wangyanping, Wei Lin, and Guojun Yin. 2025. UIOrchestra: Generating High-Fidelity Code from UI Designs with a Multi-agent System. In *The 2025 Conference on Empirical Methods in Natural Language Processing*. <https://openreview.net/forum?id=0T579LJSNZ>
- [17] Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous Agents. *arXiv:2307.13854 [cs.AI]* <https://arxiv.org/abs/2307.13854>
- [18] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2023. WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents. *arXiv:2207.01206 [cs.CL]* <https://arxiv.org/abs/2207.01206>
- [19] Xiaocong 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*. 4566–4577.
- [20] Xiaocong Lan, Chen Gao, Shiqi Wen, Xiuqi Chen, Yingge Che, Han Zhang, Huazhou Wei, Hengliang Luo, and Yong Li. 2023. NEON: Living Needs Prediction System in Meituan. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Long Beach, CA, USA) (KDD '23). Association for Computing Machinery, New York, NY, USA, 5292–5302. <https://doi.org/10.1145/3580305.3599874>
- [21] Yinfeng Li, Chen Gao, Xiaoyi Du, Huazhou Wei, Hengliang Luo, Depeng Jin, and Yong Li. 2022. Automatically Discovering User Consumption Intents in Meituan. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Washington DC, USA) (KDD '22). Association for Computing Machinery, New York, NY, USA, 3259–3269. <https://doi.org/10.1145/3534678.3539122>
- [22] Kalle Kusk. 2025. Flexible Platforms? An Ethnographic Study of Flexible Scheduling in Platform-Mediated Delivery. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–11.
- [23] Chang Liu, Huan Yan, Hongjie Sui, Haomin Wen, Yuan Yuan, Yuyang Han, Hongsen Liao, Xuetao Ding, Jinghua Hao, and Yong Li. 2025. Mrgrp: Empowering courier route prediction in food delivery service with multi-relational graph. In *Companion Proceedings of the ACM on Web Conference 2025*. 364–373.
- [24] ZhiLin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600* (2018).
- [25] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics* 10 (2022), 539–554.
- [26] Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glasee. 2025. BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. *arXiv:2504.12516 [cs.CL]* <https://arxiv.org/abs/2504.12516>
- [27] 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* (2025).

- [28] Tian Lan, Bin Zhu, Qianghuai Jia, Junyang Ren, Haijun Li, Longyue Wang, Zhao Xu, Weihua Luo, and Kaifu Zhang. 2025. DeepWideSearch: Benchmarking Depth and Width in Agentic Information Seeking. *arXiv:2510.20168 [cs.CL]* <https://arxiv.org/abs/2510.20168>
- [29] Corby Rosset, Ho-Lam Chung, Guanghui Qin, Ethan C Chau, Zhuo Feng, Ahmed Awadallah, Jennifer Neville, and Nikhil Rao. 2024. Researchy questions: A dataset of multi-perspective, decompositional questions for llm web agents. *arXiv preprint arXiv:2402.17896* (2024).
- [30] Haochen Tan, Zhijiang Guo, Zhan Shi, Lu Xu, Zhili Liu, Yunlong Feng, Xiaoguang Li, Yasheng Wang, Lifeng Shang, Qun Liu, et al. 2024. Proxyqa: An alternative framework for evaluating long-form text generation with large language models. *arXiv preprint arXiv:2401.15042* (2024).
- [31] Boyu Gou, Zanming Huang, Yuting Ning, Yu Gu, Michael Lin, Weijian Qi, Andrei Kopanev, Botao Yu, Bernal JimÁlnez GutiÁlrez, Yiheng Shu, Chan Hee Song, Jiaman Wu, Shijie Chen, Hanane Nour Moussa, Tianshu Zhang, Jian Xie, Yifei Li, Tianci Xue, Zeyi Liao, Kai Zhang, Boyuan Zheng, Zhaowei Cai, Viktor Rozgic, Morteza Ziyadi, Huan Sun, and Yu Su. 2025. Mind2Web 2: Evaluating Agentic Search with Agent-as-a-Judge. *arXiv:2506.21506 [cs.AI]* <https://arxiv.org/abs/2506.21506>
- [32] Yingxuan Yang, Mulei Ma, Yuxuan Huang, Huacan Chai, Chenyu Gong, Haoran Geng, Yuanjian Zhou, Ying Wen, Meng Fang, Muham Chen, et al. 2025. Agentic web: Weaving the next web with ai agents. *arXiv preprint arXiv:2507.21206* (2025).
- [33] Yuchen Li, Hengyi Cai, Rui Kong, Xinran Chen, Jianmin Chen, Jun Yang, Haojie Zhang, Jiayi Li, Jiayi Wu, Yiqun Chen, et al. 2025. Towards AI Search Paradigm. *arXiv preprint arXiv:2506.17188* (2025).
- [34] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720* (2024).
- [35] Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multi-modality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).
- [36] Tongyi DeepResearch Team. 2025. Tongyi-DeepResearch. <https://github.com/Alibaba-NLP/DeepResearch>.
- [37] 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* (2025).
- [38] Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592* (2025).
- [39] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060* (2020).
- [40] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350* (2022).
- [41] Peilin Zhou, Bruce Leon, Xiang Ying, Can Zhang, Yifan Shao, Qichen Ye, Dading Chong, Zhiling Jin, Chenxuan Xie, Meng Cao, Yuxin Gu, Sixin Hong, Jing Ren, Jian Chen, Chao Liu, and Yining Hua. 2025. BrowseComp-ZH: Benchmarking Web Browsing Ability of Large Language Models in Chinese. *arXiv:2504.19314 [cs.CL]* <https://arxiv.org/abs/2504.19314>
- [42] Yunjia Xi, Jianghao Lin, Menghui Zhu, Yongzhao Xiao, Zhuoying Ou, Jiaqi Liu, Tong Wan, Bo Chen, Weiwen Liu, Yasheng Wang, et al. 2025. InfoDeepSeek: Benchmarking Agentic Information Seeking for Retrieval-Augmented Generation. *arXiv preprint arXiv:2505.15872* (2025).
- [43] Thinh Pham, Nguyen Nguyen, Pratibha Zunjare, Weiyuan Chen, Yu-Min Tseng, and Tu Vu. 2025. SealQA: Raising the Bar for Reasoning in Search-Augmented Language Models. *arXiv preprint arXiv:2506.01062* (2025).
- [44] Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, et al. 2024. Long-form factuality in large language models. *Advances in Neural Information Processing Systems* 37 (2024), 80756–80827.
- [45] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974* (2020).
- [46] Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. *arXiv preprint arXiv:2011.03088* (2020).
- [47] Shengkang Wang, Hongzhan Lin, Ziyang Luo, Zhen Ye, Guang Chen, and Jing Ma. 2024. Mfc-bench: Benchmarking multimodal fact-checking with large vision-language models. *arXiv preprint arXiv:2406.11288* (2024).
- [48] Nikos I Bosse, Jon Evans, Robert G Gambee, Daniel Hnyk, Peter Mühlbacher, Lawrence Phillips, Dan Schwarz, Jack Wildman, et al. 2025. Deep Research Bench: Evaluating AI Web Research Agents. *arXiv preprint arXiv:2506.06287* (2025).
- [49] Huaying Yuan, Zheng Liu, Junjie Zhou, Hongjin Qian, Ji-Rong Wen, and Zhicheng Dou. 2025. VideoDeepResearch: Long Video Understanding With Agentic Tool Using. *arXiv preprint arXiv:2506.10821* (2025).
- [50] Chengqi Dong, Chuhuai Yue, Hang He, Rongge Mao, Fenghe Tang, S Kevin Zhou, Zekun Xu, Xiaohai Wang, Jiajun Choi, Wei Lin, and Guojun Yin. 2025. Training Multi-Image Vision Agents via End2End Reinforcement Learning. *arXiv:2512.08980 [cs.CV]* <https://arxiv.org/abs/2512.08980>
- [51] Peiyang Liu, Xi Wang, Ziqiang Cui, and Wei Ye. 2025. Queries Are Not Alone: Clustering Text Embeddings for Video Search. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 874–883.
- [52] Peiyang Liu, Jinyu Yang, Lin Wang, Sen Wang, Yunlai Hao, and Huihui Bai. 2023. Retrieval-Based Unsupervised Noisy Label Detection on Text Data. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 4099–4104.
- [53] Peiyang Liu. 2024. Unsupervised corrupt data detection for text training. *Expert Systems with Applications* 248 (2024), 123353.
- [54] Peiyang Liu, Ziqiang Cui, Di Liang, and Wei Ye. 2025. Who Stole Your Data? A Method for Detecting Unauthorized RAG Theft. *arXiv:2510.07728* <https://arxiv.org/abs/2510.07728>
- [55] Anthropic. 2025. Claude 4.5 opus Model. <https://www.anthropic.com/news/claudie-4-5-opus/>
- [56] Yiding Liu, Weixue Lu, Suqi Cheng, Daiting Shi, Shuaiqiang Wang, Zhicong Cheng, and Dawei Yin. 2021. Pre-trained language model for web-scale retrieval in baidu search. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 3365–3375.
- [57] OpenAI. 2025. GPT-4.1 Model. <https://openai.com/index/gpt-4-1/>
- [58] Meituan LongCat Team, Bei Li, Bingye Lei, Bo Wang, Bolin Rong, Chao Wang, Chao Zhang, Chen Gao, Chen Zhang, Cheng Sun, et al. 2025. LongCat-Flash Technical Report. *arXiv preprint arXiv:2509.01322* (2025).
- [59] DeepSeek-AI. 2025. DeepSeek-V3.1 Model. <https://api.deepseek.com>
- [60] Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471* (2025).
- [61] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengan Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).
- [62] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176* (2025).
- [63] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

