

WebReal: Benchmarking DeepSearch Agents on Real-world User Information Needs

Yifei Wang^{1,2,3}, Shangwen Lv³, Zhiyuan Xiong³, Haoyuan Zhang³, Yue Zhang³,
Rex Yueqi Wang³, Runquan Xie³, Linjing Li^{1,2†}, Zhanhui Kang³, Daniel Dajun Zeng^{1,2}

¹ State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences

² School of Artificial Intelligence, University of Chinese Academy of Sciences

³ Tencent, AI Data

[†] Corresponding author.

Abstract

Evaluating web browsing agents demands benchmarks that faithfully reflect how real users seek information online. We argue that existing benchmarks suffer from a **structural misalignment** with genuine user needs: their reverse-engineered, puzzle-style queries exercise a narrow slice of web-searching capabilities while leaving the majority of authentic information needs untested. To characterize this gap, we develop an **operation-oriented taxonomy** comprising six cognitive operations: *direct & Implicit retrieval*, *chain traversal*, *document extraction*, *computation*, *disambiguation*, and *temporal linking*. Applying this taxonomy reveals severe distributional skew in existing benchmarks, over **64%** of BrowseComp queries reduce to disambiguation alone. To close this gap, we introduce **WebReal**, a benchmark of **507** human-verified QA pairs derived from authentic user requests, spanning **15** domains across all six operation types. Experiments across state-of-the-art agents reveal that WebReal exposes capability profiles and ranking shifts invisible to existing benchmarks: models with near-identical aggregate scores exhibit divergent strengths across operation types, and performance degrades sharply with reasoning complexity in operation-type-specific patterns. These findings demonstrate that WebReal enables a finer-grained decomposition of web search capability than any prior benchmark. Our benchmark is publicly available [here](#).

1 Introduction

The rapid evolution of large language models into tool-augmented web browsing agents has fundamentally reshaped how people seek information online. By interacting with search engines and navigating live web content, these models transcend the limitations of their pre-trained knowledge, enabling continuous updates and access to real-time information. However, despite rapid advances, evaluating their ability to retrieve, filter, and reason over dynamic web content remains a key challenge.

Existing benchmarks rely heavily on manual curation, reverse-engineering queries from known answers, or constructing knowledge graphs from Wikipedia (**summarized in Table 1**). These construction strategies converge on a common failure mode: they produce puzzle-style tasks that are far removed from the information needs real users actually express. Real users do not ask “which person simultaneously satisfies these seven constraints”; they verify policy clauses, compute financial ratios, or trace cross-domain connections anchored to a shared time point. Such authentic information needs are inherently diverse, and it is precisely this diversity that gives rise to a wide spectrum of web operation types, including document-grounded extraction, retrieval-augmented computation, and temporal linking. Entity-centric puzzle tasks, however, exercise only a narrow slice of this operational space, systematically crowding out the operations that dominate genuine user needs. This

creates a **structural misalignment** between what benchmarks measure and what agents actually need to do, leaving the true capability of web browsing agents in handling real-world information needs largely uncharted.

To address both shortcomings, we introduce **WebReal**, an **authentic-user-intent-grounded diagnostic benchmark** derived from information-seeking requests submitted by real users, which naturally reflect the full diversity of real-world user needs. Rather than attempting an unbiased sampling of raw search logs, WebReal aims to systematically diagnose agent capabilities across the full spectrum of complex, real-world operations. Starting from a collection of such requests and grounded in established frameworks from information science (Dervin, 1983; Taylor, 1968; Marchionini, 2006), we summarize an **operation-oriented taxonomy** organized around a central question: *what core cognitive operation must an agent perform to derive the answer?* We identify six operation types, **T1** Direct Retrieval, **T2** Chain Traversal, **T3** Document Extraction, **T4** Computation, **T5** Disambiguation, and **T6** Temporal Linking, each corresponding to a distinct cognitive bottleneck that emerges naturally from genuine user needs. Applying this taxonomy to existing benchmarks reveals a stark imbalance: over 64% of BrowseComp queries reduce to T5 alone (Shannon entropy 1.53 bits), while T3 and T4 together account for 32% of authentic queries yet are virtually absent from current benchmarks.

WebReal is constructed from the user requests through a five-stage pipeline that grounds open-ended queries into uniquely-answerable items, obfuscates named entities to eliminate shortcut retrieval, and filters out items solvable without genuine web browsing, yielding 507 human-verified query-answer pairs spanning 15 domains. Benchmarking six commercial web agents and multiple open-source frameworks reveals that **WebReal** poses genuine challenges to all current systems, with the best-performing agent achieving only 22.9%. Fine-grained analysis uncovers that web access yields highly uneven gains across operation types, that T6 Temporal Linking remains an open challenge even for the strongest systems, and that aggregate scores obscure fundamentally divergent competency profiles across models. These findings expose critical gaps in current agent capabilities and provide a principled diagnostic framework for guiding future development.

Our contributions are summarized as follows:

- **A new benchmark grounded in real user requests.** **WebReal** is the first deepsearch benchmark derived from authentic user information-seeking requests, capturing operational diversity that synthetic construction methods structurally cannot replicate.
- **An operation-oriented taxonomy for web information seeking.** We propose a six-type taxonomy (T1–T6) grounded in information science, and show that existing benchmarks collapse onto a single operation type (64% T5, $H=1.54$ bits), exposing a fundamental structural misalignment with real user needs.
- **A rigorous and scalable construction pipeline with quality guarantees.** We design a principled five-stage pipeline that transforms open-ended user requests into uniquely-answerable benchmark items, enabling an **LLM-assisted and human-verified construction pipeline**.
- **Comprehensive empirical analysis.** Evaluating 11 state-of-the-art agents, we find that no system dominates uniformly across operation types, with each model exposing a distinct failure mode: loop-prone models collapse on depth-requiring tasks while termination-prone models fail on tasks demanding sustained retrieval, revealing that DeepSearch capability is far more heterogeneous than aggregate scores suggest.

2 Related Work

Web browsing benchmarks are critical for evaluating tool-augmented agents, but existing ones suffer from fundamental **structural misalignment**: their construction fails to capture authentic user operational diversity, focusing instead on synthetic puzzle-style tasks. We analyze three dominant paradigms and their misalignments.

Reverse-engineering from known answers. This paradigm (BrowseComp (Wei et al., 2025), BrowseComp-ZH (Zhou et al., 2025)) crafts queries from pre-selected answers, biasing tasks

Table 1: Comparison of web information-seeking benchmarks. ✓ = yes; ✗ = no; ~ = partial.

Benchmark	Query Source	Size	Real-user Query	Entity Obfusc.	Answer Uniqueness	Oracle Filter	Multi-domain	Human Verified	Auto Scalable
REDSearcher (Chu et al., 2026)	Webpage&Wiki Research Tree synthesis	1,266	✗	✗	✓	✗	~	✓	✗
BrowseComp (Wei et al., 2025)	Human reverse-engineering	1,266	✗	✗	✓	✗	~	✓	✗
BrowseComp-ZH (Zhou et al., 2025)	Human reverse-engineering	289	✗	✗	✓	✗	~	✓	✗
WebWalkerQA (Wu et al., 2025b)	Crawl-then-generate (GPT-4o)	680	✗	✗	✓	✗	~	✗	✓
WebDancer (Wu et al., 2025a)	Entity-driven iterative synthesis	4,000+	✗	✗	~	✗	✗	✗	✓
WebShaper (Tao et al., 2025)	Formalization-driven synthesis	500+	✗	✗	✓	✗	✗	✗	✓
SealQA (Seal-0) (Pham et al., 2025)	Expert-curated (noisy-web)	1,000+	✗	✗	✓	✗	✓	✓	✗
InfoSeek (Xia et al., 2025)	Webpage Research Tree synthesis	50,000+	✗	✗	✓	✗	✗	✗	✓
DeepSearchQA (Gupta et al., 2026)	Expert-curated	900	✗	✗	✓	✗	✗	✗	✓
FRAMES (Krishna et al., 2025)	Human annotation	824	✗	✗	✓	✗	✗	~	✗
WebReal (Ours)	Authentic User Requests	507	✓	✓	✓	✓	✓	✓	✓

