DEMYSTIFYING DEEP SEARCH: A HOLISTIC EVALUATION WITH HINT-FREE MULTI-HOP QUESTIONS AND FACTORISED METRICS

Anonymous authors
Paper under double-blind review

ABSTRACT

RAG (Retrieval-Augmented Generation) systems and web agents are increasingly evaluated on multi-hop deep search tasks, yet current practice suffers from two major limitations. First, most benchmarks leak the reasoning path in the question text, allowing models to follow surface cues rather than discover reasoning chains autonomously. Second, evaluation is typically reduced to a single pass rate, which collapses diverse behaviors into one score and obscures whether failures stem from inadequate search, poor knowledge use, or inappropriate refusal. To address these issues, we present **WebDetective**, a benchmark of hint-free multi-hop questions paired with a controlled Wikipedia sandbox that ensures full traceability of model actions, and a holistic evaluation framework that separates search sufficiency, knowledge utilization, and refusal behavior. Our evaluation of 25 state-of-the-art models reveals systematic weaknesses across all architectures: models struggle with knowledge utilization despite having sufficient evidence and demonstrate near-absent appropriate refusal when evidence is lacking. These patterns expose a fundamental gap—today's systems excel at executing given reasoning paths but fail when required to discover them. We develop an agentic workflow **EvidenceLoop** that explicitly targets the challenges our benchmark identifies, incorporating verification loops and systematic evidence tracking that improve both search and synthesis capabilities. This baseline demonstrates that **WebDetective**'s diagnostic framework can guide concrete architectural improvements, establishing our benchmark as a critical tool for developing genuinely autonomous reasoning systems rather than pattern-following agents.

1 INTRODUCTION

Web agents—systems that autonomously navigate and extract information from the internet—have emerged as critical tools for extending language models beyond their parametric knowledge. These agents must solve complex information-seeking tasks by strategically combining external search with internal knowledge, searching across multiple sources, and synthesising dispersed information Nakano et al. (2021) Yao et al. (2023). Among the evaluation scenarios for these systems, *deep search* tasks stand out as particularly challenging and important. Deep search requires finding specific, hidden facts or entities through deep reasoning, multi-step inference, and noise filtering—addressing the "I can't find it" problem that challenges even skilled human searchers Wong et al. (2025). Unlike shallow retrieval where information is directly stated, deep search demands sophisticated exploration strategies to uncover information that is not immediately accessible, with outputs typically being single facts or small entity sets that must exactly match ground truth, making evaluation unambiguous while maintaining high difficulty.

However, we identify a critical but overlooked dimension in current deep search evaluation: the presence of various forms of *hinting* embedded in question formulation that fundamentally alters the nature of the search problem. As illustrated in fig. 1, classical multi-hop QA datasets like Hotpot QA Yang et al. (2018a) exhibit what we term **Path-Hinting (PH)**, where questions linguistically narrate the reasoning chain: "Who is the husband of the stepmother of the brother of Kane Cornes?" explicitly instructs the agent to first find Kane's brother, then the brother's stepmother, then her husband—effectively converting reasoning into execution. Recent benchmarks

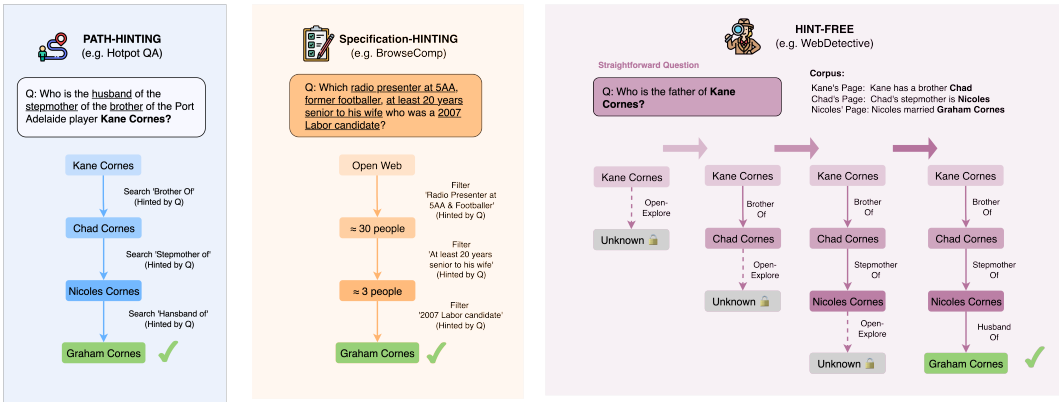

Figure 1: Comparison of different question formulations in multi-hop deep search. **Left:** Path-Hinting (PH) benchmarks such as HotpotQA embed the reasoning path directly in the question text, effectively reducing reasoning to execution. **Middle:** Specification-Hinting (SH) benchmarks such as BrowseComp obscure the target entity behind multiple attributes, testing filtering rather than autonomous exploration. **Right:** Our Hint-Free (HF) formulation in **WebDetective** removes both path and specification hints, requiring agents to autonomously discover reasoning chains within a controlled Wikipedia sandbox.