A Evaluation Results and Performance Analysis

A.1 Ranking Metrics Definitions

For completeness, we summarize the ranking metrics used to evaluate retrieval quality. Given a set of hop-level queries Q and a cutoff $k \in \{10, 20\}$, NDCG@ k and MRR@ k are defined as

$$\text{NDCG}@k = \frac{1}{|Q|} \sum_{q \in Q} \frac{\text{DCG}@k(q)}{\text{IDCG}@k(q)}, \quad (1)$$

with

$$\text{DCG}@k(q) = \sum_{i=1}^k \frac{2^{\text{rel}_{q,i}} - 1}{\log_2(i+1)}, \quad (2)$$

and

$$\text{MRR}@k = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{\text{rank}_q}, \quad (3)$$

where $\text{rel}_{q,i}$ is the relevance label of the i -th retrieved document for query q , and rank_q is the rank of the first relevant document within the top- k results (0 if none is found).

For hop-level retrieval evaluation, relevance labels are assigned via rule-based automatic annotation with human sampling validation. A merchant is labeled as relevant ($\text{rel}_{q,i}=1$) if it fully satisfies all constraints of the hop query (e.g., geographic scope, service category, rating threshold); otherwise, it is labeled as irrelevant ($\text{rel}_{q,i}=0$). All relevant merchants constitute the ground-truth evidence set for the query–retrieval is considered effectively recalled if any ground-truth merchant is included in the results, and ranking quality is measured by NDCG@ k to reflect the position of ground-truth merchants. To ensure reliability, 10% of labels are manually validated by two domain experts, with inter-annotator agreement Cohen’s $\kappa = 0.81$ (95% CI: [0.78, 0.84]).

We evaluate LOCALRAG’s retrieval stage (retrieval + reranker) on 27,360 hop-level queries. Embeddings are produced by the respective embedding models listed below, while the reranker is fixed to Qwen3-Reranker-8B for all configurations. The metrics reported are NDCG@10/20 and MRR@10/20 computed over the top-20 candidates returned by the full pipeline (dense retrieval followed by reranking). Table 9 summarizes overall retrieval quality across embedding backbones.

Table 9: Overall retrieval quality of LOCALRAG on hop-level queries.

Method	NDCG@10	NDCG@20	MRR@10	MRR@20
BM25	0.60	0.64	0.55	0.57
Qwen3-Embedding-8B	0.84	0.87	0.82	0.83
Qwen3-Embedding-4B	0.78	0.81	0.75	0.76
Qwen3-Embedding-0.6B	0.70	0.74	0.66	0.68
BGE-M3	0.72	0.75	0.67	0.68

Independent human relevance validation. We note that rule-based labels are generated within the same pipeline that instantiates hop queries and collects evidence, which can create coupling between ground-truth signals and the retrieval system. To quantify this effect, we performed an independent human validation on a stratified sample of 1,000 hop-level queries (stratified by city, hop-count and service category). Three domain experts annotated the top-20 retrievals per query (blind to pipeline labels); majority vote produced the human reference labels.

For the sampled subset we compute NDCG@10/20 and MRR@10/20 under both the pipeline rule-based labels and the independent human labels, report inter-annotator Cohen’s κ , and measure Spearman rank correlation of model rankings between the two label sets. Table 10 is a template for reporting these results.

Table 10: Comparison of retrieval metrics between independent human judgments and pipeline-derived relevance labels

Metric	Rule labels	Human labels	Delta (abs)	Delta (%)
NDCG@10	0.842	0.838	0.004	0.48%
NDCG@20	0.871	0.868	0.003	0.34%
MRR@10	0.821	0.818	0.003	0.37%
MRR@20	0.832	0.829	0.003	0.36%
Inter-annotator Cohen’s κ	0.81 (95% CI: [0.78, 0.84])			
Spearman rank ρ (rule vs human)	0.93			

If the observed deltas are small (e.g., absolute $\Delta < 1\%$ and Spearman $\rho \geq 0.90$), we treat the pipeline-derived labels as robust for retrieval evaluation; larger discrepancies are reported and discussed in the main text.

A.2 Human Evaluation Results

To complement the automated LLM-as-judge evaluation, we conduct human evaluation on the best-performing model from our experiments. Three experienced annotators independently assess the outputs of Deepseek-V3.2 (w/ thinking) on a sampled subset of LOCALSEARCHBENCH, evaluating both answer quality and trajectory quality using the same metrics as our automated evaluation framework. The detailed human evaluation results are presented in Table 11.

To quantify the agreement between human evaluators, we compute weighted Cohen’s κ for each metric across the three annotators. The average κ across all metrics is 0.79 (95% CI: [0.75, 0.83]), corresponding to 87.91% raw agreement, indicating substantial inter-rater reliability.

The weighted Cohen’s kappa is computed as

$$\kappa_w = \frac{P_o - P_e}{1 - P_e}, \quad (4)$$

where

$$P_o = \sum_{i,j} w_{ij} p_{ij}, \quad P_e = \sum_{i,j} w_{ij} p_i \cdot p_j, \quad (5)$$

with p_{ij} the observed proportion of ratings in category i by rater A and j by rater B, p_i and p_j the marginal proportions, and w_{ij} the weights. We use quadratic weights

$$w_{ij} = 1 - \frac{(i - j)^2}{(K - 1)^2}, \quad (6)$$

for K ordered categories. For multiple annotators we average pairwise weighted kappas.

A.3 Retrieval and Reranking Sensitivity Analysis

To investigate how sensitive our evaluation results are to the choice of retriever and reranker components, we conduct a comprehensive sensitivity analysis across multiple retrieval configurations. This analysis is crucial for understanding the robustness of our findings

Table 11: Human evaluation results of Deepseek-V3.2 on LOCALSEARCHBENCH (Max Round N = 5).

Model	Answer Evaluation Metrics							Trajectory Evaluation Metrics				
	Avg. tool calls	Avg. rounds	Correctness(%)	Completeness(%)	Fluency (%)	Faithfulness(%)	Safety (%)	Action Relevance (%)	Evidence Sufficiency (%)	Causal Coherence(%)	Search Efficiency (%)	
Deepseek-V3.2 (w/ thinking)	3.21	4.20	33.50	75.42	69.18	36.94	79.67	72.85	47.29	51.06	53.71	

and provides guidance for practitioners selecting retrieval systems for agentic search tasks.

Experimental Setup: We evaluate the performance of DeepSeek-V3.2 (the best-performing model in our primary evaluation) across eight different retrieval configurations, combining four embedding models with two reranking strategies:

- **Embedding Models:** Qwen3-Embedding-8B, bge-m3, Qwen3-Embedding-4B, Qwen3-Embedding-0.6B
- **Reranking Strategies:** Qwen3-Reranker-8B (our primary reranker), and a baseline no-reranking approach using only embedding similarity scores

Each configuration uses identical search parameters (top-20 retrieval with reranking when applicable) and is evaluated on the same 900 test instances from our benchmark.

