BrowseComp-ZH: Benchmarking Web Browsing Ability of Large Language Models in Chinese

Peilin Zhou[1†*] **Bruce Leon**[2†] **Xiang Ying**[3] **Can Zhang**[2] **Yifan Shao** **Qichen Ye**[4]
Dading Chong[2] **Zhiling Jin**[5] **Chenxuan Xie**[6] **Meng Cao**[7] **Yuxin Gu**[8]
Sixin Hong[2] **Jing Ren**[2] **Jian Chen**[1,9] **Chao Liu**[2] **Yining Hua**[10]

[1]Hong Kong University of Science and Technology (Guangzhou)
[2]Peking University [3]Mindverse AI [4]Alibaba Group [5]Zhejiang University
[6]Zhejiang University of Technology [7]MBZUAI [8]NIO [9]HSBC
[10]Harvard T.H. Chan School of Public Health
{zhoupalin, yifashao209, hello.crabboss, mengcaopku}@gmail.com
yingxiang@mindverse.ai, hongsixin1995hk}@gmail.com
{zhangcan, yeeeqichen, 1601213984, 1901210484}@pku.edu.cn
22415038@zju.edu.cn, jchen524@connect.hkust-gz.edu.cn
chaoliu@pku.org.cn, Yininghua@g.harvard.edu, aaron.gu1@nio.com

Abstract

As large language models (LLMs) evolve into tool-using agents, the ability to browse the web in real-time has become a critical yardstick for measuring their reasoning and retrieval competence. Existing benchmarks such as BrowseComp concentrate on English and overlook the linguistic, infrastructural, and censorship-related complexities of other major information ecosystems—most notably Chinese. To address this gap, we introduce **BrowseComp-ZH**, a high-difficulty benchmark purpose-built to comprehensively evaluate LLM agents on the Chinese web. BrowseComp-ZH consists of 289 multi-hop questions spanning 11 diverse domains. Each question is reverse-engineered from a short, objective, and easily verifiable answer (e.g., a date, number, or proper noun). A two-stage quality control protocol is applied to strive for high question difficulty and answer uniqueness.

We benchmark over 20 state-of-the-art language models and agentic search systems on our proposed BrowseComp-ZH. Despite their strong conversational and retrieval capabilities, most models struggle severely: a large number achieve accuracy rates below 10%, and only a handful exceed 20%. Even the best-performing system, OpenAI's DeepResearch, reaches just 42.9%. These results demonstrate the considerable difficulty of BrowseComp-ZH, where success demands not only effective retrieval strategies, but also sophisticated reasoning and information reconciliation—capabilities that current models still struggle to master.

Our dataset, construction guidelines, and benchmark results have been publicly released at `https://github.com/PALIN2018/BrowseComp-ZH`.

1 Introduction

As large language models (LLMs) evolve from static knowledge repositories to dynamic agents capable of using external tools, tasks that involve web browsing have emerged as a critical lens through which to evaluate their real-world reasoning and information-seeking capabilities Xi et al. (2025). By interacting with search engines and navigating live web content, LLMs can augment

*Corresponding †Co-first authors

Preprint. Under review.

their internal knowledge with up-to-date external evidence, retrieve context-specific information, and perform multi-hop reasoning across heterogeneous sources. This browsing ability extends the temporal scope of LLMs while enabling them to tackle questions that lie beyond the reach of pretraining, such as time-sensitive facts or obscure entity relations that require targeted retrieval.

While an increasing number of studies Li et al. (2023a); Fernández-Pichel et al. (2024); Fan et al. (2024); Vu et al. (2023); Lai et al. (2025); He et al. (2024); Lai et al. (2024); Xiong et al. (2024) demonstrate that web browsing greatly improves LLM performance on downstream tasks, there remains a surprising lack of direct evaluation of browsing capabilities themselves—i.e., the ability of LLMs to effectively retrieve, filter, and reason over information from the web. This evaluation is crucial for assessing the true web-browsing competence of LLMs and understanding their potential to tackle real-world tasks that require dynamic information retrieval. To address this gap, Wei et al. (2025) introduced a benchmark of reverse-designed, evidence-grounded queries that challenge English-language agents to search and reason over difficult-to-access information.

Nonetheless, Wei et al. (2025) primarily operates within the English-language web, missing the linguistic, structural, and cultural complexities inherent to other language environments. In particular, the Chinese web poses unique challenges for retrieval and reasoning: information is scattered across heterogeneous platforms (e.g., Baidu Baike, Zhihu, government portals), naming conventions are inconsistent, and search engines often fail to reliably index deep or domain-specific content. Moreover, the linguistic properties of Chinese, such as implicit referents, idiomatic expressions, and context-dependent syntax, frequently break standard keyword-based retrieval paths.

Crucially, **directly translating English browsing benchmarks into Chinese does not yield a meaningful evaluation framework**. This approach fails on several fronts: structural and idiomatic differences render many translated queries unnatural or ineffective in actual search; canonical information pathways in English (e.g., Wikipedia, IMDb) lack equivalents or exhibit vastly different structures in the Chinese web; and translated queries often collapse into keyword matches, trivializing the intended challenge. Moreover, the Chinese web presents its own set of retrieval obstacles—fragmented and platform-specific content, inconsistent naming conventions, and linguistic traits such as ellipsis, cultural references, and implicit reasoning that undermine linear search strategies. As a result, existing English-centric benchmarks struggle to generalize, leaving a critical gap in assessing real-world browsing capabilities in non-English environments. We argue that **web-based benchmarks must be natively constructed within the Chinese information ecosystem**, where the search logic, content structure, and linguistic context authentically reflect the challenges faced by agents operating in Chinese-language settings.