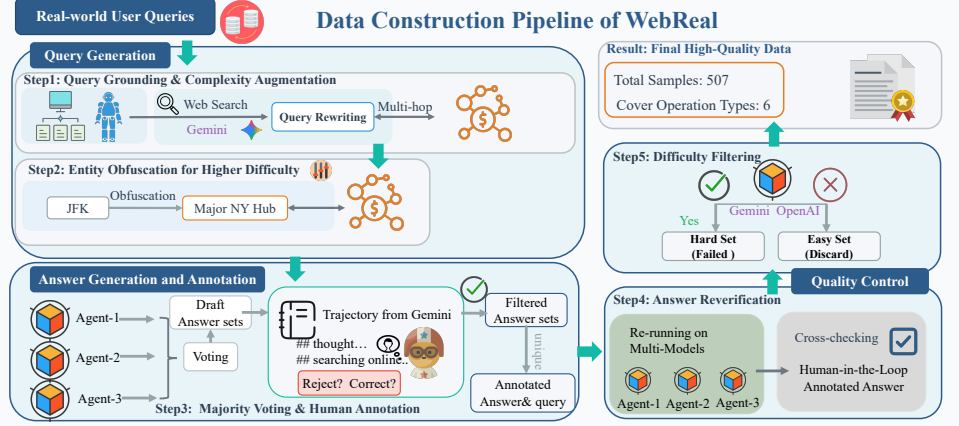


Figure 1: Overview of the five-stage data construction pipeline of WebReal. Starting from real-world user queries, the pipeline proceeds through (Φ_1) query grounding and complexity augmentation, (Φ_2) entity obfuscation, (Φ_3) majority voting and human-in-the-loop annotation, (Φ_4) multi-model answer reverification, and (Φ_5) difficulty filtering.

toward entity-disambiguation over genuine needs. Quantifiably, over 64% of BrowseComp queries are T5 (Disambiguation), while T3 (Document Extraction) and T4 (Computation), 57% of real user queries, are nearly absent, failing to test core agent capabilities.

Crawl- or entity-driven synthesis. Benchmarks like WebWalkerQA (Wu et al., 2025b) and WebDancer (Wu et al., 2025a) generate queries from crawled content/entities, over-sampling T1 (Direct Retrieval) and sidelining T2, T3, and T6. Real users perform complex tasks (e.g., policy verification, financial computation) beyond basic retrieval, so these benchmarks exacerbate misalignment between tasks and agent requirements.

Formalization- or curation-driven synthesis. This paradigm (WebShaper (Tao et al., 2025)) uses formal templates, marginalizing T3 and T6 in favor of constraint satisfaction tasks irrelevant to real needs. SealQA (Pham et al., 2025) (expert curation) and InfoSeek (Xia et al., 2025) limit operational diversity, failing to cover T1–T6 and perpetuating misalignment.

All prior benchmarks use *top-down* construction (answers, crawls, templates) instead of *bottom-up* authentic user needs, suppressing operational diversity and leaving agent capabilities untested. **This is a structural misalignment between the cognitive operations that existing benchmarks measure and those that web browsing agents must actually perform in deployment.** Our WebReal inverts this logic, using real user queries and our T1–T6 taxonomy to resolve misalignment. Its pipeline (entity obfuscation, human annotation, difficulty filtering) addresses prior scalability, realism, and coverage limits.

3 WebReal Benchmark

Let $\mathcal{Q}$ denote the universe of natural-language queries and $\mathcal{W}$ the live web corpus. A **web browsing agent** is a function $\mathcal{F} : \mathcal{Q} \times \mathcal{W} \rightarrow \mathcal{A}$ that maps a query and a browseable web environment to an answer. Formally, we begin with a seed corpus of real-world user query: $\mathcal{Q}_0 = \{q_1, q_2, \dots, q_n\}$, capturing the full spectrum of naturalistic information-seeking behaviors that users exhibit in practice. We propose a five-stage pipeline, which transforms

$\mathcal{Q}_0$ to construct a benchmark $\mathcal{D} = \{(q_i, a_i^*)\}_{i=1}^N$ where each pair is (i) grounded in real user intent, (ii) requires iterative search and reasoning to resolve, and (iii) is verifiably correct.

As shown in Figure 1, our pipeline consists of three major parts (Query Generation, Answer Generation and Annotation, Quality Control) and five stages. To ensure transparency, we summarize the retention statistics of this pipeline: starting from approximately 20,000 raw queries, Stage Φ_1 retained 3,210 queries (16.05% retention); subsequent stages further refined this set until Stage Φ_5 yielded the final 507 hard items from 1,180 candidates (43.0% retention). Detailed stage-by-stage statistics are provided in Appendix C.1.

3.1 Query Generation

Stage Φ_1 : Query Grounding and Complexity Augmentation. To capture queries that reflect genuine user information needs, we first filter $\mathcal{Q}_0$ to retain only information-seeking queries that require retrieving, filtering, or reasoning over external web content (e.g., "What is the best treatment for diabetes?"), discarding intent types such as code generation or creative writing that do not involve web retrieval. However, such real-world user queries are often *ambiguous*: they admit no single verifiable answers or even are *not understandable* without further context, making automated evaluation infeasible. To address this, we introduce **query grounding**, aiming to extend meaningful context or revise the origin query to obtain *unambiguous*, *understandable* and *verifiable* queries. Firstly, we obtain queries from logs and remove the identity features. For each seed query $q \in \mathcal{Q}_0$, we issue it to a web search engine and retrieve the content of the top- k returned web pages as context $C(q) = \psi(q)$, where ψ denotes the web-search operator. We then use this context to inject explicit constraints that reduce the answer space to a unique, verifiable response. We issue the origin query and web page contents to a powerful LLM (e.g. Gemini) and explicitly obtain rewritten queries¹. Formally, a query q is *grounded* with respect to context $C(q)$ if

$$|\mathcal{A}(q \mid C(q))| = 1, \quad (1)$$

i.e., the constrained answer space collapses to a singleton. For example, the vague query "Who runs the largest AI lab?" can be grounded to "As of Q1 2026, who is the CEO of the organization that developed GPT-4?", which admits exactly one correct answer.

Building on grounded queries, we further apply **complexity augmentation** via a rewriting function $R : \mathcal{Q} \times \mathcal{C} \rightarrow \mathcal{Q}'$, where $q' = R(q, C(q))$ and $\mathcal{Q}'$ denotes the augmented query set. This function is designed to produce queries that require synthesizing information from at least $h = 10$ distinct web sources. We perform search operators more than once and change the search keywords until we obtain an acceptable query or reach the maximum limit times (e.g. 5). Formally, we enforce the following *compositional constraint*:

$$\forall q' \in \mathcal{Q}' : \min_{\tau \in \mathcal{T}(q')} |\text{sources}(\tau)| \geq 10, \quad (2)$$

where $\mathcal{T}(q')$ is the set of valid answer trajectories for q' and $\text{sources}(\tau)$ returns the distinct URLs accessed in τ . It is important to note that this compositional constraint is evaluated against a *verified reference evidence set* rather than an absolute shortest global path, ensuring that a naturalistic search process would require synthesizing information across multiple sources. Unlike knowledge-graph-based methods that restrict multi-hop reasoning to fixed, pre-indexed entities and relations, our approach constructs questions directly from live web content, enabling broader knowledge coverage and temporal currency.