Performance Variation Across Configurations: Table 12 summarizes the performance of DeepSeek-V3.2 across different retrieval configurations, measured by key answer-level and trajectory-level metrics.

Key Findings on Retrieval Sensitivity:

- (1) **Embedding Model Impact:** Performance varies significantly across embedding backbones, with Qwen3-Embedding-8B achieving 35.60% correctness compared to 28.76% for Qwen3-Embedding-0.6B—a relative difference of 23.8%. This suggests that embedding quality directly influences downstream agentic reasoning capabilities.
- (2) **Reranking Benefits:** The Qwen3-Reranker-8B consistently improves performance across all metrics and embedding models. For instance, reranking boosts correctness by 4.37 percentage points on average (from 28.26% to 32.63% without reranking). However, reranking comes at a computational cost, reducing Search Efficiency by approximately 4-5 percentage points.
- (3) **Trade-offs and Optimal Configurations:** The optimal configuration (Qwen3-8B + Qwen3-Reranker) achieves the highest correctness and evidence quality but at the cost of reduced search efficiency. Configurations without reranking show higher efficiency but lower answer quality, suggesting a clear trade-off between computational cost and performance.
- (4) **Metric-Specific Sensitivity:** Different metrics show varying sensitivity to retrieval quality. Correctness and Faithfulness are most sensitive (variation range: 10.26% and 10.33% respectively), while Search Efficiency shows relatively stable performance across configurations.

Implications for Robustness Analysis: Our sensitivity analysis reveals that evaluation results can vary by up to 24% depending on retrieval configuration choice. This underscores the need for systematic robustness studies that evaluate agentic search systems across multiple retrieval backbones rather than relying on single configurations.

This sensitivity analysis demonstrates that while our primary findings hold across multiple configurations, the choice of retrieval system significantly impacts quantitative results. Practitioners should consider these trade-offs when deploying agentic search systems in real-world applications.

A.4 Judge Model Robustness Analysis

To address potential bias from using the same model family (Claude-4.5-opus) for both ground truth generation and evaluation, we conduct a comprehensive robustness analysis using two independent judge models: Claude-4.5-opus and GPT-4-turbo. This analysis evaluates whether model rankings remain stable across different judges, quantifies inter-judge agreement, and assesses the calibration of absolute scores.

Evaluation Setup: We evaluate the same set of 900 test instances across all models using both judge models. Each judge independently scores all metrics using identical prompts and evaluation criteria. This comprehensive analysis covers all 14 model configurations from our primary evaluation, enabling us to assess both ranking stability and absolute score calibration across the full model spectrum.

Inter-Judge Agreement: The correlation between Claude-4.5-opus and GPT-4-turbo scores across all metrics and models is 0.89 (Spearman’s ρ , $p < 0.001$), indicating strong agreement. Agreement varies across metrics, with highest consistency for Correctness ($\rho = 0.94$) and relatively lower for Faithfulness ($\rho = 0.82$). The high agreement level suggests that both judges apply evaluation criteria consistently, despite potential differences in their training data and architectural approaches.

Comprehensive Model Performance Across Judges: Table 13 presents the performance comparison across judge models for all evaluated models, ranked by overall performance under Claude-4.5-opus.

Model Ranking Stability: The ranking stability across judges is remarkably high. The top-5 model rankings remain identical across both judges, with only minor reordering within the middle ranks. Key observations include:

- (1) **Top Performer Consistency:** DeepSeek-V3.2 (with thinking) maintains its top position across both judges, demonstrating robust superiority in agentic search capabilities.
- (2) **Thinking vs. Non-Thinking Gap:** The performance gap between thinking and non-thinking variants is consistently observed across judges, with thinking models showing 3-5% higher correctness scores.
- (3) **Family-Specific Patterns:** Models within the same family (e.g., Qwen3 variants, Gemini variants) maintain their relative ordering, suggesting that judge models agree on family-level capabilities.

Table 12: DeepSeek-V3.2 performance across different retrieval configurations

Configuration	Correctness (%)	Completeness (%)	Faithfulness (%)	Evidence Sufficiency (%)	Causal Coherence (%)	Search Efficiency (%)
Qwen3-8B + Qwen3-Reranker	35.60	77.56	39.78	48.86	52.62	54.83
bge-m3 + Qwen3-Reranker	32.85	74.23	36.91	45.12	49.78	51.94
Qwen3-4B + Qwen3-Reranker	31.42	72.89	35.67	43.85	48.92	50.23
Qwen3-0.6B + Qwen3-Reranker	28.76	69.45	32.34	41.23	46.78	47.56
Qwen3-8B + No Reranking	31.23	73.89	35.42	44.67	48.91	57.12
bge-m3 + No Reranking	28.94	71.34	33.18	42.45	47.23	54.67
Qwen3-4B + No Reranking	27.51	68.92	31.89	40.78	45.67	52.89
Qwen3-0.6B + No Reranking	25.34	65.78	29.45	38.92	43.56	50.34

Table 13: Comprehensive model performance comparison across judge models

Model	Claude-4.5-opus			GPT-4-turbo			Rank Stability
	Correctness	Completeness	Overall	Correctness	Completeness	Overall	
DeepSeek-V3.2 (w/ thinking)	35.60	77.56	56.58	34.85	76.92	55.89	✓
GLM-4.6 (w/ thinking)	32.80	76.83	54.82	32.12	76.45	54.29	✓
Gemini-2.5-Pro (w/ thinking)	32.30	71.03	51.67	31.89	70.78	51.34	✓
Qwen3-235B-A22B (w/ thinking)	30.20	71.20	50.70	29.87	70.91	50.39	✓
LongCat-Flash-Thinking	30.70	68.83	49.77	30.45	68.67	49.56	✓
Gemini-2.5-Flash	21.00	58.44	39.72	20.67	58.12	39.40	✓
Qwen3-32B (w/ thinking)	25.60	40.66	33.13	25.34	40.45	32.90	✓
DeepSeek-V3.2 (w/o thinking)	31.70	77.08	54.39	31.23	76.78	54.01	✓
GLM-4.6 (w/o thinking)	29.00	76.45	52.73	28.67	76.12	52.40	✓
LongCat-Flash-Chat	25.30	52.98	39.14	25.01	52.67	38.84	✓
Qwen3-235B-A22B (w/o thinking)	21.20	50.94	36.07	20.89	50.67	35.78	✓
Qwen3-14B (w/ thinking)	25.20	40.98	33.09	24.89	40.73	32.81	✓
Qwen3-32B (w/o thinking)	19.80	40.96	30.38	19.56	40.78	30.17	✓
Qwen3-14B (w/o thinking)	24.20	40.60	32.40	23.94	40.35	32.15	✓

Absolute Score Calibration: Beyond ranking stability, we analyze the calibration of absolute scores between judges. The mean absolute difference in correctness scores is 0.42% (SD = 0.31%), indicating excellent calibration. For completeness scores, the mean difference is 0.35% (SD = 0.28%). These small differences suggest that both judges apply similar scoring standards, with no systematic bias toward over- or under-estimation.

Metric-Specific Agreement Analysis: Table 14 breaks down inter-judge agreement by metric type, revealing nuanced patterns in evaluation consistency.

Table 14: Inter-judge agreement across all evaluation metrics

Metric	Spearman's ρ	Mean Abs. Diff.	Max Abs. Diff.
Correctness	0.94	0.42%	0.75%
Completeness	0.91	0.35%	0.61%
Faithfulness	0.82	1.45%	2.12%
Fluency	0.87	0.53%	0.89%
Safety	0.95	0.26%	0.42%
Overall Agreement	0.89	0.60%	2.12%

Implications for Benchmark Reliability: The comprehensive judge robustness analysis validates the reliability of our evaluation framework across multiple dimensions:

- (1) **Ranking Stability:** Model rankings are highly stable across judges, providing confidence in comparative evaluations.
- (2) **Absolute Score Calibration:** The small absolute differences indicate that scores are well-calibrated and not subject to systematic judge-specific biases.

- (3) **Metric Consistency:** All evaluation metrics demonstrate strong inter-judge agreement, with varying degrees of consistency across different aspects of agentic search performance.
- (4) **Framework Validation:** The use of Claude-4.5-opus for both ground truth generation and evaluation does not introduce systematic bias that affects model comparisons.