话题：艺术 **Topic:** Art
问题：在中国传统艺术中，有一种特殊的绘画形式，起源于元代，盛行于清末，传说是由一位古代知名画家在酒后兴起所创。这种艺术形式在2010~2015年之间被列入某省级非物质文化遗产名录。绘制这种艺术形式需要画家精通各种画法，并且要善书各种字体。请问这种艺术形式是什么？
Question: In traditional Chinese art, there is a unique form of painting that originated in the Yuan Dynasty and became popular during the late Qing Dynasty. It is said to have been created by a famous ancient painter who was inspired by alcohol. Between 2010 and 2015, this art form was included in the provincial intangible cultural heritage list of a certain region. To paint in this style, artists must be proficient in various painting techniques and skilled in writing different types of calligraphy. What is this art form called?
答案：锦灰堆 **Answer:** Heaps of Brocade and Ash

话题：影视 **Topic:** Film & TV
问题：某知名电视剧，女二号（演员）在1993年进入演艺圈。女一号（演员）的现任丈夫是浙江湖州人。男一号（演员）6年后登上了春晚舞台。问该电视剧是什么？
Question: In a well-known TV drama, the second female lead (actress) entered the entertainment industry in 1993. The current husband of the first female lead (actress) is from Huzhou, Zhejiang. The first male lead (actor) performed on the CCTV Spring Festival Gala six years later. What is the name of this TV drama?
答案：父母爱情 **Answer:** Love of Parents

Figure 1: Two data samples from BrowseComp-ZH with their English translations.

To fill this gap, we introduce **BrowseComp-ZH**, the first benchmark specifically designed to evaluate web-enabled LLMs in the Chinese information environment. Mirroring the philosophy of the original BrowseComp, each item is created by reverse design: expert annotators (all holding at least a master's

degree and possessing both LLM expertise and topical domain knowledge) begin with a single, factual answer and craft a multi-constraint query that is hard to retrieve yet trivially verifiable. BrowseComp-ZH extends the original framework in two crucial directions: (1) Native Chinese construction. All queries, evidence chains, and browsing steps are authored in Chinese and fully localized. Every candidate query is verified on three widely-used search engines: Baidu[1], Bing[2], and Google[3], and is accepted only if the answer is not surfaced in the first-page results of any engine. (2) Two-stage quality control. Stage 1 screens for keyword retrievability as above. Stage 2 focuses on ensuring the uniqueness of the answers. We employ a human-in-the-loop approach, where multiple top-performing AI agents first perform reasoning and provide answers based on the designed questions. These results are then manually verified by annotators to check if any additional answers satisfy the constraints of the question.

The final **BrowseComp-ZH** dataset consists of 289 complex questions, each with multiple constraints and unique answers, spanning 11 diverse domains, including film, technology, law, and medicine. These questions require multi-step reasoning and are difficult to answer directly through search engines, making them ideal for evaluating the ability of Chinese LLM agents to conduct multi-hop retrieval, perform factual reasoning, and integrate online information. Fig. 1 illustrates two representative example drawn from BrowseComp-ZH. Based on this dataset, we benchmark the performance of **more than 20 systems**, encompassing open-source models (e.g., DeepSeek R1 Guo et al. (2025), Qwen-2.5-72B-Instruct Yang et al. (2024)), closed-source APIs (e.g., GPT-4o OpenAI (2024a), Claude-3.7 Anthropic (2025), Gemini-2.5-Pro Google (2024a)), as well as AI search agents (e.g., DeepResearch OpenAI (2025a), DeepSeek DeepSeek (2025), Perplexity Perplexity (2025), and Doubao ByteDance (2025)).

Our evaluation offers a nuanced characterization of model capabilities across different system designs. Based on our analysis, we highlight several key findings:

1. Naive large language models, irrespective of parameter scale or training data size, exhibit consistently poor performance on the benchmark, with accuracies often remaining below 10%. Representative examples include Qwen2.5-72B-Instruct, Llama 4, and GPT-4o.
2. Models endowed with inherent reasoning capabilities exhibit consistent improvements in accuracy. For example, DeepSeek-R1 surpasses DeepSeek-V3 by 14.5%, while Claude-3.7-Sonnet outperforms Claude-3.5 by 12.2%.
3. Agentic systems augmented with external retrieval mechanisms tend to achieve comparatively higher scores. Notably, OpenAI's DeepResearch and Doubao (Deep Search) demonstrate that well-orchestrated retrieval and reasoning pipelines can substantially enhance performance, achieving accuracies of 42.9% and 26.0%, respectively.
4. While well-designed retrieval pipelines can significantly improve performance, not all systems benefit from retrieval integration. For instance, enabling web search for DeepSeek-R1 results in a substantial decline in performance, with accuracy dropping from 23.2% in the direct-answer (no web access) setting to 7.6% when web search is enabled.

These findings suggest that BrowseComp-ZH not only challenges models' browsing and reasoning capabilities but also exposes their limitations in processing, evaluating, and aligning retrieved information with internal representations. Detailed performance statistics across model categories are summarized in Table 1.