Stage Φ_2 : Entity Obfuscation for Higher Difficulty. A grounded query q' may still exist shortcuts (only one or two constraints can obtain the right answers.) via verbatim keyword search over its named entities, bypassing the multi-step reasoning the task is designed to elicit. To eliminate this shortcut, we introduce an entity obfuscation operator $E : \mathcal{Q}' \rightarrow \mathcal{Q}''$ that replaces named entity $e \in \mathcal{E}(q')$ with an indirect description $\hat{e} = \varphi(e)$ under a paraphrase map $\varphi : \mathcal{E} \rightarrow \mathcal{E}$, subject to two constraints: (a) $\hat{e} \neq e$ (the surface form is

¹We will remove the queries which are meaningless or too easy by expert human annotators.

altered), and (b) $\text{resolve}(\hat{e}) = e$ (the referent remains uniquely recoverable via contextual reasoning). The resulting obfuscated query set is

$$\mathcal{Q}'' = E(\mathcal{Q}') = \{q'' : q'' = E(q'), q' \in \mathcal{Q}'\}. \quad (3)$$

For instance, “JFK” is replaced by “the major hub serving New York”, requiring the agent to first resolve the indirect reference to John F. Kennedy International Airport before proceeding with retrieval. By construction, E strictly increases reasoning complexity without altering the answer target, rendering entity resolution an obligatory prerequisite.

3.2 Answer Generation and Annotation

Stage Φ_3 : Majority Voting & Human Annotation For each query $q'' \in \mathcal{Q}''$, we prompt an ensemble of M search agents to generate candidate answers in parallel. The top- K most frequent responses under majority voting form the **draft answer set** $\mathcal{A}^d = \{a_1^d, a_2^d, a_3^d\}$ ($K = 3$), filtering out idiosyncratic low-confidence outputs.

We then introduce a **human-in-the-loop annotation** stage to verify answer uniqueness. Each annotator is provided with the query q'' , the draft set $\mathcal{A}^d$, and a Gemini-generated reference trajectory $\tau^G = (\theta_1, s_1, \theta_2, s_2, \dots)$ of interleaved chain-of-thought steps θ_i and web-search actions s_i . The trajectory serves as a lightweight reference that reduces annotator workload by surfacing relevant evidence, without prescribing a specific answer. The annotator first identifies the subset of *valid* draft answers: $\mathcal{A}^v = \{a^d \in \mathcal{A}^d : a^d \text{ is a correct answer to } q''\}$.

If $|\mathcal{A}^v| \geq 2$, the query is deemed *ambiguous* and discarded, as it violates the uniqueness requirement. If $|\mathcal{A}^v| = 1$, the annotator independently derives their own answer a^h and compares it with the sole valid draft. If a^h agrees with $\mathcal{A}^v$, the pair is retained with $a^* = a^h$. If $|\mathcal{A}^v| = 0$, no draft answer is correct, the annotator’s own label is adopted directly as the gold answer, giving $a^* = a^h$. In both retained cases the gold label originates from the human annotator, ensuring answer quality is not contingent on model correctness. To ensure the reliability of this process, we conducted an inter-annotator agreement (IAA) study. Two independent annotators plus a third arbitrator evaluated the data, yielding a raw agreement of 89% (Cohen’s Kappa = **0.85**) for operation-type labeling and 92% (Cohen’s Kappa = **0.88**) for correctness verification. The output of Φ_3 is $\mathcal{D}^3 = \{(q'', a^*) : |\mathcal{A}^v| \leq 1\}$.

Stage Φ_4 : Answer Reverification We subject all such pairs to an automated **answer reverification** step. Concretely, each pair (q'', a^*) drawn from the $|\mathcal{A}^v| = 0$ subset of $\mathcal{D}^3$ is independently re-evaluated by all M agents. Each model i judges whether a^* is correct, producing a binary verdict $v_i \in \{0, 1\}$, and the pair is retained if and only if a strict majority approves, i.e., $\sum_{i=1}^M v_i > M/2$. Pairs failing this check are discarded, and the output is $\mathcal{D}^4 = \{(q'', a^*) \in \mathcal{D}^3 : \sum_{i=1}^M v_i > M/2\}$.

Stage Φ_5 : Difficulty Filtering We use powerful frontier models (e.g., Gemini and GPT-4) as an oracle evaluator G to partition $\mathcal{D}^4$ by difficulty. Define the **oracle correctness indicator**: $d(q'') = \mathbf{1}[G(q'', \mathcal{W}) \neq a^*]$. Queries where $d(q'') = 1$ (the oracle fails) are assigned to the **Hard Set** $\mathcal{D}_H$ and retained. Queries where $d(q'') = 0$ are assigned to the **Easy Set** $\mathcal{D}_E$ and discarded². Formally:

$$\mathcal{D}_H = \{(q'', a^*) \in \mathcal{D}^4 : d(q'') = 1\}, \quad \mathcal{D}_E = \mathcal{D}^4 \setminus \mathcal{D}_H. \quad (4)$$

The final benchmark is $\mathcal{D}_{\text{final}} = \mathcal{D}_H$. This ensures that every item *strictly exceeds the capability of the filtering oracle*, guaranteeing meaningful discrimination between agent architectures. We further validated the robustness of this LLM-as-judge filtering against human judgment: an evaluation on 150 sampled responses showed a **95.3%** overall agreement rate between the LLM judge and human experts (96.1% precision, 94.5% recall), confirming its reliability as an automated filter. To address concerns regarding potential architectural bias introduced by relying on a single oracle, we conducted a sensitivity analysis using GPT-4o as an alternative

²This set means that the question can be answered only by the LLM itself without the search engines.

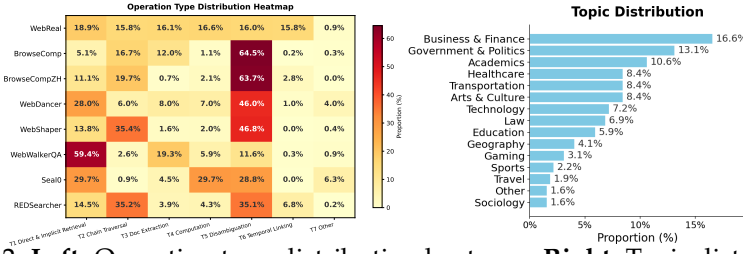


Figure 2: Left: Operation type distribution heatmap. Right: Topic distribution.

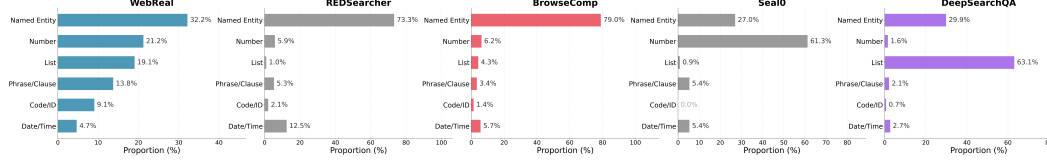


Figure 3: Answer type distribution across five benchmarks. REDSearcher and BrowseComp concentrate heavily on Named Entity (73.3% and 79.0%).

oracle. The results demonstrate an **86.5%** overlap in the resulting Hard Sets between Gemini and GPT-4o, proving that the filtering criterion successfully identifies inherently complex, non-trivial queries in a model-agnostic manner.

3.3 Benchmark Composition and Statistics

①**Operational balance.** As shown in Figure 2 Left, WebReal maintains a near-uniform distribution across all six operation types (15.8–18.9%), in stark contrast to BrowseComp (64.5% T5) and WebWalker (59.4% T5), confirming that genuine user needs span the full operational taxonomy rather than collapsing onto a single task type. This structural bias in existing benchmarks renders them inadequate for evaluating the full spectrum of agentic web search capabilities that real-world deployment demands. ②**Topical breadth.** As shown in Figure 2 Right, WebReal spans 15 domains with no single domain exceeding 17%, reflecting the naturally distributed topical diversity of real user requests. ③**Answer type diversity.** Unlike existing benchmarks dominated by a single type (BrowseComp: 79.0% Named Entity; DeepSearchQA: 63.1% List), WebReal distributes across six answer types, mirroring the heterogeneous responses that real users seek, shown in Figure 3.