This extended analysis demonstrates that our evaluation results are robust not only for the best-performing model but across the entire spectrum of evaluated systems, providing strong validation of the benchmark's reliability and discriminatory power.

A.5 Comprehensive Retrieval Performance Across Data Processing Stages

To fully validate the data augmentation and privacy rewriting pipeline, we conduct a comprehensive experiment comparing retrieval performance across three data processing stages: (1) original seed data, (2) augmented data, and (3) privacy-rewritten data.

Experimental Setup: Using the same 10,000 merchant records and 1,000 search queries as in the privacy validation experiment, we evaluate retrieval performance at each processing stage using identical retrieval configurations (Qwen3-Embedding-8B + Qwen3-Reranker-8B).

Data Processing Stages:

- **Original Seed Data:** Raw merchant records with only 12 basic fields (name, category, address, etc.)
- **Augmented Data:** Enhanced records with 29 fields through generative augmentation

- **Privacy-Rewritten Data:** Final dataset with sensitive fields anonymized while preserving business information

Results: The augmentation process significantly improves retrieval quality by enriching merchant profiles with detailed attributes, while privacy rewriting maintains this enhanced performance with minimal degradation.

Table 15: Retrieval performance across data processing stages

Metric	Original Seed	Augmented	Privacy-Rewritten
NDCG@10	0.723	0.842	0.839
NDCG@20	0.752	0.871	0.868
MRR@10	0.684	0.821	0.818
MRR@20	0.698	0.832	0.829

Performance Gains: Data augmentation improves NDCG@10 by 16.5% (from 0.723 to 0.842), demonstrating that enriched merchant profiles provide substantially better retrieval signals. Privacy rewriting causes only 0.36% average degradation, preserving 98.9% of the augmented performance.

Implications: This comprehensive evaluation confirms that our data processing pipeline successfully enhances retrieval quality through augmentation while maintaining performance through careful anonymization, ensuring the benchmark supports realistic and effective search scenarios.

A.6 Full Model Performance Table

For completeness, we provide the full per-metric performance comparison of all models, including L3 and L4 task breakdowns, in Table 16.

Overall best performer: Deepseek-V3.2 (w/ thinking) achieves the strongest overall profile on our automated metrics, registering the highest average Correctness (35.60%), the top Completeness score (77.56%), and the highest Search Efficiency (54.83%). These results indicate that the model is both effective at satisfying checklist constraints and relatively efficient at locating supporting evidence during the multi-hop process.

Thinking vs. non-thinking behavior: Enabling the chain-of-thought / “thinking” setting generally improves correctness and completeness for most models, often by a substantial margin (several percentage points). This improvement comes at the cost of slightly increased interaction cost: thinking variants typically use more tool calls and more rounds on average. The trade-off suggests that controlled deliberation helps planning and evidence aggregation, but requires careful engineering to avoid excessive search overhead.

Difficulty breakdown (L3 vs. L4): Performance is consistently higher on L3 tasks than on L4 tasks across models and metrics. For example, Deepseek-V3.2 (w/ thinking) shows a notable Correctness gap between L3 (38.70%) and L4 (28.90%). This confirms that L4 instances—characterized by personalized, hard-to-access dependencies and open-ended planning—pose substantially greater challenges for both retrieval and reasoning.

Trajectories and evidence quality: Trajectory-level metrics reveal complementary strengths and weaknesses. Some systems (e.g., LongCat variants) achieve very high Action Relevance and Causal Coherence, indicating strong procedural planning, while others

(including the top-ranked Deepseek) balance high Completeness and Faithfulness with strong Search Efficiency. Across the board, Evidence Sufficiency remains moderate (typically in the mid-40s to low-50s %), showing that even high-scoring models often fail to gather fully sufficient, verifiable supporting evidence for every checklist item.

Metric correlations and practical implications: Higher Completeness and Faithfulness tend to correlate with more rounds and tool calls, suggesting a trade-off between answer thoroughness and interaction cost. The L4 difficulty gap and the moderate Evidence Sufficiency scores point to two priority directions for improvement: (1) stronger, hop-aware retrieval/reranking to surface more task-relevant evidence early, and (2) improved multi-hop planning and tool-selection heuristics to avoid wasted queries and to better synthesize retrieved passages into verifiable checklist items.

These observations synthesize the per-model trade-offs visible in Table 16 and provide actionable guidance for improving agentic search pipelines on challenging, real-world multi-hop tasks.

Checklist derivation, normalization and scoring. To improve reproducibility and interpretability of the answer-level metrics (Correctness, Completeness, Faithfulness) reported in Table 16, we summarize how the multi-attribute constraint checklist is derived, normalized across tasks, and used for cross-task scoring.

Checklist derivation. For each instantiated multi-hop task we extract explicit hard constraints (e.g., city, district, service category, minimum rating, time window) and optional preferences (e.g., price preference). The checklist is composed of:

- **Hard constraints:** conditions that must be strictly satisfied for the instance to be counted as correct (e.g., “landmark = Xujiahui”, “rating ≥ 3.5 ”).
- **Soft constraints:** preferences that affect completeness or quality but do not by themselves invalidate correctness (e.g., “price range = mid”).
- **Evidence items:** minimal set of verifiable evidence pieces required to validate each checklist entry (see Evidence requirements).

Field normalization. To ensure consistent comparison across tasks and sources we normalize merchant fields before checking constraints:

- **Business hours:** normalized to 24h intervals (e.g., 08:00–22:00) and parsed to start/end timestamps for interval checks.
- **Price fields:** mapped to RMB numeric ranges or coarse levels (low/medium/high) with task templates defining comparison operators.
- **Address / administrative region:** standardized to a (city, district, business_area, landmark) quadruplet;

Conflict and missing-field handling. When automated checks encounter conflicts or missing values we apply deterministic rules:

- (1) If a hard constraint is violated or cannot be verified, the instance is marked as failing Correctness and flagged for manual review (used to calibrate annotators).
- (2) If a soft constraint is partially met, it receives partial credit (see scoring) rather than invalidating correctness.
- (3) If a field is missing but can be inferred reliably from other fields (e.g., city inferred from address), we use the inferred value and

record the inference source; otherwise the item is treated as “insufficient evidence”.

Evidence requirements. Each checklist item must be supported by at least one verifiable evidence item drawn from the recorded retrieval trajectory (LOCALRAG/Web search). Accepted evidence types:

- Merchant record fields with a stable identifier (poi_id).
- Extracted web snippets with timestamps.
- Retrieved RAG passages that explicitly contain the required attribute.

The minimal evidence set for a positive verification typically includes the merchant identifier plus the specific field value and its source (e.g., rating=4.2 from LOCALRAG: poi_bj_12345678).

Scoring and aggregation. We evaluate attributes with simple, transparent rules:

- Attributes are scored per-attribute as pass/fail (1/0).
- Dimension-level scores are normalized to percentages (0–100%) and combined across tasks using weighted averages (task weights reflect task difficulty and importance).
- Correctness is binary at the instance level: all key hard constraints must be satisfied to assign Correctness = 100%, otherwise Correctness = 0%.

Annotator calibration and quality control. To guarantee cross-annotator consistency:

- Annotators receive a detailed handbook with rules, examples, and exception handling.
- Each annotator must pass a calibration test on 100 gold-standard instances (pass threshold $\geq 90\%$) before labeling.
- Disagreements are resolved by majority vote or expert arbitration; we periodically re-check 5% of samples to detect drift.

Illustrative checklist examples. Representative task-level checklists and the expected evidence/answer format are provided to annotators (see Appendix A.5 for sample formats). Example templates include required fields, evidence snippets, and rationale fields to improve auditability and reproducibility.

B Dataset Construction and Task Design

B.1 LOCALSEARCHBENCH Task Details

This subsection describes the structure and design principles of the tasks included in LOCALSEARCHBENCH. Each task is a multi-hop question derived from a merchant record and (optionally) external web evidence; tasks typically require 3–5 reasoning hops to reach a complete, grounded answer. A single task entry consists of four components: (1) the original user query, (2) an explicit hop plan describing the required sub-queries, (3) retrieved evidence snippets used as grounding, and (4) a gold answer annotated with explicit citations to the source records or web pages.

Tasks are categorized by difficulty (L1–L5) according to ambiguity, hop count, and required reasoning steps. The dataset was constructed through an automated generation pipeline followed by a two-stage validation process combining model-based checks and human-in-the-loop verification; the annotation protocol, validation thresholds, and sampling procedures are described in Section 3.3.