2 Related Work

With the increasing ability of large language models (LLMs) to use external tools, recent research has focused on evaluating their capacity to retrieve and reason over real-world information. Representative works such as WebGPT Nakano et al. (2021), Toolformer Schick et al. (2023), and ReAct Yao et al. (2023) explore how LLMs leverage search engines, tool usage, and reasoning strategies to tackle complex question-answering tasks. In parallel, retrieval-augmented generation (RAG) frameworks Guu et al. (2020); Lewis et al. (2020) have been widely adopted in QA Wiratunga et al.

[1] `https://www.baidu.com`

[2] `https://cn.bing.com`

[3] `https://www.google.com`

(2024), summarization Edge et al. (2024), and fact verification Martin et al. (2024), serving as a mechanism for injecting external knowledge into LLMs.

To assess retrieval capabilities, a variety of widely used English benchmarks have been proposed, including TriviaQA Joshi et al. (2017), HotpotQA Yang et al. (2018), FEVER Thorne et al. (2018), KILT Petroni et al. (2020), and GAIA Mialon et al. (2023). These datasets cover multi-hop reasoning, knowledge-intensive QA, and fact checking, typically relying on structured sources like Wikipedia and StackExchange. However, since many answers can be retrieved via simple keyword searches, these benchmarks often fail to evaluate an agent's ability to plan complex search trajectories and synthesize information across documents. Wei et al. (2025) addresses this by reverse-designing queries from known answers with multiple retrieval constraints, requiring multi-hop search and cross-page reasoning. While it offers finer-grained evaluation for web browsing agents, it remains confined to the English web and lacks generalizability to non-English environments with fragmented platforms and diverse linguistic structures.

However, extending such evaluations to Chinese poses unique challenges. Several retrieval-related datasets have emerged for the Chinese web, but each presents notable limitations. Hu et al. (2024) evaluates Chinese web search agents but lacks rigorous control over task difficulty and answer accessibility. Xu et al. (2024) focuses on dynamic, time-sensitive QA tasks but does not emphasize multi-hop retrieval. Lyu et al. (2025) evaluates RAG systems under the CRUD (Create, Read, Update, Delete) paradigm, yet primarily emphasizes generation quality over retrieval path validation. Liu et al. (2023) and Li et al. (2023b) focus on domain-specific medical QA, but do not evaluate open-ended retrieval or browsing strategies.

To address these limitations in Chinese retrieval benchmarking, BrowseComp-ZH introduces three key innovations: (1) reverse-designed tasks in native Chinese to avoid translation artifacts; (2) rigorous multi-step validation to ensure high retrieval difficulty and answer verifiability; and (3) broad model coverage for evaluating both open-source and proprietary agents. It establishes a high-difficulty benchmark for Chinese web retrieval, supporting systematic evaluation of agents' ability to navigate fragmented, unstructured, and linguistically diverse Chinese information sources.

3 The BrowseComp-ZH Dataset

3.1 Dataset Construction

Inspired by Wei et al. (2025), we adopt a reverse construction strategy: each task begins with a factual answer, from which an elaborate, multi-constraint query is crafted to make direct retrieval non-trivial. To ensure high-quality annotation, we recruited 10 expert contributors (with Master's or PhD degrees) who have extensive experience in both LLM usage and web search. The overall process comprises following two stages:

Stage 1: Topic and Answer Selection Each annotator selects at least 5 topics from a predefined list spanning *Film & TV*, *Technology*, *Art*, *History*, *Sports*, *Music*, *Geography*, *Policy & Law*, *Medicine*, *Video Games*, and *Academic Research*, based on their personal interests. For each topic, they identify several factual answers (e.g., person names, dates, titles, institutions) that meet two criteria: (1) The selected facts must be objective statements that can be independently verified through reliable sources, without the need for interpretation or inference; and (2) they must be concrete and specific enough to exclude overly generic or widely known common-sense facts.

Stage 2: Reverse Question Design Building on the selected factual answers, annotators construct complex queries that require integrating contextual cues and external knowledge. The design follows three key principles:

- **Multi-constraint design**: Each question combines temporal, spatial, categorical, or descriptive conditions to ensure answer uniqueness; not all conditions are required, but at least two dimensions are typically combined;
- **Non-trivial retrieval**: Annotators test each query on Baidu, Bing, and Google, using three distinct keyword combinations per search engine. If the correct answer appears on the first page of any search engine, the query will be revised or further constrained;

- **Evidence traceability**: Each sample includes at least one authoritative source URL that validates the logical connection between the query constraints and the target answer.

To further discourage shortcut-based resolution, all questions are tested using GPT-4o and DeepSeek, both operating in web-enabled search mode. If both models consistently retrieve the correct answer with minimal effort, the query will be revised to increase its complexity by methods including introducing implicit constraints or obfuscating key lexical signals. This process yields 480 preliminary samples, which are subsequently reviewed via the quality control procedure detailed in Section 3.2.

3.2 Quality Control

To ensure the rigor and challenge of BrowseComp-ZH, we implement a two-stage quality control protocol: one focusing on question difficulty and the other on answer uniqueness.

Stage 1: Question Difficulty Validation Although annotators are required to check whether a question can be quickly solved by search engines during the design process, variations in search ability and prior knowledge could introduce inconsistencies, leading to the inclusion of overly simple questions. To address this, we conduct a cross-checking phase:

- Each annotator validates questions written by others using only search engines (no LLMs);
- A strict 10-minute time limit is applied per question;
- If the answer is found within the time limit, the task is labeled *low difficulty*;
- If the answer is not found and the question structure is logical and verifiable, it is labeled *high difficulty*.