4 Experiments

4.1 Experimental Setup

Evaluation Backbones. We evaluate three categories of agents to assess info-seeking capabilities. (1) **No-Agency Models** rely solely on parametric knowledge without any web access, serving as a knowledge-only baseline; we evaluate GPT-5.2, Gemini-2.5-Pro, and Doubao-Seed-1.6 (Seed et al., 2025). (2) **Commercial Deep-Research Systems** deploy proprietary built-in search tools within their respective web platforms, including Doubao-DR, Gemini-DR, OpenAI-DR, and Tongyi-DR. (3) **Open-source Agentic Frameworks** adopt the ReAct paradigm (Yao et al., 2023), which organizes each trajectory as an interleaved sequence of (*thought*, *action*, *observation*) tuples:

$$\mathcal{H}_T = (q, (\tau_0, a_0, o_0), (\tau_1, a_1, o_1), \dots, (\tau_T, a_T, o_T), y). \quad (5)$$

At each step, a single LLM issues actions drawn from two tools: a web search and browsing tools, with observations truncated upon reaching the maximum context length. This design covers the full task lifecycle, from query planning and iterative retrieval to final answer synthesis, and is instantiated with both general-purpose foundation models (Qwen-3-8B/32B (Yang et al., 2025)) and deepsearch-optimized models (WebDancer-32B (Wu et al., 2025a), WebShaper-32B (Tao et al., 2025), Tongyi-DR-30B-A3B (Team, 2025)).

Table 2: Pass@1 of different backbones on WebReal across six operation types. On BrowseComp, OpenAI-DR achieves 54.9%, confirming WebReal’s higher difficulty.

Backbone	WebReal						Avg.
	T1 Direct & Implicit	T2 Chain Traversal	T3 Doc Extraction	T4 Computation	T5 Disambig- uation	T6 Temporal Linking	
No Agency (No-thinking)							
GPT-5.2	11.8	9.4	5.0	0.9	6.3	23.5	7.4
Gemini-2.5-Pro	10.0	8.5	3.0	2.8	9.5	11.8	6.8
GPT-4o	9.1	5.7	1.0	0.0	3.2	5.9	4.0
No Agency (Thinking)							
Kimi-K2.5	12.8	15.2	3.0	8.6	15.9	18.8	11.0
DeepSeek-R1	9.2	14.2	3.0	8.7	9.5	17.6	9.2
Qwen3-235B-A22B	8.2	13.2	2.0	6.6	6.3	17.6	7.8
Kimi-K2-thinking	10.4	12.0	3.0	7.9	9.8	17.3	6.2
DR Commercial Systems							
OpenAI-DR	30.9	20.8	16.8	21.7	22.2	29.4	22.9
Qwen3.5-Plus-DR	23.8	20.6	12.3	12.5	14.0	26.7	17.7
Gemini-DR	20.0	19.8	11.9	15.1	12.7	17.6	16.3
Doubao-DR	15.5	18.9	12.9	12.3	17.5	11.8	15.1
Qwen3-Max-DR	18.2	14.2	12.0	12.3	15.9	17.6	14.5
Kimi-K2-thinking	15.0	10.9	14.7	11.9	13.2	14.3	13.4
Open-sourced Agentic Frameworks (ReAct)							
WebShaper-32B	17.3	18.9	17.8	8.5	15.9	17.6	15.7
Tongyi-DR-30B	11.8	13.2	15.8	10.4	12.7	11.8	12.7
WebDancer-32B	10.9	14.2	9.9	9.4	14.3	11.8	11.5
Qwen3-32B	10.0	11.3	6.9	10.4	9.5	17.6	9.9
Qwen3-8B	12.7	9.4	5.9	6.6	11.1	5.9	8.9

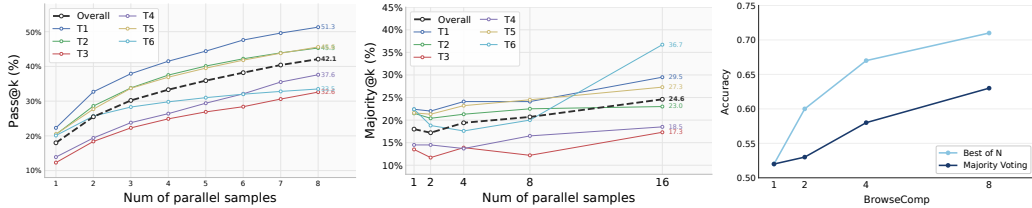


Figure 4: Pass@k (left) and Majority@k (middle) on WebReal, on BrowseComp (right).

Evaluation Metrics. We report **Pass@1** accuracy, evaluating each query with a single inference run. Answer correctness is determined by an LLM-as-judge protocol, which assesses semantic equivalence between the predicted answer and the gold answer.

4.2 Main Results

The main results are shown in Table 2. We have the following observations:

WebReal Poses Genuine Challenges to State-of-the-Art Systems. Even the strongest system, OpenAI-DR, achieves only 22.9% overall, and performance remains consistently low across all six operation types, with no system exceeding 30.9 on any single type. T3 (Doc Extraction) and T4 (Computation) prove universally difficult: no model of any category surpasses 17.8 on T3 or 21.7 on T4, indicating that deep document grounding and multi-step numerical reasoning remain unsolved regardless of system design. These results reveal critical capability gaps that existing benchmarks structurally overlook.

Web Access Provides Heterogeneous Benefits Across Operation Types. Web access yields the largest gains on T3 and T4, where no-agency models score near zero while the best

Table 3: **Left:** Fine-grained error type distribution (% of failed samples) across four open-source agents. Failure mode is detailed in Table 4 in Appendix; **Right:** Per-operation-type search query statistics. Mean and median are computed over all instances; Corr. and Incorr. denote the mean query count for correctly and incorrectly answered instances, respectively.

Error Type	Tongyi WS. WD. Qwen3				Type	Mean	Med.	Corr.	Incorr.	Gap
Ineffective Loop	43.7	48.7	0.7	0.7	T1	7.86	5.0	6.00	8.14	+2.14
Premature Termination	10.4	12.4	46.5	61.2	T2	6.10	5.0	4.64	6.38	+1.74
Poor Query Formation	30.1	37.0	50.2	36.1	T3	10.34	7.0	11.50	10.09	-1.41
Context Limit	10.7	0.0	0.0	0.0	T4	8.55	6.0	6.55	8.89	+2.34
Redundant Tool Call	4.4	0.3	0.7	0.0	T5	7.19	6.0	5.75	7.43	+1.68
Insufficient Coverage	0.5	0.8	1.5	1.9	T6	4.53	4.0	7.00	4.15	-2.85

commercial system reaches 16.8 and 21.7, confirming these tasks demand precise retrieval beyond any parametric knowledge. By contrast, Temporal Linking shows the smallest gap: no-agency thinking models already reach 18.8, suggesting that strong temporal reasoning partially substitutes for live retrieval. For T1 and T2, web access provides consistent but moderate gains, while T5 (Disambiguation) remains difficult even with retrieval, indicating that the bottleneck lies in entity resolution rather than information access.