such as BrowseComp Wei et al. (2025) and WebShaper Tao et al. (2025) attempt to address this limitation through **Specification-Hinting (SH)**, where questions obscure the target entity behind multiple indirect attributes rather than naming it directly. For instance, instead of directly asking about 'Graham Cornes', the question specifies "Which radio presenter at 5AA, former footballer, at least 20 years senior to his wife who was a 2007 Labor candidate?"—creating enough constraints to uniquely identify the target through sophisticated filtering rather than exploratory reasoning. The widespread presence of these hints in existing benchmarks remains understudied and unaddressed, yet it fundamentally shapes what capabilities are actually being evaluated.

While Path-Hinting (PH) benchmarks are largely solved by modern systems, Specification-Hinting (SH) benchmarks remain challenging due to their demands for sustained exploration over long trajectories and filtering through deliberately obfuscated information. These benchmarks effectively test an agent's ability to maintain context, handle noise, and perform complex constraint satisfaction—important capabilities for robust web agents. However, they leave a different but equally critical aspect untested: *the ability to autonomously discover which connections matter, generate hypotheses about potential paths, and adaptively explore the information space without guidance*. When agents receive either an explicit path (PH) or a unique signature (SH), they operate with substantial scaffolding that may not be available in real-world scenarios. Furthermore, existing evaluations suffer from a critical limitation: they typically report only aggregate pass rates, collapsing diverse failure modes into a single metric. This obscures crucial distinctions—an agent that searches exhaustively but fails to connect evidence exhibits fundamentally different limitations than one that gives up prematurely or misuses its parametric knowledge. Without understanding these failure modes, it becomes hard to diagnose system weaknesses or guide improvements.

In this work, we introduce **WebDetective**, a benchmark that fundamentally rethinks hint-free deep search evaluation through the co-design of questions and their evaluation environment. First, we design **Hint-Free (HF) Multi-Hop** questions that provide neither path narration nor attribute fingerprints—straightforward questions like "Who is the father of Kane Cornes?" require agents to autonomously discover relevant contexts and reasoning chains. Second, and critically, we develop a controlled Wikipedia sandbox that prevents shortcuts by selectively revealing information only when agents follow the correct reasoning path. For instance, if answering a question requires connecting information through multiple intermediate facts, our sandbox ensures these connections cannot be bypassed—the agent must discover each link in the chain sequentially. This co-designed system creates an evaluation environment where we can guarantee what knowledge an agent must have discovered to succeed.

This co-design—where the sandbox enforces that agents must discover each step of the reasoning chain—uniquely enables a comprehensive two-level evaluation framework that provides precise attribution of failure modes in multi-hop reasoning. Because our sandbox guarantees that successful task completion requires finding and connecting all necessary intermediate facts, we can definitively separate knowledge sufficiency (from both search and parametric sources) from generation quality. When an agent succeeds, we know it must have discovered the complete reasoning chain; when it fails, we can pinpoint whether it stopped searching too early, found the right information but failed to connect it, or appropriately refused to answer when evidence was insufficient. Such fine-grained diagnostics are only possible because our benchmark's architecture ensures that there is only one path to the correct answer, and we can observe exactly how far along that path each agent progresses.

Additionaly, we further design an agentic workflow baseline that explicitly incorporates context retention, memory management, and verification steps, offering a first attempt at addressing the unique challenges posed by hint-free deep search. Through our diagnostic evaluation framework, we uncover fundamental brittleness in current systems when reasoning paths must be discovered rather than given, exposing critical gaps between existing capabilities and the requirements of genuine autonomous deep search.

2 THE **WEBDETECTIVE** BENCHMARK

2.1 HINT-FREE MULTI-HOP QUESTION ANSWERING

Existing multi-hop QA benchmarks Yang et al. (2018b); Chen et al. (2019) systematically embed hints h into their question formulations that fundamentally alter the search problem. We identify two prevalent types of hint embedding:

Path-Hinting (PH): The question linguistically encodes the reasoning chain, where $h_{PH} = \text{Encode}(\mathcal{R})$ directly reveals the reasoning structure. For example, in "Who is the husband of the stepmother of the brother of Kane Cornes?", the hint h_{PH} explicitly decomposes the reasoning into sequential steps: find brother $\rightarrow$ find stepmother $\rightarrow$ find husband. The agent's task reduces from discovering the reasoning path to merely executing the already-specified h_{PH}.