In this stage, annotators identified 76 simple samples, and after filtering them out, we are left with 404 high-difficulty candidates.

Stage 2: Answer Uniqueness Validation This stage focuses on ensuring the uniqueness of the answers, meaning that there is only one correct and unambiguous answer that satisfies all the constraints of each question. We employ a human-in-the-loop approach as follows:

- Multiple top-performing AI agents refer to the models that perform best on the original 404 high-difficulty candidates based on their ability to reason and retrieve accurate answers. AI agents, including OpenAI DeepSearch, Perplexity, Doubao (with deep reasoning), OpenAI O1, and Gemini 2.5-Pro, first generate reasoning processes and answers for each question.
- These results are then manually verified by annotators to check if any alternative answers satisfy the constraints of the question.
- If any alternative answer meets all task constraints (e.g., factual accuracy, specificity, verifiability) but differs from the original answer, the task is considered ambiguous and rejected.

This process eliminates 115 ambiguous samples, resulting in a final benchmark of 289 validated questions that maintains high levels of difficulty and answer verifiability.

3.3 Data Statistics

In this section, we present statistics on the topic distribution, question and answer length distribution of our curated BrowseComp-ZH dataset.

Topic distribution. Fig. 2 presents the distribution of samples across 11 topic domains in the BrowseComp-ZH dataset. As illustrated, the most represented categories include *Film & TV* (15.6%), *Art* (13.8%), and *Geography* (12.8%), reflecting the diverse interests of annotators and a broad coverage of Chinese web content. The dataset also features *Music* (11.1%), *History*(10.0%), and *Medicine* (9.0%). Conversely, topics like *Policy & Law* (3.5%) and *Academic Papers* (2.4%) have fewer samples, likely due to the complexity of sourcing factual answers from these areas. The distribution underscores the multi-disciplinary nature of the dataset, aimed at evaluating language models across a wide array of knowledge domains.

QA Length distribution. Fig. 3 illustrates the distribution of question and answer lengths in the BrowseComp-ZH dataset. As shown in Fig. 3 (a), the question length predominantly ranges between

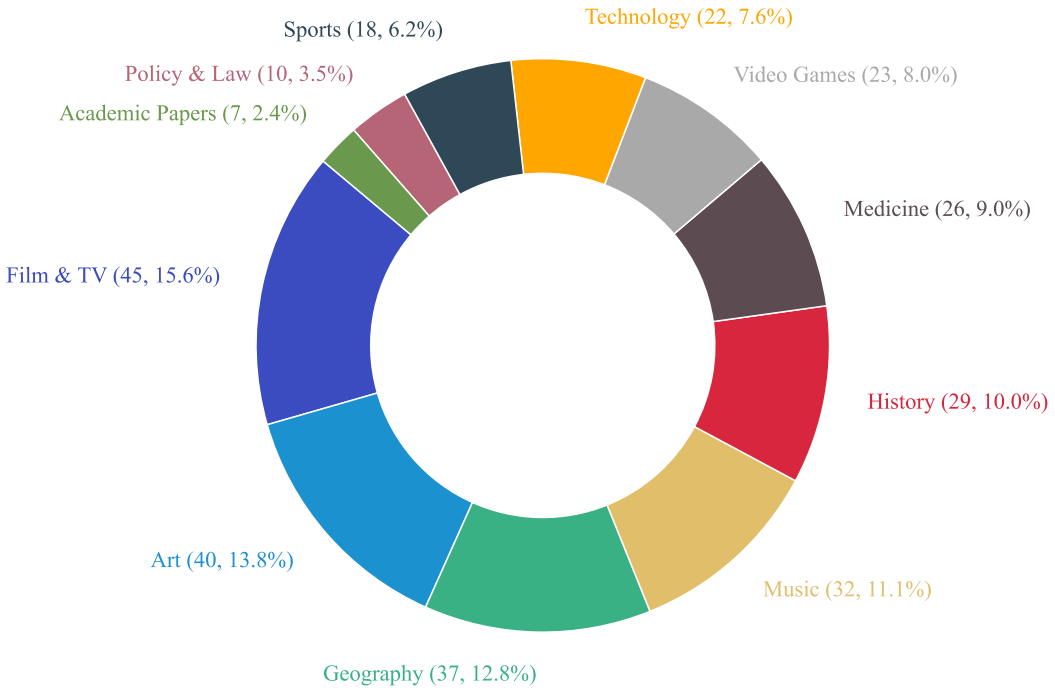
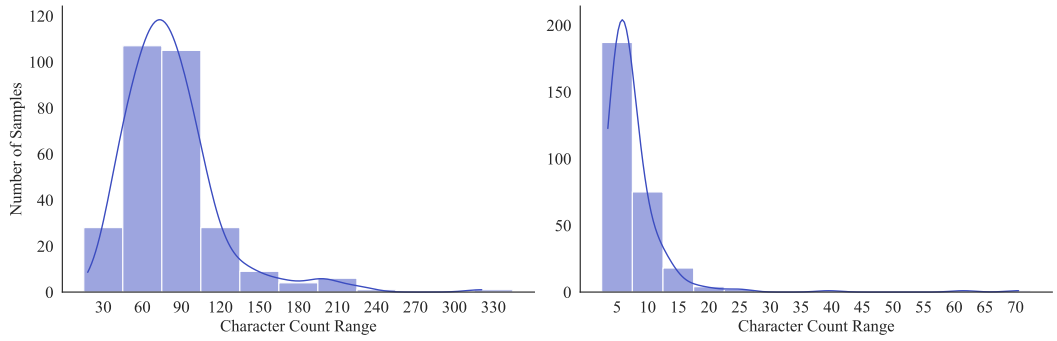


Figure 2: Distribution of samples across 11 topic domains in BrowseComp-ZH dataset.