Fine-Grained Decomposition Reveals Divergent Competency Profiles. Decomposing results across operation types reveals that no system dominates uniformly, and competency profiles diverge sharply by model category. OpenAI-DR leads on all six types but its margin narrows on T3 and T4, where even the best system stays below 22, suggesting these tasks impose a ceiling that retrieval alone cannot lift. Within the ReAct paradigm, WebShaper-32B excels at retrieval-heavy types (T1–T3: 17.3–17.8) but collapses on T4 (8.5), exposing a hard boundary between retrieval proficiency and numerical reasoning. Qwen3 models show the inverse: relatively competitive on T2 and T6 yet weak on T3 and T4, consistent with their premature-termination tendency on document-intensive tasks. Across all groups, T3 and T4 rank as the two hardest types, pointing to deep document grounding and multi-step computation as the critical unsolved challenges for current DeepSearch systems.

4.3 Test-time Scaling

As shown in Figure 4, $\text{pass}@k$ rises steadily from 18.0% to 42.1% at $k=8$, confirming that correct answers are within reach and that the primary bottleneck is consistency rather than capability. Yet $\text{Majority}@k$ grows far more slowly (24.6% at $k=16$), revealing a persistent $\text{Pass}@k$ – $\text{Majority}@k$ gap driven by high retrieval-path sensitivity. This divergence is strongly task-dependent: T3 shows negligible majority voting gain, indicating that document extraction is largely deterministic, whereas T6 surges from 18.5% to 36.7%, suggesting that temporal linking converges reliably once the correct retrieval path is identified. The markedly wider gap relative to BrowseComp directly reflects WebReal’s origin in authentic user requests: **real information needs are inherently more path-sensitive than synthetic ones**, exposing answer consistency as a critical yet underexplored dimension of agentic web search that current systems fail to address.

4.4 Search Behavior and Failure Patterns

DeepSearch Behavior of Agents. Table 3 (Left) reveals distinct behavioral profiles among the evaluated agents. **Tongyi** falls into a *deep-trap* pattern: it issues the most queries per instance (8.03 on average) yet achieves only moderate accuracy, as ineffective loops (43.7%) exhaust the context budget before a valid answer is produced, making it the only model to incur Context Exceedance errors (10.7%). **WebShaper** similarly suffers from high loop rates (48.7%), but its dominant failure is Poor Query Formulation, where structurally repetitive queries fail to deepen retrieval coverage. **WebDancer** exhibits a *shallow-abandon* pattern: it abandons search after a single unsuccessful attempt in nearly half of all failures, returning empty answers rather than pursuing iterative refinement. **Qwen3-32B** share a *conservative-abort* tendency dominated by Premature Termination (61.2%, respectively), and scaling from 8B to 32B yields negligible improvement, indicating that the bottleneck is behavioral rather than parametric. Collectively, these profiles expose two orthogonal failure axes: *over-*

persistence, where agents cannot escape unproductive loops, and *under-persistence*, where agents abandon retrieval prematurely, underscoring the need for training signals that reward adaptive query reformulation over both extremes.

Failure Pattern Across Operation Type. As shown in Figure 5, error patterns decompose along two orthogonal dimensions: model-intrinsic behavioral tendencies and task-driven amplification. Model-intrinsic patterns remain stable across operation types: loop-prone models (Tongyi, WebShaper) consistently incur Ineffective Loop errors while termination-prone models (WebDancer, Qwen3) default to premature abandonment, confirming that execution control failures are primarily behavioral rather than task-contingent (Table 3, Left).

Task structure, however, selectively amplifies these tendencies.

T3 and T6 universally exacerbate poor query formation: successful trajectories require *more* queries than failed ones ($\Delta = -1.41$ and -2.85 , respectively), yet all models exhaust their budget on shallow reformulations, a pattern reflected in elevated dominant error rates across both loop-prone and termination-prone models in Figure 5.

T4 and T5 amplify each model’s intrinsic failure mode in opposing directions: T4 drives WebDancer’s error rate to its peak (94%) via premature abandonment, while loop-prone models accrue more queries on failures than successes ($\Delta = +2.34$) yet still fail, exposing numerical synthesis as the true bottleneck beyond retrieval; T5 inverts this pattern, pushing WebShaper to its peak (76%) through redundant loops while WebDancer reaches its lowest error rate (39%). In sum, T3 and T6 reveal a shared gap in *adaptive query construction*, whereas T4 and T5 point to distinct targets of *numerical reasoning* and *loop escape* that require non-overlapping training interventions.

Implications for DeepSearch Development. First, execution control is the dominant bottleneck, yet polarized into over-persistence and under-persistence, requiring trajectory-level training signals that jointly penalize both extremes. Second, T3 and T6 call for rewards that credit progressive retrieval deepening, since successful trajectories require *more* queries than failed ones yet all models exhaust their budget on shallow reformulations. Third, T4 demands training that couples retrieval with numerical reasoning, as issuing more queries still fails to recover accuracy, confirming the bottleneck lies in arithmetic synthesis rather than search volume. Fourth, parameter scaling provides no remedy since Qwen3-32B replicates the error profile of Qwen3-8B almost exactly, underscoring the need for operation-type-specific trajectory optimization over capacity scaling.

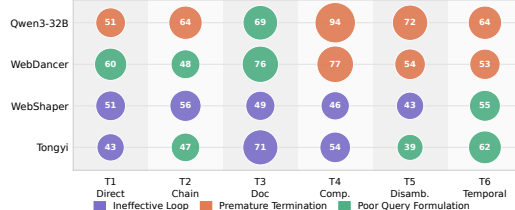


Figure 5: Error analysis across operation types.

4.5 Additional Analysis

The Effect of Answer Type. Figure 6 reveals two universal patterns: Named Entity is consistently the easiest type (OpenAI-DR: 28.9%), as named entities are well-indexed and directly retrievable, while Date/Time and Phrase/Clause are universally the hardest, requiring precise temporal anchoring and verbatim span extraction respectively. Beyond these global trends, answer type exposes model-specific strengths invisible to aggregate scores: OpenAI-DR leads on Named Entity (28.9%) and Number (19.1%), while Tongyi achieves its highest accuracy on Named Entity (23.8%) yet collapses on Phrase/Clause (7.9%), a divergence that aggregate rankings would obscure. The consistent difficulty ordering across all systems confirms that answer type reflects an intrinsic query property, constituting an independent diagnostic axis absent from prior benchmarks.

Performance across Different Reasoning Hops. As shown in Figure 7, all agents exhibit consistent accuracy degradation as reasoning hops increase, dropping from roughly 40% at 1–3 hops to below 29% at 10+, confirming that multi-hop reasoning over live web content remains a critical bottleneck.

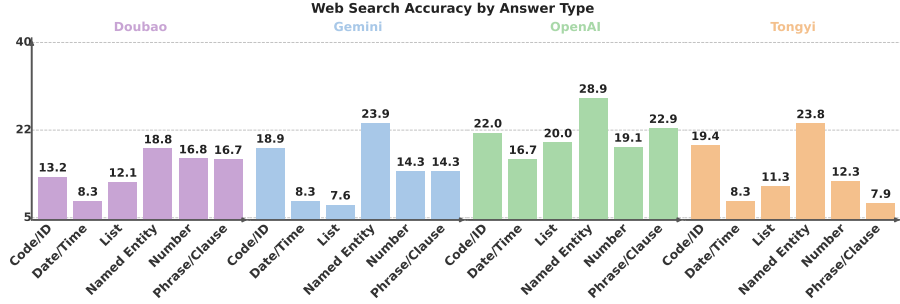


Figure 6: Accuracy (%) of commercial systems broken down by answer type.