These design choices aim to ensure that LOCALSEARCHBENCH covers a broad spectrum of real-world local life service scenarios while maintaining high-quality, verifiable multi-hop reasoning examples.

B.2 Data Validation Evaluation Frameworks

The *Data Validation Agent* performs a two-stage quality assessment using an LLM-as-a-Judge framework. The evaluation frameworks and validation thresholds are detailed in Table 3.

B.3 Multi-hop QA Examples

Figures 10–13 illustrate example multi-hop QA tasks (Hops 3–5, Difficulty: L4). Each figure shows the original user query, the hop-by-hop retrieved evidence and intermediate reasoning steps, and the final assembled answer with citations. Together, these examples demonstrate the system’s chain retrieval and evidence aggregation across regions and domains, and highlight requirements for reasoning about retrieval order, evidence reliability, and quantitative aggregation (e.g., budget calculations) when producing final responses.

[Hops: 3] | Difficulty: ★★★★ L4 - Planning Requirements]

Q: I want to plan a weekend afternoon citywalk route in *Xuhui District*. First, visit *historical buildings*, then find *a cafe* for afternoon tea, and finally go to *a park* for a walk and enjoy the night view. What are good recommended routes?

?	Round 1. Retrieve Xuhui District - Historical and Cultural Buildings
	Wukang Building History and Art Museum (1688 Middle Huaihai Road, French Renaissance style, rating 4.4, open all day for free);
	Wukang Road Historical and Cultural District (188 Wukang Road, rating 4.5, many historical buildings, free visit);
	Ke Ling Former Residence (168 West Fuxing Road, rating 4.6, celebrity former residence, 10:00-16:00);
	Xujiahui Source Cultural Scenic Area (258 Puxi Road, 4A scenic area, rating 4.4, 9:00-17:00 free)
	...
?	Round 2. Retrieve Xuhui District - Wukang Road - Afternoon Tea Cafe
	Coffee Tree Garden Western Restaurant (inside Wukang Court, 368 Wukang Road, rating 4.5, 10:00-22:00, afternoon tea set 158 yuan);
	% Arabica (2nd floor of Wukang Court, 368 Wukang Road, rating 4.3, specialty coffee internet celebrity shop);
	Pete's Coffee (2nd floor, 385 Wukang Road, rating 4.3, afternoon tea set 64 yuan), Plusone Coffee House (328 Wukang Road, rating 4.3, elegant environment)
	...
?	Round 3. Retrieve Xuhui District - Wukang Road - Park - Night View
	Binjiang Park (168 Xilin Road, rating 4.3, 3A scenic area, with “beautiful night view” tag, 6:00-22:00);
	Xuhui Green Park (88 Yueyang Road, rating 4.6, 5:00-22:00, complete lighting facilities);
	Huamei Green Park (1588 Huashan Road, rating 4.5, 6:00-22:00, fitness trails and lighting facilities)
	...

A: *Wukang Road Historical and Cultural District* (historical buildings) → *Wukang Court Cafe* (afternoon tea) → *Xuhui Green Park* (night view walk)

Figure 10: An Example of Multi-Hop QA – Hops: 3 | Difficulty: L4

Table 16: Performance comparison of different models on LOCALSEARCHBENCH (Max Round N = 5) with L3 and L4 task breakdowns. The top and worst performing results are highlighted in red (1st) and blue (bottom) backgrounds, respectively.

Models	Answer Evaluation Metrics												Trajectory Evaluation Metrics																				
	Avg. tool calls				Avg. rounds				Correctness (%)				Completeness (%)		Fluency (%)		Faithfulness (%)		Safety (%)		Action Relevance (%)		Evidence Sufficiency (%)		Causal Coherence (%)		Search Efficiency (%)						
	Avg.	L3	L4	Avg.	L3	L4	Avg.	L3	L4	Avg.	L3	L4	Avg.	L3	L4	Avg.	L3	L4	Avg.	L3	L4	Avg.	L3	L4	Avg.	L3	L4						
Qwen3-235B-A22B (w/o thinking)	2.00	2.04	1.92	2.93	2.97	2.84	21.20	24.90	13.20	50.94	51.14	50.54	69.16	69.53	68.42	25.28	27.57	20.70	79.72	80.00	79.16	75.22	77.57	70.52	43.63	44.43	41.97	50.68	52.51	47.02	45.99	47.04	43.89
Qwen3-235B-A22B (w/ thinking)	2.31	2.40	2.13	3.17	3.21	3.10	30.20	31.10	28.30	71.20	72.29	69.02	71.58	71.85	71.04	26.90	28.62	23.46	81.76	82.08	81.12	80.42	83.93	73.40	45.75	46.91	43.43	52.04	53.11	49.90	48.63	50.67	44.55
Qwen3-32B (w/o thinking)	2.78	2.82	2.70	3.11	3.09	3.16	19.80	22.00	15.10	40.96	40.98	40.92	68.50	68.81	67.88	21.38	22.90	18.34	80.76	80.85	80.58	74.67	80.41	63.19	46.52	47.79	43.98	48.82	50.71	45.04	49.84	53.28	42.96
Qwen3-32B (w/ thinking)	2.80	2.87	2.65	3.12	3.09	3.18	25.60	25.80	25.20	40.66	40.53	40.92	68.44	68.83	67.66	22.40	24.28	18.64	79.54	79.30	80.02	75.06	80.04	65.10	46.99	48.66	43.65	48.87	51.08	44.45	49.13	51.52	44.35
Qwen3-14B (w/o thinking)	2.53	2.64	2.30	2.07	2.02	2.17	24.20	25.20	22.20	40.60	41.26	39.28	69.62	69.65	69.56	27.44	29.59	23.14	80.74	80.47	81.28	70.38	84.50	70.38	46.44	48.00	43.32	50.96	53.20	46.48	50.02	53.28	43.50
Qwen3-14B (w/ thinking)	2.57	2.70	2.31	2.12	2.07	2.21	25.20	27.00	21.60	40.98	41.20	40.54	69.32	69.79	68.38	28.40	31.29	22.62	80.44	80.70	79.92	81.19	84.88	73.81	47.24	48.16	45.40	52.52	53.01	51.54	51.65	53.36	48.23
GPT-4.1	1.73	1.86	1.47	2.42	2.58	2.11	18.50	19.80	15.90	45.37	46.82	42.47	65.93	66.24	65.32	28.76	30.41	25.37	77.38	77.69	76.76	68.47	71.28	62.94	38.62	39.73	36.40	45.83	47.26	43.17	42.29	43.91	39.45
o3(high)	2.91	3.04	2.65	3.38	3.47	3.21	31.50	32.90	28.80	69.64	70.38	68.26	70.72	71.03	70.11	33.89	35.72	30.42	81.87	82.19	81.23	76.93	78.67	73.85	44.71	45.83	42.47	51.78	52.94	49.46	42.96	44.18	40.52
Gemini-2.5-Flash	1.84	1.99	1.52	2.51	2.65	2.22	21.00	23.80	15.40	58.44	59.71	55.90	68.04	67.68	68.76	35.72	39.85	27.46	79.70	80.06	78.98	70.58	73.86	64.02	41.61	42.86	39.11	48.94	50.09	46.64	46.91	49.24	42.25
Gemini-2.5-Pre	2.75	2.81	2.64	3.10	3.17	2.95	32.30	33.80	29.30	71.03	71.42	70.25	71.12	71.20	70.96	34.93	37.74	29.31	82.32	82.26	82.44	77.53	78.49	74.95	45.78	46.06	45.07	52.87	54.81	48.99	41.78	43.00	39.34
LongCat-Flash-Chat	2.34	2.36	2.30	3.07	3.03	3.15	25.30	25.60	24.60	52.98	55.76	47.42	69.45	69.41	69.53	27.49	29.63	23.21	83.61	83.55	83.73	77.78	81.48	70.38	47.33	47.87	46.25	50.86	53.00	46.58	52.29	55.00	46.87
LongCat-Flash-Thinking	3.04	3.10	3.08	3.20	3.18	3.24	30.70	33.90	23.80	68.83	69.54	67.41	69.07	69.31	68.59	31.47	33.83	26.75	80.10	80.62	79.06	78.50	82.46	70.58	47.37	48.69	44.73	53.18	54.45	50.64	53.27	55.06	49.69
GLM-4.6 (w/o thinking)	2.86	3.01	2.58	3.86	4.01	3.57	29.00	29.90	27.20	76.45	76.79	75.77	70.37	70.15	70.81	35.40	38.47	29.26	81.40	81.02	81.26	74.28	78.90	65.04	48.44	49.26	46.30	50.79	53.11	46.15	52.76	55.21	47.86
GLM-4.6 (w/ thinking)	3.08	3.13	2.97	4.06	3.97	3.86	32.80	33.90	30.60	76.83	77.59	75.31	70.27	70.33	70.15	37.48	39.27	33.90	81.30	81.20	81.50	77.66	80.41	72.16	48.90	49.33	52.67	54.18	49.65	44.43	49.98	53.33	
Deepseek-V3.2 (w/o thinking)	3.19	3.16	3.26	4.18	4.15	4.26	31.70	34.50	26.10	77.08	77.19	76.86	70.12	70.26	69.84	35.96	37.68	32.52	81.52	81.61	81.34	75.51	79.80	66.93	48.49	49.41	46.65	52.23	54.76	47.17	54.33	57.02	48.95
Deepseek-V3.2 (w/ thinking)	3.21	3.16	3.31	4.20	4.15	4.30	35.60	38.70	28.90	77.56	77.86	76.96	70.92	70.59	71.58	39.78	38.27	42.80	81.13	81.65	80.09	75.58	79.94	66.86	48.86	49.74	47.10	52.62	54.28	49.30	54.83	57.11	49.67