(a) Question Length Distribution

(b) Answer Length Distribution

Figure 3: Distribution of question and answer lengths in the BrowseComp-ZH dataset.

60 to 90 characters, with the majority of questions falling within this interval. This suggests that questions are designed to be succinct but information-dense, requiring substantial detail to challenge the models. In Fig. 3 (b), the answer length distribution shows that answers are generally short, typically falling within the range of 5 to 10 characters. This aligns with the design principle of providing precise and verifiable answers, ensuring that the responses are direct and concise. These length distributions reflect the high information density of the dataset, emphasizing both complexity in the questions and simplicity in the answers.

4 Benchmarks

4.1 Models

We evaluate a wide range of state-of-the-art open-source and proprietary models, as well as mainstream AI search products on BrowseComp-ZH, aiming to provide a diverse, comprehensive, and instructive benchmark.

- **Open-source models:** DeepSeek-V3 Liu et al. (2024), DeepSeek-R1 Guo et al. (2025), Qwen2.5-72B-Instruct Yang et al. (2024), Qwen3-235B-A22B Yang et al. (2024), QwQ-32B Qwen_Team (2024), and LlaMa4 (maverick-instruct-basic) Meta (2024).
- **Closed-source models:** GPT-4o OpenAI (2024a), O1 OpenAI (2024b), O4-mini OpenAI (2025b), Claude-3.5-Sonnet() Anthropic (2024), Claude-3.7-Sonnet(20250219) Anthropic (2025), Gemini-2.0-Flash() Google (2024b), Gemini-2.5-Pro(preview-03-25) Google (2024a), and Qwen2.5-MAX (qwen-max-2025-01-25) Qwen_Team (2025)
- **AI search products:** DeepResearch OpenAI (2025a), Grok-3 xAI (2025), Perplexity (Research mode) Perplexity (2025), Doubao (with both deep search and standard modes) ByteDance (2025), Kimi (deep think version) MoonShot_AI (2025), Yuanbao (Hunyuan Model) Tencent (2025), and DeepSeek (with both deep think and standard modes) DeepSeek (2025).

4.2 Setting

For models such as O1 and Claude-3.7, inference is performed with fixed temperature and top-p settings, as these parameters are not user-configurable. For models that support custom decoding parameters, we follow the configuration used in DeepSeek R1 Guo et al. (2025), setting temperature to 0.6 and top-p to 0.95. For AI search products, we recruited human annotators to perform GUI-based interactions and recorded the outputs for subsequent analysis.

Since the questions in BrowseComp-ZH benchmark are designed to elicit concise factual answers, evaluating the correctness of model outputs is straightforward. For both open-source and closed-source models, which are generally capable of accurately following the provided instructions, we extract answers from the model responses using regular expressions and score them using GPT-4o. The grading prompts are adapted from those used in the original BrowseComp benchmark Wei et al. (2025). Full grading prompts are included in Appendix 5.

In contrast, AI search products typically undergo additional post-training to align with product-specific features, which often reduces their ability to follow instructions precisely. Therefore, we employ human annotators to manually extract the answers from these systems and verify their consistency with the ground truth.

The instructions provided to both open-source and closed-source models, as well as AI search products, are detailed in Appendix 4.

4.3 Metrics

To quantitatively evaluate model performance, we report both accuracy and calibration error. For the calculation of calibration error, we partition the predicted probabilities into five bins: [0–0.2), [0.2–0.4), [0.4–0.6), [0.6–0.8), and [0.8–1.0]. For each bin, we compute the absolute difference between the average predicted probability and the actual accuracy, and then calculate a weighted average across all bins. The ECE is formally defined as:

$$\mathrm{ECE} = \sum_{i=1}^{B} \frac{n_i}{N} |acc(i) - conf(i)|$$

where B is the number of bins, n_i is the number of samples in the i-th bin. $acc(i)$ is the empirical accuracy of the i-th bin, $conf(i)$ is the average predicted confidence in the i-th bin.

4.4 Performance

As shown in the performance comparison of models and AI search products on our benchmark (Tab. 1), several systems, such as Qwen2.5-72B-Instruct (6.6% accuracy), GPT-4o (6.2%), and Claude-3.5-Sonnet (5.5%), exhibited limited performance, consistently achieving low accuracy scores across tasks, highlighting the challenging nature of the dataset we proposed. Among all evaluated systems, OpenAI's DeepResearch achieved the highest accuracy (42.9%), followed by OpenAI's O1 (29.1%) and Google's Gemini-2.5-Pro (27.3%). Interestingly, even models without browsing capabilities, such as DeepSeek R1, O1, and Gemini-2.5-Pro, which rely solely on their internal world knowledge, managed to achieve accuracies exceeding 20%, demonstrating the strength of