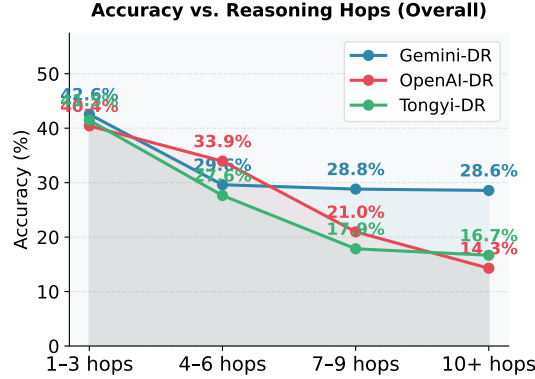


Figure 7: Accuracy vs. reasoning hops on commercial DeepSearch systems. All agents exhibit consistent degradation as hop count increases.

5 Conclusion

We introduced WebReal, a benchmark grounded in authentic user requests that exposes a structural misalignment in existing evaluation: prior benchmarks collapse onto disambiguation (>64% T5), while document extraction and computation are dominant in real user needs yet virtually absent from current benchmarks. Evaluation across 11 agents reveals that the best system achieves only 22.9%, failures polarize into over-persistence and under-persistence, and parameter scaling provides no remedy. These results suggest that progress requires operation-type-specific trajectory optimization rather than larger models or more search calls, and we hope future work will develop training and evaluation strategies with finer-grained coverage across all operation types.

Ethics Statement

WebReal is constructed from anonymized user queries collected from a commercial search platform. All queries underwent a rigorous filtering process to remove personally identifiable information, sensitive content, and any data that could be traced back to individual users prior to annotation and release. Human annotators involved in the labeling pipeline were compensated fairly and worked under clearly defined guidelines. No annotator was asked to engage with content that could cause psychological harm.

The benchmark is intended solely for academic research on agentic web search evaluation. We do not foresee direct misuse risks, as WebReal contains only information-seeking questions and factual answers drawn from publicly accessible web sources.

LLM Usage Statement

Large language models were used in this work in the following capacities: data construction (query grounding, rewriting, and answer generation via Gemini), automated evaluation (LLM-as-judge for answer correctness assessment), and minor assistance in paper writing.

(grammar and phrasing). All research ideas, experimental design, taxonomy construction, and conclusions are entirely the authors’ own. All model-generated outputs were verified by human annotators before inclusion in the benchmark.

References

- Zheng Chu, Xiao Wang, Jack Hong, Huiming Fan, Yuqi Huang, Yue Yang, Guohai Xu, Chenxiao Zhao, Cheng Xiang, Shengchao Hu, Dongdong Kuang, Ming Liu, Bing Qin, and Xing Yu. Redsearcher: A scalable and cost-efficient framework for long-horizon search agents, 2026. URL <https://arxiv.org/abs/2602.14234>.
- Brenda Dervin. An overview of sense-making research: Concepts, methods and results to date. In *Annual Meeting of the International Communication Association*, Dallas, TX, 1983.
- Nikita Gupta, Riju Chatterjee, Lukas Haas, Connie Tao, Andrew Wang, Chang Liu, Hidekazu Oiwa, Elena Gribovskaya, Jan Ackermann, John Blitzer, Sasha Goldshtein, and Dipanjan Das. Deepsearchqa: Bridging the comprehensiveness gap for deep research agents, 2026. URL <https://arxiv.org/abs/2601.20975>.
- Satyapriya Krishna, Kalpesh Krishna, Anhad Mohanane, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation, 2025. URL <https://arxiv.org/abs/2409.12941>.
- Gary Marchionini. Exploratory search: From finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006. doi: 10.1145/1121949.1121979.
- Thanh Pham, Nguyen Nguyen, Pratibha Zunjare, Weiyuan Chen, Yu-Min Tseng, and Tu Vu. Sealqa: Raising the bar for reasoning in search-augmented language models. *arXiv preprint arXiv:2506.01062*, 2025.
- ByteDance Seed, :, Jiaze Chen, Tiantian Fan, Xin Liu, Lingjun Liu, Zhiqi Lin, Mingxuan Wang, Chengyi Wang, Xiangpeng Wei, Wenyuan Xu, Yufeng Yuan, Yu Yue, Lin Yan, Qiyang Yu, Xiaochen Zuo, Chi Zhang, Ruofei Zhu, Zhecheng An, Zhihao Bai, Yu Bao, Xingyan Bin, Jiangjie Chen, Feng Chen, Hongmin Chen, Riwei Chen, Liangqiang Chen, Zixin Chen, Jinsong Chen, Siyan Chen, Kaiyuan Chen, Zhi Chen, Jin Chen, Jiecao Chen, Jinxin Chi, Weinan Dai, Ning Dai, Jiahui Dai, Shihan Dou, Yantao Du, Zhengyin Du, Jianhui Duan, Chen Dun, Ting-Han Fan, Jiazhan Feng, Junda Feng, Ziyuan Feng, Yuwei Fu, Wenqi Fu, Hanjie Fu, Hao Ge, Hongyi Guo, Mingji Han, Li Han, Wenhao Hao, Xintong Hao, Qianyu He, Jerry He, Feng He, Wen Heng, Zehua Hong, Qi Hou, Liang Hu, Shengding Hu, Nan Hu, Kai Hua, Qi Huang, Ziyue Huang, Hongzhi Huang, Zihao Huang, Ting Huang, Wenhao Huang, Wei Jia, Bin Jia, Xiaoying Jia, Yuhua Jiang, Haobin Jiang, Ziheng Jiang, Kaihua Jiang, Chengquan Jiang, Jianpeng Jiao, Xiaoran Jin, Xing Jin, Xunhao Lai, Zheng Li, Xiang Li, Liyi Li, Hongkai Li, Zheng Li, Shengxian Wan, Ya Wang, Yunshui Li, Chenggang Li, Niuniu Li, Siyu Li, Xi Li, Xiao Li, Aoyan Li, Yuntao Li, Nianning Liang, Xinnian Liang, Haibin Lin, Weijian Lin, Ye Lin, Zhicheng Liu, Guanlin Liu, Guanlin Liu, Chenxiao Liu, Yan Liu, Gaohong Liu, Juncal Liu, Chundian Liu, Deyi Liu, Kaibo Liu, Siyao Liu, Qi Liu, Yongfei Liu, Kang Liu, Gan Liu, Boyi Liu, Rui Long, Weiqiang Lou, Chenwei Lou, Xiang Luo, Yao Luo, Caiping Lv, Heyang Lv, Bole Ma, Qianli Ma, Hongzhi Ma, Yiyuan Ma, Jin Ma, Wenchang Ma, Tingting Ma, Chen Mao, Qiyang Min, Zhe Nan, Guanghan Ning, Jinxiang Ou, Haojie Pan, Renming Pang, Yanghua Peng, Tao Peng, Lihua Qian, Lihua Qian, Mu Qiao, Meng Qu, Cheng Ren, Hongbin Ren, Yong Shan, Wei Shen, Ke Shen, Kai Shen, Guangming Sheng, Jinlong Shi, Wenlei Shi, Guang Shi, Shuai Shuai Cao, Yuxin Song, Zuquan Song, Jing Su, Yifan Sun, Tao Sun, Zewei Sun, Borui Wan, Zihan Wang, Xiaohui Wang, Xi Wang, Shuguang Wang, Jun Wang, Qinlong Wang, Chenyuan Wang, Shuai Wang, Zihan Wang, Changbao Wang, Jiaqiang Wang, Shihang Wang, Xuwu Wang, Zaiyuan Wang, Yuxuan Wang, Wenqi Wang, Taiqing Wang, Chengzhi Wei, Houmin Wei, Ziyun Wei, Shufa Wei, Zheng Wu, Yonghui Wu, Yangjun Wu, Bohong Wu, Shuang Wu, Jingqiao Wu, Ning Wu, Shuangzhi Wu, Jianmin Wu, Chenguang Xi, Fan Xia, Yuqiao Xian, Liang Xiang, Boren Xiang, Bowen Xiao, Zhen Xiao, Xia Xiao, Yongsheng Xiao, Chao Xin, Shulin Xin, Yuwen Xiong, Jingjing Xu, Ziwen Xu, Chenyin Xu, Jiayi Xu, Yifan Xu,