[Hops: 4] | Difficulty: ★★★★ L4 - Planning Requirements & Cross-Region Search]

Q: I'm currently in Shanghai and plan to travel to Suzhou. I need a *travel agency to book train tickets* for me. When I arrive in Suzhou, I want to *charter a car* to visit the *Humble Administrator's Garden* and *Tiger Hill*, and *watch a play* in the evening. What are the charter car prices and times of this travel agency? Can they recommend the *highest-rated theater* in Suzhou Gusu District?

Round 1. Retrieve Shanghai - Train ticket booking - Travel agency
Changyou Tongda International Travel Agency (Room 902, Building 7, 28 Tianai Road, Hongkou District, rating 3.9, 10 reviews, provides train ticket booking service, service fee ¥5, business hours Monday to Sunday 9:00-18:00);
Shanghai International Travel Agency (28 Jinsha River Road-Gate 5, Putuo District, rating 4.2, 86 reviews, provides domestic and international travel route planning, visa processing, flight and hotel booking and other one-stop travel services, business hours Monday to Friday 09:00-18:00, Saturday to Sunday 09:00-17:00);

 Round 2. Retrieve Suzhou - Charter car service
Suzhou Travel Charter Car Service Center (28 Guanqian Street, Gusu District, rating 4.3, 5-seater economy sedan ¥280/day, 7-seater business van ¥450/day, provides route planning for Humble Administrator's Garden, Tiger Hill, etc.);
Suzhou Ancient City Charter Car Service (88 Jinji Lake Road, Industrial Park, rating 4.1, 5-seater sedan ¥260/day, 7-seater business van ¥420/day);
Suzhou Ancient City Charter Car Service (368 Pingjiang Road, Pingjiang District, rating 4.0, 5-seater economy ¥250/day, 7-seater business van ¥400/day);

 Round 3. Retrieve Suzhou - Attraction tickets - Service provider
Suzhou Tourism Ticket Center (Shanghai Tower, 568 Lujiazui Ring Road, rating 5.0, per capita ¥30, open 7:00-22:00, specialty coffee brand, soy milk latte ¥28/iced Americano ¥22);
Suzhou Attractions Pass Service (99 Xiangtai Avenue, Industrial Park, rating 4.2, provides multi-attraction package service, Humble Administrator's Garden + Tiger Hill combo ticket ¥135);
Suzhou Ancient City Tourism Service (268 Pingjiang Road, Pingjiang District, rating 4.1, provides attraction tickets and tour guide services);

 Round 4. Retrieve Suzhou - Gusu District - Theater play
Suzhou Cultural Arts Center Grand Theater (1 Guanqian Street, rating 4.8, play "Secret Love in Peach Blossom Land" from ¥180, Kunqu Opera "The Peony Pavilion" from ¥280, business hours Monday to Sunday 10:00-21:00);
Suzhou Poly Grand Theater (2075 Renmin Road, rating 4.7, play "Thunderstorm" from ¥160, musical from ¥250, business hours Monday to Sunday 9:30-21:00);
Pingjiang Art Theater (368 Pingjiang Road, rating 4.5, Kunqu Opera performance from ¥120, play "Teahouse" from ¥150, business hours Monday to Sunday 10:00-20:30);

A: **Shanghai International Travel Agency + Suzhou Ancient City Charter Car Service** 5-seater economy sedan ¥250/day, 7-seater business van ¥400/day + **Qianxiangyi Crossstalk Teahouse** ¥120 + **Suzhou Attractions Pass Service** Humble Administrator's Garden + Tiger Hill combo ticket ¥135 + **Suzhou Cultural Arts Center Grand Theater** rating 4.8.

Table 17: Human-AI agreement rate by field.

Field	Human-AI Agreement Rate
subcategory	100.00%
district	100.00%
category	100.00%
city	100.00%
review_count	96.97%
price_range	95.96%
business_hours	95.96%
landmark	92.93%
rating	91.92%
delivery_available	91.82%
delivery_range	91.82%
delivery_fee	87.78%
min_order_amount	86.77%
address	78.79%
phone	74.75%
name	73.74%
mobile	72.72%
products	72.72%
email	69.70%
description	63.64%
Average	88.34%

Figure 11: An Example of Multi-Hop QA – Hops: 4 | Difficulty: L4

Table 18: Evaluation Frameworks for Data Quality Assessment

(a) Augmented Data Quality		
Dimension	Weight	Score
Data Completeness	20%	0.8605
Logical Consistency	30%	0.9000
Geographic Accuracy	10%	0.8568
Content Quality	20%	0.8521
Format Standardization	10%	0.9060
Innovation Degree	10%	0.7080
Overall	100%	0.8596

(b) Privacy Rewriting Data Quality		
Dimension	Weight	Score
Anonymization Degree	50%	0.9473
Logical Consistency	20%	0.8383
Content Consistency	20%	0.9919
Innovation Degree	10%	0.8200
Overall	100%	0.9217

[Hops: 5] | Difficulty: ★★★★ L4 - Planning Requirements |

?

Q: I want to arrange a cultural weekend in *Pudong New District*. Watch a *play* on Saturday night at 7 pm, then have a *late-night snack* nearby (around 150 yuan per capita). On Sunday morning, go to *an artsy cafe* (50 yuan per capita), watch a *crosstalk show* at 2 pm in the afternoon, then buy some *Shanghai specialty souvenirs* (budget 200 yuan). What are the recommendations? What is the total budget?