Model	Category	Reasoning	Browsing	Accuracy	Calibration Error(%)	Enterprise
DeepSeek-V3	Open-Source	N	N	8.7%	72	DeepSeek
DeepSeek-R1	Open-Source	Y	N	23.2%	59	DeepSeek
Qwen2.5-72B-Instruct	Open-Source	N	N	6.6%	62	Alibaba
Qwen3-235B-A22B(Non-Thinking)	Open-Source	N	N	8.0%	80	Alibaba
Qwen3-235B-A22B(Thinking)	Open-Source	Y	N	13.2%	67	Alibaba
QwQ-32B	Open-Source	Y	N	11.1%	64	Alibaba
LlaMa4	Open-Source	N	N	4.8%	70	Meta
GPT4o	Closed-Source	N	N	6.2%	73	OpenAI
O1	Closed-Source	Y	N	29.1%	52	OpenAI
O4-mini	Closed-Source	Y	N	15.2%	42	OpenAI
Claude-3.5-Sonnet	Closed-Source	N	N	5.5%	78	Anthropic
Claude-3.7-Sonnet	Closed-Source	Y	N	17.7%	71	Anthropic
Gemini-2.0-Flash	Closed-Source	N	N	6.9%	74	Google
Gemini-2.5-Pro	Closed-Source	Y	N	27.3%	59	Google
Qwen2.5-MAX	Closed-Source	N	N	7.6%	78	Alibaba
OpenAI DeepResearch	AI Search Product	-	Y	**42.9%**	9	OpenAI
Grok3 (Research)	AI Search Product	-	Y	12.9%	39	xAI
Perplexity (Research)	AI Search Product	-	Y	22.6%	53	Perplexity
Doubao (Deep Search)	AI Search Product	-	Y	26.0%	61	ByteDance
Doubao (Standard)	AI Search Product	-	Y	18.7%	37	ByteDance
Kimi (Deep Think)	AI Search Product	-	Y	8.0%	58	Moonshot
Yuanbao (Hunyuan Model)	AI Search Product	-	Y	12.2%	56	Tencent
DeepSeek (Deep Think)	AI Search Product	-	Y	7.6%	65	DeepSeek
DeepSeek (Standard)	AI Search Product	-	Y	4.8%	66	DeepSeek

Table 1: Performance of various models and AI search products on BrowseComp-ZH dataset.

high-capacity language models. Overall, AI search products equipped with retrieval mechanisms outperformed other categories, followed by closed-source APIs, while open-source models performed the worst on this challenging benchmark. These results demonstrate the effectiveness of our benchmark in differentiating models across a wide range of performance levels.

4.5 Analysis of Reasoning Ability of LLM

Since the tasks in our proposed BrowseComp-ZH benchmark require both extensive world knowledge and strong reasoning abilities, we further investigate the impact of reasoning on model performance. As summarized in Tab. 1, we compare models with and without explicit reasoning mechanisms, including DeepSeek-V3 versus DeepSeek-R1, Claude-3.5-Sonnet versus Claude-3.7-Sonnet, and Gemini-2.0-Flash versus Gemini-2.5-Pro. Across all comparisons, models enhanced with reasoning capabilities consistently demonstrate substantial performance gains. For example, DeepSeek-R1 achieves an accuracy of 23.2%, markedly improving upon DeepSeek-V3's 8.7%, while Claude-3.7-Sonnet outperforms Claude-3.5-Sonnet (17.7% vs. 5.5%).

A similar pattern emerges among AI search products. Doubao (Deep Search) achieves a nearly 8% absolute improvement over Doubao (Standard) (26.0% vs. 18.7%), highlighting the benefits of enhanced reasoning in facilitating more accurate and iterative retrieval for complex queries.

However, this trend is less evident in DeepSeek's AI search products. Notably, all versions of DeepSeek's search system perform only a single round of retrieval, limiting the extent to which improved reasoning capabilities can be leveraged. Consequently, DeepSeek's Deep Search variant does not exhibit a significant performance advantage over its standard counterpart.

4.6 Analysis of AI Search Products

Currently, AI search systems can be broadly categorized into two types: those that perform a single retrieval to answer a user's query (e.g., Kimi, Tencent Yuanbao based on the Hunyuan model, and DeepSeek), and those that conduct multiple rounds of retrieval, iteratively refining or expanding the search based on the query and intermediate results (e.g., DeepResearch, Perplexity, and Doubao). Statistical analysis shows that systems employing multi-round retrieval achieve significantly higher accuracy, with DeepResearch reaching 42.9%, Doubao (Deep Search) achieving 26.0%, and Perplexity attaining 22.6%, compared to single-retrieval systems such as Kimi (8.0%), Yuanbao (12.2%), and DeepSeek (7.6%). This trend aligns with the nature of tasks in BrowseComp-ZH, which

often involve multi-faceted queries, making it challenging to obtain accurate answers through a single retrieval operation.

Notably, we observe a counterintuitive phenomenon: enabling search functionality for DeepSeek-R1 leads to a substantial decline in performance, with accuracy dropping from 23.2% in the direct-answer (no web access) setting to 7.6% when web search is enabled. We hypothesize that this degradation arises because, without effective alignment mechanisms, the model may rely on less reliable retrieved content, which in turn overrides its more accurate internal knowledge.

This observation highlights a critical challenge for large language models: effectively reconciling retrieved evidence with internal representations remains non-trivial. Furthermore, integrating retrieval capabilities without robust post-retrieval reasoning and alignment strategies may, in some cases, hinder rather than enhance model performance.

4.7 Calibration Analysis

We also evaluate the model's calibration, which measures the alignment between predicted confidence scores and actual accuracy. Following the methodology adopted in BrowseComp, we require models to provide confidence estimates alongside their predictions during evaluation. As shown in Tab. 1, integrating search functionality results in increased calibration errors. For instance, the calibration error for DeepSeek-R1 increases from 59% in the direct-answer setting to 65% when search is enabled. This trend is consistent with the observations reported in BrowseComp.