Specification-Hinting (SH): The question obscures the target entity behind excessive constraints, where $h_{SH} = \{s_1, s_2, ..., s_k\}$ progressively narrows the search space to a unique answer. For instance, "Which radio presenter at 5AA, former footballer, at least 20 years senior to his wife who was a 2007 Labor candidate?" provides constraints $h_{SH} = \{\text{radio presenter at 5AA}, \text{former footballer}, \text{20+ years senior to wife}, \text{wife was 2007 Labor candidate}\}$ that collectively fingerprint Graham Cornes. While this creates search challenges through constraint matching and noise filtering, the fundamental task becomes constraint satisfaction—locate any entity matching all specifications in h_{SH}—rather than discovering which connections matter for reasoning.

In contrast, we propose **Hint-Free (HF) Multi-Hop** question answering where $h = \emptyset$. Formally, given a question q and a knowledge corpus $\mathcal{C}$, an agent must find an answer a^* by discovering and composing a sequence of evidence pieces $\mathcal{E} = \{e_1, e_2, ..., e_n\}$ from $\mathcal{C}$. Each evidence piece e_i represents an atomic fact extracted from entity v_i's web page that contains related information to v_{i+1}, forming a reasoning chain $v_0 \rightarrow v_1 \rightarrow ... \rightarrow v_n$ where v_0 is the starting entity (mentioned in q) and v_n yields the answer a^*. The reasoning function $\mathcal{R} : \mathcal{E} \rightarrow a^*$ defines how these evidence pieces must be composed—through logical inference, relationship transitivity, or domain-specific reasoning—to derive the final answer from the collected facts.

The key distinction is that HF Multi-Hop questions provide neither reasoning scaffolding nor excessive specifications. Questions like "Who is the father of Kane Cornes?" contain only the essential information needed to identify what is being asked, without revealing how to find it. For this example, answering requires discovering e_1 (Kane Cornes has brother Chad Cornes), e_2 (Chad's stepmother is Nicole Cornes), e_3 (Nicole's husband is Graham Cornes), and composing them via familial reasoning to derive a^* = Graham Cornes. Critically, the agent must independently discover both the evidence chain $\mathcal{E}$ and the reasoning function $\mathcal{R} : \mathcal{E} \rightarrow a^*$ without guidance from hints. This formulation captures a fundamental capability: given a straightforward information need, can an agent autonomously discover the reasoning structure required to find the answer?

2.2 THE CO-DESIGN PRINCIPLE

While hint-free questions eliminate linguistic scaffolding, we observe that question design alone is insufficient to ensure genuine multi-hop reasoning. In open corpora or live web environments, even well-designed hint-free questions permit shortcuts that bypass the intended reasoning chain. Consider our example "Who is the father of Kane?"—in Wikipedia or web search, direct co-occurrences of "Kane" and "Graham Cornes" may exist in unrelated contexts, or intermediate entities like "Chad Cornes" could be found through direct search, allowing agents to bypass the intended reasoning path.

This shortcut problem is fundamental: in any open corpus, both answers and intermediate entities are typically accessible through multiple paths. This accessibility makes it impossible to determine whether an agent genuinely discovered the reasoning chain or simply leveraged shortcuts, prior knowledge, or lucky searches.

To address this, we introduce a **co-designed evaluation system** where questions and their environment are jointly constructed to enforce reasoning path discovery. Our key insight is to create a controlled sandbox with selective entity masking. For a reasoning chain with entities $v_0 \to v_1 \to ... \to v_n$ (where v_0 appears in the question and v_n yields the answer), we **mask each intermediate entity** v_i everywhere in the corpus except on the Wikipedia page of v_{i-1}. Formally:

$$\boxed{\text{Entity } v_i \text{ is discoverable} \iff \text{agent visits page}(v_{i-1})}$$

This creates a strict sequential dependency where each step in the reasoning chain can only be accessed through the previous step.

Returning to our example: "Chad Cornes" is masked throughout the entire corpus except on Kane's Wikipedia page. An agent cannot find Chad through search or cross-references—it must visit Kane's page to discover that Kane has a brother Chad (e_1). Similarly, "Nicole Cornes" only appears on Chad's page, revealing she is Chad's stepmother (e_2). Finally, the connection to "Graham Cornes" exists only on Nicole's page, identifying him as her husband (e_3). The sandbox enforces that reaching the answer requires following the exact chain: Kane $\to$ Chad $\to$ Nicole $\to$ Graham.