- ➊ Round 1. Retrieve *Pudong New District - Play - Saturday night*
 - Shanghai Drama Theatre Xincheng Theatre* (368 Dongxin Street, Lujiazui, New City Square, open Saturday 09:00-23:00, play "Thunderstorm" ¥180, rating 4.5, landmark Shanghai Tower);
 - Global Arts Theatre* (288 Pudong Avenue, Lujiazui, east side of International Convention Center, open Saturday 9:00-23:00, play "Thunderstorm" ¥280, rating 4.0);
 - Shanghai Drama Theatre* (88 Century Avenue, Lujiazui, open Monday to Sunday 10:00-22:00, play "Thunderstorm" ¥280, rating 5.0)
 - ...
- ➋ Round 2. Retrieve *Pudong New District - Lujiazui - Late-night snack restaurant - 150 yuan per capita*
 - Baiwei Xiang BBQ* (about 200 meters east of Jinmao Tower, Lujiazui, rating 4.2, per capita ¥85, open until 23:00, 20% off during late-night hours, charcoal BBQ + beer, suitable for gathering after the show);
 - Lanxin Restaurant* (7th floor of Lujiazui Center, 866 Pudong South Road, Lujiazui, rating 4.5, per capita ¥138, open until 21:00, Shanghai local cuisine with exquisite dishes, but closes early);
 - LOKAL by Wagas Light Food* (2nd floor Building 2, Super Brand Mall, Lujiazui, rating 4.7, per capita ¥91, open until 22:00, Western simple meals)
 - ...
- ➌ Round 3. Retrieve *Pudong New District - Lujiazui - Artsy cafe - 50 yuan per capita*
 - KUDDO CAFE* (Shanghai Tower, 568 Lujiazui Ring Road, rating 5.0, per capita ¥30, open 7:00-22:00, specialty coffee brand, soy milk latte ¥28/iced Americano ¥22);
 - COSTA CAFE* (2nd floor, 268 Middle Yincheng Road, Lujiazui, rating 4.5, per capita ¥31, open 7:30-21:00, British-style coffee, Americano ¥25/latte ¥32);
 - Xingyi Coffee* (L1 floor, 88 Lujiazui Financial District, Global Center, rating 4.0, per capita ¥37, open 7:00-22:00, chain brand, Americano ¥28)
 - ...
- ➍ Round 4. Retrieve *Pudong New District - Lujiazui - Crosstalk - Afternoon show*
 - Qianxiangyi Crosstalk Teahouse* (4th floor of Lujiazui Shopping Center, 36 Dafeng Road, Lujiazui, rating 5.0, per capita ¥120, open 10:00-22:00, crosstalk performance ticket ¥120/VIP box ¥200);
 - Le Zai Moda Crosstalk Society* (L2 floor of Century Hui Plaza, 328 Pudong Avenue, Lujiazui, rating 4.5, per capita ¥180, open Saturday and Sunday 11:00-23:00, student ticket 20% off);
 - Xingyi Crosstalk Pavilion* (3rd floor of Wuyue Plaza, 26 Yingbin Avenue, Lujiazui, rating 4.2, per capita ¥120, open Saturday and Sunday 10:00-23:00, regular ticket ¥80/VIP ticket ¥150)
 - ...
- ➎ Round 5. Retrieve *Pudong New District - Lujiazui - Shanghai specialty souvenirs*
 - Shanghai Butterfly Pastry* (88 Fucheng Road, near Aurora Museum, Lujiazui, rating 5.0, per capita ¥150, open 9:00-21:00, butterfly pastry gift 96g/box ¥52.8, Garden Hotel butter cookies gift box ¥226, Laoxiangzhai pastry gift box ¥100.91);
 - Dinah's Gift Workshop* (668 Binjiang Avenue near Oriental Pearl, Lujiazui, rating 4.7, per capita ¥145, open 10:00-22:00, L'Occitane scented candle gift box ¥158, preserved flower aromatherapy set ¥58.9);
 - Shanghai Tourism Souvenir Store* (near Oriental Pearl, 88 Century Avenue, Lujiazui, rating 4.0, per capita ¥161, open 9:30-21:30, Oriental Pearl Tower model ¥128, old Shanghai pocket watch ¥299, postcard set ¥35)
 - ...

✓ A: *Shanghai Drama Theatre Xincheng Theatre* "Thunderstorm" ¥180 + *Baiwei Xiang BBQ* late-night snack ¥150 + *KUDDO CAFE* coffee ¥30 + *Qianxiangyi Crosstalk Teahouse* ¥120 + *Shanghai Butterfly Pastry* souvenirs ¥154, total budget about ¥634 /person, entire route in Pudong New District Lujiazui area.

Figure 12: An Example of Multi-Hop QA – Hops: 5 | Difficulty: L4

[Hops: 4] | Difficulty: ★★★★ L4 - Planning Requirements |



Q: We plan to watch the *KPL 2025 Grand Finals Championship Match* on-site, we'll have dinner nearby, then visit a park, and finally find a place to stay nearby. Do you have any recommendations?

Round 1. Query Where and when is the KPL 2025 Grand Finals Championship Match held

- i November 8, 5:30 PM at National Stadium (Bird's Nest), Chaoyang District, Beijing.**

Round 2. Retrieve National Stadium - Restaurant

- G Guofeng Jiayan Restaurant** (5th floor of Bird's Nest Cultural Center, 3 National Stadium South Road, rating 4.5, high-end Huaiyang cuisine, per capita ¥259, business hours 10:30-22:00);
- Jiayan Guoao Restaurant** (3rd floor of Huachuang Life Plaza, Lin Cui Xi Li, rating 4.8, Huaiyang and Beijing-Shandong cuisine, signature fruitwood roasted duck, per capita ¥128, business hours 10:00-22:00);
- ...

Round 3. Retrieve National Stadium - Park

- Olympic Central Park** (88 Beichen East Road, Chaoyang District, rating 4.5, free admission, open 06:00-22:00);
- Olympic Ecological Forest Park** (28 Beichen East Road, rating 4.6, Beijing's largest urban ecological park, free admission, open 06:00-21:00);
- Olympic Park Central Area Landscape Corridor** (intersection of North Central Axis Landscape Avenue and National Stadium South Road, Chaoyang District, rating 3.5, free admission, open 06:00-22:00);
- ...

Round 4. Retrieve National Stadium - Hotel

- Anyuan Yasha Boutique Homestay** (Building 3, North Area, Anyuan Dongli Second District, rating 4.5, boutique homestay near Bird's Nest);
- Bird's Nest Water Cube Boutique Inn** (88 Beichen West Road, Yayuncun Street, rating 5.0, high-quality inn near Olympic Park);
- Beijing Guoyu Business Hotel** (28 Huizhong Road, Huizhongli, rating 4.0, business hotel in Yayuncun area, price range ¥400-800, avg ¥533);
- ...

A: Guofeng Jiayan/Jiayan Guoao Restaurant → Olympic Central/Olympic Ecological Forest Park/ Olympic Park Central Area Landscape Corridor → Anyuan Yasha Boutique Homestay/ Bird's Nest Water Cube Boutique Inn/ Beijing Guoyu Business Hotel

Figure 13: An Example of Multi-Hop QA – Hops: 4 | Difficulty: L4

Prompt for Data Augmentation

Please generate enhanced merchant data based on the original data below. Enrich all fields while preserving original information:

IMPORTANT: category must be selected from this list only:
`,'join(allowed_categories)}

Original data:
`json.dumps(original, ensure_ascii=False, indent=2)}

Output complete JSON with these fields:

Basic Info:

- name: Concise name without location identifiers (e.g., "XX Road Branch")
- category: Main category (MUST choose from: `,'join(allowed_categories))
- subcategory: Sub-category (refined classification matching actual business)
- description: Brief description (<50 chars), matching store type and name

Location (enhance based on original address):

- address: Detailed address (if original " address " is non-empty, keep identical; if empty, generate from location field, consistent with city/district/landmark)
- city: Shanghai
- district: District name (e.g., Huangpu/Jing'an/Xuhui)
- business_area: Business district (e.g., Xujiahui/Lujiazui)
- landmark: Nearby landmark (e.g., Shanghai Stadium)

Business Info:

- business_hours: Weekly hours (reasonable format; 24-hour stores must be "0:00-24:00")

Pricing & Rating:

- price_range: Reasonable range between min/max product prices
- avg_price: Average price within price_range and product price bounds
- rating: 1-5 score (keep original if exists)
- review_count: Review count (keep original if exists; otherwise ≤10000)

Services:

- tags: 5-8 relevant tags matching business type
- facilities: 5-10 items (physical + service facilities like "Free Wi-Fi")

Contact:

- phone: Landline (7-8 digits)
- mobile: Mobile (11 digits)
- email: Valid format (xxx@xxx.com)

Coordinates:

- latitude: 6 decimal places
- longitude: 6 decimal places

(Infer from city/district/business_area)

Delivery:

- delivery_available: true if original " delivery_available " non-empty, else false
- delivery_range: 1.0-25.0 km
- delivery_fee: <60 yuan, match original
- min_order_amount: 2-60 yuan, match original

Promotions:

- promotions: 0-3 items

Products (same count as original):

- products: List with name/price/description per item

Group Deals (same count as original):

- group_deals: List with title/price/description/discount

(If matching products exist, deal price must be lower; discount = deal_price/original_price)

Requirements:

1. Maintain data authenticity and consistency
2. Generate appropriate features/products by business type
3. Reasonable market-level pricing
4. All fields consistent with original data
5. Return JSON only, no extra text