5 Conclusion and Discussion

This study introduces BrowseComp-ZH, the first benchmark specifically designed to evaluate the web browsing and reasoning capabilities of large language models (LLMs) in the Chinese information environment. Inspired by the BROWSECOMP benchmark, we construct challenging question-answer pairs that require multi-hop retrieval, information filtering, and logical reasoning to derive concise, factual answers. To ensure the high difficulty of each question and the uniqueness of each answer, we implement a rigorous two-stage quality control pipeline that includes a three-engine keyword validation process and human-in-the-loop verification, ensuring that answers are both difficult to retrieve and unambiguous.

We construct 289 high-quality samples across 11 diverse topics, including *Film & TV*, *History*, *Technology*, *Medicine*, and more. Using these tasks, we conduct extensive evaluations on over 20 models and AI search products, encompassing open-source LLMs, closed-source APIs, and AI search products. As shown in Tab. 1, most standalone LLMs—such as GPT-4o, Qwen2.5-72B, and Llama-4—achieve limited accuracy, highlighting the difficulty of the benchmark. Models with stronger reasoning abilities, such as O1 (29.1%) and Gemini-2.5-Pro (27.3%), demonstrate substantial improvements, underscoring the critical role of reasoning for complex question answering. AI search products employing multi-turn retrieval, including DeepResearch (42.9%) and Doubao (Deep Search) (26.0%), further outperform purely parametric models, illustrating the effectiveness of test-time scaling through iterative retrieval. These results reflect the inherent challenges of the Chinese web environment, where fragmented information and inconsistent indexing complicate single-shot search strategies.

Limitations. Despite the innovations in BrowseComp-ZH's design and quality control, several limitations remain. First, the current dataset is relatively small; increasing both the sample size and the diversity of question types would improve its representativeness. Second, although we apply rigorous validation to ensure answer uniqueness, it cannot be fully guaranteed. In particular, the dynamic nature of the web means that factual answers may evolve or become inconsistent over time, making stability and reproducibility an ongoing challenge.

Future Work In future work, we plan to incorporate a broader range of questions to enable more comprehensive and accurate evaluations of the models. We will also conduct an in-depth analysis of their reasoning mechanisms and search strategies. Furthermore, additional case studies will be carried out to examine failure cases across different models. Finally, we aim to explore methods to further improve the models' browsing and reasoning capabilities, such as leveraging post-training techniques like reinforcement learning.

References

Anthropic. 2024. Introducing Claude 3.5 Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet.

Anthropic. 2025. Introducing Claude 3.7 Sonnet and Claude Code. https://www.anthropic.com/news/claude-3-7-sonnet.

ByteDance. 2025. https://www.doubao.com/chat/.

DeepSeek. 2025. https://chat.deepseek.com/.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130* (2024).

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

Marcos Fernández-Pichel, Juan C Pichel, and David E Losada. 2024. Search Engines, LLMs or Both? Evaluating Information Seeking Strategies for Answering Health Questions. *arXiv preprint arXiv:2407.12468* (2024).

Google. 2024a. Gemini 2.5: Our most intelligent AI model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/.

Google. 2024b. Introducing Gemini 2.0: our new AI model for the agentic era. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/.

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

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*. PMLR, 3929–3938.

Guangxin He, Zonghong Dai, Jiangcheng Zhu, Binqiang Zhao, Qicheng Hu, Chenyue Li, You Peng, Chen Wang, and Binhang Yuan. 2024. Zero-Indexing Internet Search Augmented Generation for Large Language Models. *arXiv preprint arXiv:2411.19478* (2024).

Chuanrui Hu, Shichong Xie, Baoxin Wang, Bin Chen, Xiaofeng Cong, and Jun Zhang. 2024. Level-Navi Agent: A Framework and benchmark for Chinese Web Search Agents. *arXiv preprint arXiv:2502.15690* (2024).

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551* (2017).

Hanyu Lai, Xiao Liu, Iat Long Iong, Shuntian Yao, Yuxuan Chen, Pengbo Shen, Hao Yu, Hanchen Zhang, Xiaohan Zhang, Yuxiao Dong, et al. 2024. AutoWebGLM: A Large Language Model-based Web Navigating Agent. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5295–5306.

Hanyu Lai, Xiao Liu, Hao Yu, Yifan Xu, Iat Long Iong, Shuntian Yao, Aohan Zeng, Zhengxiao Du, Yuxiao Dong, and Jie Tang. 2025. WebGLM: Towards an Efficient and Reliable Web-Enhanced Question Answering System. *ACM Transactions on Information Systems* (2025).

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

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jingyuan Wang, Jian-Yun Nie, and Ji-Rong Wen. 2023a. The web can be your oyster for improving large language models. *arXiv preprint arXiv:2305.10998* (2023).

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023b. Huatuo-26m, a large-scale chinese medical qa dataset. *arXiv preprint arXiv:2305.01526* (2023).

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, et al. 2023. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems* 36 (2023), 52430–52452.

Yuanjie Lyu, Zhiyu Li, Simin Niu, Feiyu Xiong, Bo Tang, Wenjin Wang, Hao Wu, Huanyong Liu, Tong Xu, and Enhong Chen. 2025. Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models. *ACM Transactions on Information Systems* 43, 2 (2025), 1–32.

Andreas Martin, Hans Friedrich Witschel, Maximilian Mandl, and Mona Stockhecke. 2024. Semantic Verification in Large Language Model-based Retrieval Augmented Generation. In *Proceedings of the AAAI Symposium Series*, Vol. 3. 188–192.