- Wei Xu, Yufei Xu, Shikun Xu, Shipeng Yan, Shen Yan, Qingping Yang, Xi Yang, Tianhao Yang, Yuehang Yang, Yuan Yang, Ximing Yang, Zeyu Yang, Guang Yang, Yifan Yang, Xuesong Yao, Bairen Yi, Fan Yin, Jianian Yin, Ziqiang Ying, Xiangyu Yu, Hongli Yu, Song Yu, Menghan Yu, Huan Yu, Siyu Yuan, Jun Yuan, Yutao Zeng, Tianyang Zhan, Zheng Zhang, Yun Zhang, Mofan Zhang, Wang Zhang, Ru Zhang, Zhi Zhang, Tianqi Zhang, Xinyi Zhang, Zhexi Zhang, Sijun Zhang, Wenqiang Zhang, Xiangxiang Zhang, Yongtao Zhang, Yuyu Zhang, Ge Zhang, He Zhang, Yue Zhang, Renjie Zheng, Ningxin Zheng, Zhuolin Zheng, Yaowei Zheng, Chen Zheng, Xiaoyun Zhi, Wanjun Zhong, Cheng Zhong, Zheng Zhong, Baoquan Zhong, Xun Zhou, Na Zhou, Huan Zhou, Hang Zhu, Defa Zhu, Wenjia Zhu, and Lei Zuo. Seed1.5-thinking: Advancing superb reasoning models with reinforcement learning, 2025. URL <https://arxiv.org/abs/2504.13914>.
- Zhengwei Tao, Jialong Wu, Wenbiao Yin, Junkai Zhang, Baixuan Li, Haiyang Shen, Kuan Li, Liwen Zhang, Xinyu Wang, Yong Jiang, et al. Webshaper: Agentically data synthesizing via information-seeking formalization. *arXiv preprint arXiv:2507.15061*, 2025.
- Robert S. Taylor. Question-negotiation and information seeking in libraries. *College & Research Libraries*, 29(3):178–194, 1968.
- Tongyi DeepResearch Team. Tongyi deepresearch: A new era of open-source ai researchers. <https://github.com/Alibaba-NLP/DeepResearch>, 2025.
- Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.
- Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Gang Fu, Yong Jiang, et al. Webdancer: Towards autonomous information seeking agency. *arXiv preprint arXiv:2505.22648*, 2025a.
- Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and Fei Huang. Webwalker: Benchmarking LLMs in web traversal. In *Workshop on Reasoning and Planning for Large Language Models*, 2025b. URL <https://openreview.net/forum?id=cVI9lAfkUk>.
- Ziyi Xia, Kun Luo, Hongjin Qian, and Zheng Liu. Open data synthesis for deep research, 2025. URL <https://arxiv.org/abs/2509.00375>.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023. URL <https://arxiv.org/abs/2210.03629>.
- Peilin Zhou, Bruce Leon, Xiang Ying, Can Zhang, Yifan Shao, Qichen Ye, Dading Chong, Zhi Hua Jin, Chenxuan Xie, Meng Cao, Yu Gu, Sixin Hong, Jing Ren, Jian Chen, Chao Liu, and Yining Hua. Browsecomp-zh: Benchmarking web browsing ability of large language models in chinese. *ArXiv*, abs/2504.19314, 2025. URL <https://api.semanticscholar.org/CorpusID:278165208>.

A Data Construction

A.1 Filtering Information-Seeking Requests

Given the breadth of intent types present in user requests, we apply a filtering process to isolate genuine information-seeking queries that require web retrieval before benchmark construction.

Web Retrieval Intent Filtering. Rather than enumerating all possible non-information-seeking categories, we adopt a binary classification approach: each raw query $q \in \mathcal{Q}_{\log}$ is classified as either **Information Seeking** or **Non-Information Seeking** using a prompted LLM classifier (The classification prompt is shown in Fig. 8). A query is classified as **Information Seeking** if and only if its resolution is both necessary and sufficient to require retrieving factual content from external web sources, with no reliance on user-provided documents, symbolic computation, or parametric knowledge as a substitute. Concretely, a query is discarded if it can be resolved by any of the following: (i) operating on user-provided content such as files, code, or documents (e.g., *"Fix the bug in my script"*, *"Summarise this PDF"*); (ii) pure mathematical or logical derivation that requires no external facts (e.g., *"Solve this integral"*, *"Prove this theorem"*); (iii) code generation or software engineering tasks whose resolution depends on programming knowledge rather than retrieved information (e.g., *"Write a function to sort a list"*); or (iv) subjective or opinion-based requests that admit no verifiable factual answer retrievable from the web (e.g., *"What do you think of my idea?"*). Only queries whose resolution is contingent on retrieving specific, externally hosted factual content are retained for subsequent pipeline stages.

Answerability Pre-screening. Retained queries are further screened to remove cases that are structurally incompatible with benchmark construction regardless of intent, including purely temporal queries whose answer changes continuously and cannot be pinned to a stable ground truth (e.g., *"What is the current Bitcoin price?"*), hyper-local or private queries whose answer depends on information unavailable on the public web (e.g., *"What time does my local pharmacy close?"*), and ambiguous queries where grounding via web context fails to reduce $|\mathcal{A}(q \mid \mathcal{C}(q))|$ to a singleton even after augmentation, as defined in Equation 1.

Binary Intent Judgement Prompt

You are an intent classifier for a web retrieval benchmark.

Your task is to determine whether a given user query requires accessing external web content to be resolved.

(1). A query requires web retrieval if it asks about real-world facts, current events, specific entities, domain-specific knowledge, or any information that cannot be answered from general parametric knowledge alone and necessitates consulting external sources such as websites, documents, or databases.

(2). A query does NOT require web retrieval if it can be fully resolved through logical reasoning, mathematical computation, code generation, manipulation of user-provided content, or general world knowledge that does not depend on specific up-to-date or source-specific information.

Question: {question}

Respond with Yes if the query requires web retrieval, or No if it does not. Output the label only, with no explanation or additional text.

Answer

Figure 8: Intent Filtering prompt.

C Construction Pipeline Statistics and Validation

C.1 Stage-by-Stage Construction Audit

The WebReal construction pipeline is an LLM-assisted and human-verified process. Human annotators participate in verifying query validity, evidence sufficiency, final answers, operation labels, and overall sample quality at each stage. Table 5 reports the inputs, outputs, retention rates, and human validation scope at each stage.

B Taxonomy of failures

We classify the failures of models into six types, the detailed explanations are shown in Table 4.

Error Type	Description
Ineffective Loop	The agent repeatedly issues semantically equivalent queries that retrieve no new information, cycling without progress until the context budget is exhausted.
Premature Termination	The agent exits the search loop before gathering sufficient evidence, producing an empty or unsupported answer despite the information being retrievable.
Poor Query Formation	The agent constructs structurally or semantically ill-formed queries that fail to surface the target content, preventing deeper retrieval coverage regardless of search volume.
Context Limit	The trajectory is forcibly terminated upon reaching the maximum context length before a valid answer can be synthesized, typically caused by verbose or redundant intermediate observations.
Redundant Tool Call	The agent invokes the same tool with identical or near-identical inputs across multiple steps, yielding duplicate observations that consume context without advancing the search.
Insufficient Coverage	The agent retrieves partial evidence from some required sources but fails to access all pages or documents necessary to assemble a complete and correct answer.

Table 4: Taxonomy of failure modes observed in open-source agentic search trajectories. Each error type is mutually exclusive and collectively exhaustive within the annotation scheme; a single failed trajectory is assigned its *primary* failure mode based on the dominant pattern in the search log.