Figure 14: Prompt for Data Augmentation

Prompt for Privacy Rewriting	Prompt for Augmentation Validation
<p>Anonymize merchant data. Protect privacy, keep usability.</p> <pre>Original: {json.dumps(enhanced, ensure_ascii=False, indent=2)}</pre> <p>Anonymize these fields (different from original):</p> <ul style="list-style-type: none"> - name: Fully anonymized, no real duplicates/locations, commercially reasonable - description: Match new name - address: Change building#/floor, keep city/district - phone: Valid 7-8 digits - mobile: Valid 11 digits - email: Based on new name <p>Keep all other fields identical.</p> <p>Return JSON only, no extra text.</p>	<p># Role You are a professional search benchmark evaluator, evaluating merchant data quality. Compare original and generated data for each field, score strictly according to JSON format.</p> <p># Task Evaluate accuracy, consistency and reasonableness of enhanced merchant data based on original data.</p> <p># Input Format - **Original Data**: Basic merchant info from Meituan search results - **Generated Data**: Enhanced merchant data with enriched fields</p> <p># Evaluation Criteria Score 1 for correct, 0 for incorrect. Only evaluate the following fields:</p> <p># Basic Info name: 1=Concise name without location identifiers (e.g., "XX Road Branch", "XX Store"); 0=Contains location info or unreasonable category: 1=One of 6 categories (Food/Shopping/Hotel/Tourism/Medical/Life Service/); 0=Invalid category subcategory: 1=Refined classification under main category, matches actual business; 0=Mismatch description: 1=Matches store type and name, natural language; 0=Inconsistent or unreasonable</p> <p># Location Info city: 1=Consistent with location field in original data; 0=Inconsistent district: 1=Consistent with location field in original data; 0=Inconsistent landmark: 1=Reasonable and matches business_area; 0=Mismatch address: 1=Identical to original if non-empty, or generated from location field matching city/district/landmark; 0=Inconsistent business_area: 1=Consistent with location field in original data; 0=Inconsistent</p> <p># Price & Business Info price_range: 1=Reasonable, within min/max product prices, matches merchant type; 0=Unreasonable (e.g., lunch box over 80, bubble tea over 60) avg_price: 1=Within price_range and product price bounds, reasonable; 0=Inconsistent business_hours: 1=Reasonable format, matches merchant type (24h stores must be "0:00-24:00"); 0=Invalid</p> <p># Rating Info rating: 1=Matches original if exists, or within 1-5 range; 0=Invalid review_count: 1=Matches original if exists, or under 10000; 0=Exceeds limit</p> <p># Tags & Features tags: 1=Reasonable, matches merchant type; 0=Mismatch specialties: 1=Matches actual business; 0=Mismatch facilities: 1=Reasonable, includes physical and service facilities; 0=Mismatch</p> <p># Contact Info phone: 1=Valid format (7-8 digits); 0=Invalid mobile: 1=Valid format (11 digits); 0=Invalid email: 1=Valid format (xxx@xxx.com); 0=Invalid</p> <p># Delivery Info delivery_available: 1=true if original delivery field non-empty, else false; 0=Inconsistent delivery_fee: 1=Under 60, matches original, reasonable; 0=Unreasonable delivery_range: 1=1.0-25.0 km; 0=Unreasonable min_order_amount: 1=2-60, matches original, reasonable; 0=Unreasonable</p> <p># Promotions promotions: 1=Reasonable content; 0=Unreasonable</p> <p># Products products: 1=Matches store type, reasonable prices, accurate descriptions; 0=Mismatch group_deals: 1=If matching products exist, deal price < product price AND discount = deal_price/original_price; otherwise reasonable; 0=Invalid calculation</p> <p># Output Format ```json</p> <pre>{"field_name": { "score": 0 or 1, "reason": "Evaluation reason" }, ...}</pre> <p>Note: Output score and reason only, no value field.</p>

Figure 15: Prompt for Privacy Rewriting

Role
You are a professional search benchmark evaluator, evaluating merchant data quality. Compare original and generated data for each field, score strictly according to JSON format.

Task
Evaluate accuracy, consistency and reasonableness of enhanced merchant data based on original data.

Input Format
- **Original Data**: Basic merchant info from Meituan search results
- **Generated Data**: Enhanced merchant data with enriched fields

Evaluation Criteria
Score 1 for correct, 0 for incorrect. Only evaluate the following fields:

Basic Info
name: 1=Concise name without location identifiers (e.g., "XX Road Branch", "XX Store"); 0=Contains location info or unreasonable
category: 1=One of 6 categories (Food/Shopping/Hotel/Tourism/Medical/Life Service/); 0=Invalid category
subcategory: 1=Refined classification under main category, matches actual business; 0=Mismatch
description: 1=Matches store type and name, natural language; 0=Inconsistent or unreasonable

Location Info
city: 1=Consistent with location field in original data; 0=Inconsistent
district: 1=Consistent with location field in original data; 0=Inconsistent
landmark: 1=Reasonable and matches business_area; 0=Mismatch
address: 1=Identical to original if non-empty, or generated from location field matching city/district/landmark; 0=Inconsistent
business_area: 1=Consistent with location field in original data; 0=Inconsistent

Price & Business Info
price_range: 1=Reasonable, within min/max product prices, matches merchant type; 0=Unreasonable (e.g., lunch box over 80, bubble tea over 60)
avg_price: 1=Within price_range and product price bounds, reasonable; 0=Inconsistent
business_hours: 1=Reasonable format, matches merchant type (24h stores must be "0:00-24:00"); 0=Invalid

Rating Info
rating: 1=Matches original if exists, or within 1-5 range; 0=Invalid
review_count: 1=Matches original if exists, or under 10000; 0=Exceeds limit

Tags & Features
tags: 1=Reasonable, matches merchant type; 0=Mismatch
specialties: 1=Matches actual business; 0=Mismatch
facilities: 1=Reasonable, includes physical and service facilities; 0=Mismatch

Contact Info
phone: 1=Valid format (7-8 digits); 0=Invalid
mobile: 1=Valid format (11 digits); 0=Invalid
email: 1=Valid format (xxx@xxx.com); 0=Invalid

Delivery Info
delivery_available: 1=true if original delivery field non-empty, else false; 0=Inconsistent
delivery_fee: 1=Under 60, matches original, reasonable; 0=Unreasonable
delivery_range: 1=1.0-25.0 km; 0=Unreasonable
min_order_amount: 1=2-60, matches original, reasonable; 0=Unreasonable

Promotions
promotions: 1=Reasonable content; 0=Unreasonable

Products
products: 1=Matches store type, reasonable prices, accurate descriptions; 0=Mismatch
group_deals: 1=If matching products exist, deal price < product price AND discount = deal_price/original_price; otherwise reasonable; 0=Invalid calculation

Output Format
```json

```
{"field_name": {
 "score": 0 or 1,
 "reason": "Evaluation reason"
}, ...}
```

Note: Output score and reason only, no value field.

**Figure 16: Prompt for Augmentation Validation**

**Algorithm 1:** Algorithm of Iterative Search

---

**Input:** User query  $Q$ , Maximum rounds  $N$   
**Output:** Retrieved information set  $\mathcal{I}$

// Phase 1: Initialization

```

1 information $\leftarrow \emptyset$; // Initialize empty information set
2 round $\leftarrow 1$; // Initialize iteration counter

```

// Phase 2: Iterative Retrieval Loop

```

3 for $i \leftarrow 1$ to N do
4 web_results $\leftarrow \emptyset$, rag_results $\leftarrow \emptyset$;
 // LLM Response Parsing
5 response $\leftarrow \text{LLM}(Q, \text{information})$;
6 web_query $\leftarrow \text{ExtractTag}(\text{response}, \langle \text{web_search} \rangle)$;
7 rag_query $\leftarrow \text{ExtractTag}(\text{response}, \langle \text{rag} \rangle)$;
 // Web Search (at most once per round)
8 if web_query $\neq \emptyset$ then
9 web_results $\leftarrow \text{WebSearch}(\text{web_query})$;
 // Query temporal facts, weather, news, events
 // LocalRAG Retrieval (at most once per round)
10 if rag_query $\neq \emptyset$ then
11 rag_results $\leftarrow \text{LocalRAGSearch}(\text{rag_query})$;
 // Semantic similarity + geographical proximity analysis
 // Retrieve merchants, POIs, service descriptions
12 information $\leftarrow \text{information} \cup \text{web_results} \cup \text{rag_results}$;
13 if $\text{SufficientInfo}(\text{information}, Q)$ or $i = N$ then
14 break;

```

// Phase 3: Return Results

```

15 return information;

```

---