Meta. 2024. LLaMA 4. `https://www.llama.com/models/llama-4`.

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

MoonShot_AI. 2025. `https://kimi.moonshot.cn/`.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332* (2021).

OpenAI. 2024a. hello-gpt-4o. `https://openai.com/index/hello-gpt-4o/`.

OpenAI. 2024b. Introducing OpenAI o1. `https://openai.com/o1/`.

OpenAI. 2025a. Introducing deep research. `https://openai.com/index/introducing-deep-research/`.

OpenAI. 2025b. Introducing OpenAI o3 and o4-mini. `https://openai.com/index/introducing-o3-and-o4-mini/`.

Perplexity. 2025. `https://www.perplexity.ai/`.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2020. KILT: a benchmark for knowledge intensive language tasks. *arXiv preprint arXiv:2009.02252* (2020).

Qwen_Team. 2024. QwQ: Reflect Deeply on the Boundaries of the Unknown. `https://qwenlm.github.io/blog/qwq-32b-preview/`

Qwen_Team. 2025. Qwen2.5-Max: Exploring the Intelligence of Large-scale MoE Model. `https://qwenlm.github.io/blog/qwen2.5-max/`.

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

Tencent. 2025. https://yuanbao.tencent.com/chat/.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. *arXiv preprint arXiv:1803.05355* (2018).

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. 2023. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214* (2023).

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. *arXiv preprint arXiv:2504.12516* (2025).

Nirmalie Wiratunga, Ramitha Abeyratne, Lasal Jayawardena, Kyle Martin, Stewart Massie, Ikechukwu Nkisi-Orji, Ruvan Weerasinghe, Anne Liret, and Bruno Fleisch. 2024. CBR-RAG: case-based reasoning for retrieval augmented generation in LLMs for legal question answering. In *International Conference on Case-Based Reasoning*. Springer, 445–460.

xAI. 2025. https://grok.com/.

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

Haoyi Xiong, Jiang Bian, Yuchen Li, Xuhong Li, Mengnan Du, Shuaiqiang Wang, Dawei Yin, and Sumi Helal. 2024. When search engine services meet large language models: visions and challenges. *IEEE Transactions on Services Computing* (2024).

Zhikun Xu, Yinghui Li, Ruixue Ding, Xinyu Wang, Boli Chen, Yong Jiang, Hai-Tao Zheng, Wenlian Lu, Pengjun Xie, and Fei Huang. 2024. Let llms take on the latest challenges! a chinese dynamic question answering benchmark. *arXiv preprint arXiv:2402.19248* (2024).

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115* (2024).

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).

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

A Technical Appendices and Supplementary Material

A.1 Instructions for model prediction

In the evaluation, we followed the BrowseComp instructions for model predictions. Additionally, for both open-source and closed-source models, which may exhibit a refusal to answer (often stating the lack of search capabilities), we included an instruction that prompts the models to rely on their intrinsic knowledge for providing answers.

Instruction for Open-source/Closed-source Model:

ZH:
如果你回复的问题需要借助外部资源，请根据你自身的知识储备给出具体答案，而不是拒答后让用户自行查询。
你的回复应遵循以下格式：
Explanation: {你对最终答案的解释}
Exact Answer: {你简洁的最终答案}
Confidence: {你对答案的置信度得分在 0% 到 100% 之间}

EN:
If the question you are responding to may require external resources, please provide a concrete answer based on your own knowledge base, rather than refusing to answer and directing the user to search independently.
Your response should follow the format below:
Explanation: {Your reasoning behind the final answer}
Exact Answer: {A concise statement of your final answer}
Confidence: {Your confidence in the answer, on a scale from 0% to 100%}

Instruction for AI Search Product:

ZH:
你的回复应遵循以下格式：
Explanation: {你对最终答案的解释}
Exact Answer: {你简洁的最终答案}
Confidence: {你对答案的置信度得分在 0% 到 100% 之间}

EN:
Your response should follow the format below:
Explanation: {Your reasoning behind the final answer}
Exact Answer: {A concise statement of your final answer}
Confidence: {Your confidence in the answer, on a scale from 0% to 100%}

Figure 4: Instruction for model prediction.

A.2 Instruction for grading

We adopt the same grading prompt as used in BrowseComp and employ GPT-4o for the grading process.

Judge whether the following [response] to [question] is correct or not based on the precise and unambiguous [correct_answer] below.

[question]: {question}
[response]: {response}

Your judgement must be in the format and criteria specified below:

extracted_final_answer: The final exact answer extracted from the [response]. Put the extracted answer as 'None' if there is no exact, final answer to extract from the response.
[correct_answer]: {correct_answer}

reasoning: Explain why the extracted_final_answer is correct or incorrect based on [correct_answer], focusing only on if there are meaningful differences between [correct_answer] and the extracted_final_answer. Do not comment on any background to the problem, do not attempt to solve the problem, do not argue for any answer different than [correct_answer], focus only on whether the answers match.

correct: Answer 'yes' if extracted_final_answer matches the [correct_answer] given above, or is within a small margin of error for numerical problems. Answer 'no' otherwise, i.e. if there if there is any inconsistency, ambiguity, non-equivalency, or if the extracted answer is incorrect.

confidence: The extracted confidence score between 0% and 100% from [response]. Put 100 if there is no confidence score available.

Figure 5: Prompt for model grading.