Stage	# In	# Out	Retention	Human Validation Scope
Φ_1 : Grounding & Augmentation	~20k	3,210	~16%	Validity and privacy screening (sampled)
Φ_2 : Entity Obfuscation	3,210	2,850	88.8%	Intent preservation check (sampled)
Φ_3 : Annotation & Voting	2,850	1,420	49.8%	100% checked for correctness and format
Φ_4 : Answer Reverification	1,420	1,180	83.1%	100% checked for evidence sufficiency
Φ_5 : Difficulty Filtering	1,180	507	43.0%	Final quality control and labeling (100%)

Table 5: Construction audit showing human validation coverage at each pipeline stage. Stages Φ_3 through Φ_5 receive 100% human review.

The overall retention rate from raw requests to the final benchmark is approximately 2.5%, reflecting the stringent multi-stage quality control. The primary sources of removal are non-information-seeking queries in Φ_1 (84% removed), inter-model disagreement in Φ_3 (50% removed), and oracle-based difficulty filtering in Φ_5 (57% removed).

C.2 Intent Preservation Audit

To verify that the multistage transformations do not produce an artificial hard set disconnected from genuine user needs, we conducted an intent preservation audit. Three expert annotators independently reviewed a random subset of 100 final queries against their corresponding raw requests, rating each pair on a 3-point scale: *Fully Preserved*, *Partially Preserved*, or *Intent Altered*. Results are reported in Table 6.

Rating	Proportion
Fully Preserved	82%
Partially Preserved (constraints added, core intent intact)	12%
Intent Altered	6%

Table 6: Intent preservation audit results (n=100). 94% of final queries preserve the core information-seeking intent.

Table 7 provides representative examples of the transformation from raw requests to final benchmark queries.

Raw User Request	Final Benchmark Query	Operation Type
“Who runs the largest AI lab?”	“As of Q1 2026, who is the CEO of the organization that developed GPT-4?”	T1: Direct Retrieval
“JFK flights”	“Which airlines operate direct flights from the major hub serving New York to LHR?”	T2: Chain Traversal
“latest iPhone price”	“What is the official retail price of the iPhone 16 Pro Max (256GB) listed on Apple’s US online store as of March 2026?”	T3: Document Extraction
“how much does shipping cost to Europe”	“For a 2kg package from Shenzhen to Berlin via DHL Express, what is the total declared shipping cost including fuel surcharge as of Q1 2026?”	T4: Computation

Table 7: Representative transformation examples from raw requests to final benchmark items, with assigned operation types.

C.3 Complexity Augmentation Formalism

The complexity augmentation constraint (Eq. 2) requiring multiple evidence sources is applied over the *verified reference evidence set* as an operational construction criterion. It does not require the enumeration of all possible valid trajectories. In practice, the construction team collects a reference evidence set for each query and verifies that it contains at least the required number of independent sources. Queries failing this threshold are either augmented with additional constraints or discarded. Table 8 reports the distribution of source counts in the final benchmark.

	Min	Median	Mean	Max
# Evidence sources per query	3	7	8.2	24
# Unique domains per query	2	4	4.6	12

Table 8: Distribution of evidence source counts in the final WebReal benchmark (n=507).

D Quality Control and Reliability

D.1 Oracle Sensitivity Analysis

In Stage Φ_5 , Gemini is used solely for difficulty filtering to remove trivial cases (i.e., queries answerable without web search). It does not generate gold answers or determine the model rankings reported in our experiments. To assess whether this single oracle introduces model-specific bias, we performed an oracle sensitivity audit using an alternative strong model (GPT-4o) on a subset of 200 queries. Table 9 reports the comparison.

Metric	Value	Interpretation
Retained-set overlap (Jaccard)	86.5%	High agreement on nontriviality
Max operation-type shift	1.8 pp	Negligible distributional change
Max domain shift	1.4 pp	Negligible distributional change
Rank correlation (Spearman ρ)	0.97	Model rankings preserved

Table 9: Oracle sensitivity audit comparing Gemini and GPT-4o as difficulty filters (n=200).

These results indicate that difficulty filtering robustly identifies nontrivial queries independent of the specific oracle model used.

D.2 Inter-Annotator Agreement

Each operation-type label and correctness verification task was independently performed by two trained annotators following a detailed annotation guideline. Disagreements were resolved by a third senior annotator through discussion. Table 10 reports the agreement statistics.

Annotation Task	# Examples	# Annotators	Raw Agreement	Cohen’s κ
Operation-type labeling	507	2 (+1 adjudicator)	89%	0.85
Correctness verification	507	2 (+1 adjudicator)	92%	0.88
Domain classification	507	2	94%	0.91

Table 10: Inter-annotator agreement statistics. All tasks achieve substantial to near-perfect agreement ($\kappa > 0.80$).

D.3 LLM-as-Judge Calibration

To validate the reliability of the LLM-as-judge component used for automated evaluation, we conducted a human adjudication study on a representative subset of 150 model responses stratified by answer type. Table 11 reports the agreement between the LLM judge and human experts.

Answer Type	# Examples	Agreement	Precision / Recall
Exact match (entity, number)	62	97.8%	98.4% / 97.2%
Short phrase	48	94.5%	95.0% / 93.8%
Multi-part answer	40	92.5%	93.8% / 91.2%
Overall	150	95.3%	96.1% / 94.5%

Table 11: LLM-as-judge calibration against human adjudication, stratified by answer type.

Agreement is highest for exact-match answers and slightly lower for multi-part answers, where partial credit decisions introduce some subjectivity.

D.4 Per-Operation and Per-Domain Statistics

Table 12 provides the exact counts and percentages of queries across operation types and domains in the final benchmark. Bootstrap stability analysis (95% confidence intervals over 1,000 resamples) confirms that the performance rankings reported in our experiments are statistically stable despite the focused size of the benchmark.

Operation Type	Count	%	Top Domains	Count	%
T1: Direct Retrieval	86	17.0%	Technology	85	16.8%
T2: Chain Traversal	90	17.8%	Finance & Business	78	15.4%
T3: Document Extraction	80	15.8%	Entertainment	65	12.8%
T4: Computation	79	15.6%	Health & Medicine	58	11.4%
T5: Disambiguation	96	18.9%	Science	52	10.3%
T6: Temporal Linking	76	15.0%	Others (10 domains)	169	33.3%
Total	507	100%	Total	507	100%

Table 12: Per-operation and per-domain distribution in WebReal. Shannon entropy of operation types is 2.38 bits (maximum 2.58), indicating near-uniform coverage.

E Benchmark Governance and Reproducibility

E.1 Contamination Mitigation

WebReal employs multiple structural safeguards against test-set contamination:

- **Entity obfuscation:** Surface entities are systematically replaced, preventing keyword-matching shortcuts from pre-training memorization.
- **Temporal grounding:** Queries are anchored to specific time windows, rendering cached parametric knowledge insufficient.
- **Versioned releases:** The benchmark follows a versioned release protocol, enabling periodic refresh of a subset of queries to counteract data leakage over time.

E.2 Evaluation Protocol

Web-agent evaluation is sensitive to the search engine, timing, geolocation, and agent configurations. Table 13 specifies the standardized evaluation settings used in all experiments reported in this paper.

Parameter	Setting
Search API	Bing Search API
Context limit	128k tokens
Maximum tool-call turns	30
Browsing timeout per page	10 seconds
Decoding temperature	0.0 (greedy)
Region	US-East
Evaluation timestamp	January 2026

Table 13: Standardized evaluation protocol. All agents are evaluated under identical environmental conditions.

E.3 Data Governance and Privacy

The raw user queries originate from real-world search logs. The released benchmark contains only sanitized and transformed queries that have passed through automated PII detection and removal, entity obfuscation, and human review for sensitive content exclusion. Raw queries and user identifiers are excluded from all public releases.