This masking mechanism eliminates shortcuts and provides strong evaluation guarantees. When an agent succeeds, we know definitively that it discovered the complete reasoning chain by visiting each required page in sequence. When it fails, we can precisely diagnose where the breakdown occurred—did it never visit Kane's page (insufficient exploration)? Did it find Chad but fail to visit his page (failed to recognize relevance)? Or did it reach Nicole but couldn't extract the answer (synthesis failure)? This fine-grained attribution is only possible because our controlled environment ensures a unique, traceable path to each answer, transforming multi-hop QA evaluation from probabilistic assessment to deterministic verification.

2.3 BEYOND PASS RATES: A DIAGNOSTIC EVALUATION FRAMEWORK

Traditional multi-hop QA evaluation reduces agent performance to a single pass rate, obscuring the diverse failure modes that occur in complex reasoning tasks. An agent that searches exhaustively but fails to synthesize evidence exhibits fundamentally different limitations than one that refuses prematurely or hallucinates from parametric knowledge. Our co-designed sandbox, with its guaranteed unique reasoning paths, enables unprecedented diagnostic precision in distinguishing these failure modes.

We introduce a two-level evaluation framework that separates *knowledge sufficiency* from *generation quality*. First, we assess whether an agent possesses the requisite knowledge—either through successful search or parametric memory—to answer the question. Second, conditioned on knowledge sufficiency, we evaluate the agent's ability to either correctly synthesize an answer or appropriately refuse when evidence is insufficient. This decomposition reveals that seemingly similar pass rates can mask vastly different underlying capabilities.

Knowledge Discovery Metrics. We assess whether agents acquire necessary information through two complementary metrics. **Knowledge Sufficiency** determines if an agent possesses all required evidence $\mathcal{E} = \{e_1, ..., e_n\}$ for answering. We track which evidence the agent discovered through search by monitoring visited pages in our sandbox. For any missing evidence $e_i \notin \mathcal{E}_{\text{found}}$, we probe

the model's parametric knowledge with targeted queries (e.g., "Kane Cornes has brother ____?"). An instance is knowledge-sufficient when the agent has all evidence either from search or parametric memory. The **Search Score** extends this by crediting models that efficiently combine partial search with parametric knowledge—recognizing that if an entity discovered through search has a meaningful relationship to the answer stored in parametric memory, this represents legitimate reasoning that demonstrates efficient knowledge utilization.

Generation Quality Metrics. Given knowledge sufficiency assessment, we partition instances along two dimensions: knowledge-sufficient ($\mathcal{S}$) vs. insufficient ($\mathcal{I}$), and attempted answer ($\mathcal{A}$) vs. refusal ($\mathcal{R}$). This creates critical regions revealing different capabilities. **Good Refusal (GR)** measures appropriate abstention when lacking evidence through an F1 score (F1_{GR}) that balances precision and recall—high recall indicates the agent avoids hallucination by refusing most knowledge-insufficient cases, while high precision ensures refusals are justified rather than overcautious. **Knowledge Utilization (KU)** assesses synthesis of correct answers when evidence is available, also measured via F1 score (F1_{KU})—high recall means the agent leverages available evidence effectively, while high precision indicates that answer attempts are grounded in sufficient knowledge rather than speculation. These F1 formulations capture the complementary nature of both capabilities: an ideal agent achieves high scores in both metrics rather than trading one for the other.

We combine these into a unified **Generation Score**: $\text{GenScore} = \frac{\text{F1}_{\text{GR}}+\text{F1}_{\text{KU}}}{2} \cdot \text{KnowledgeScore}$. The knowledge sufficiency weighting is crucial—without it, models could game the evaluation by performing minimal search and refusing all questions, achieving high Good Refusal scores while providing no value. This design ensures models must demonstrate both effective evidence discovery and appropriate handling of that evidence.

Knowledge Degradation Analysis. For instances where models achieve knowledge sufficiency yet fail to generate correct answers, we conduct diagnostic tests to understand why evidence possession doesn't translate to correct synthesis. The **Knowledge Forget** test reveals when models cannot leverage parametric knowledge to fill gaps in the full question context, despite correctly answering individual knowledge probes. The **Lead-astray** test identifies when accumulated search context—failed attempts, irrelevant pages, exploration noise—disrupts the model's ability to synthesize answers it could produce from clean evidence alone.

Unlike simple pass rates that collapse diverse behaviors into a single number, our metrics provide actionable diagnostics: low Knowledge Scores reveal inadequate search strategies, poor Good Refusal indicates over-confident hallucination, weak Knowledge Utilization exposes synthesis failures, and high Knowledge Degradation rates pinpoint where models struggle to maintain coherence across extended reasoning chains. This diagnostic precision, enabled by our co-designed evaluation environment, illuminates the specific capabilities required for robust multi-hop reasoning. See Appendix A.1 for complete mathematical formulations.

2.4 Dataset Construction

To instantiate our hint-free multi-hop QA benchmark, we develop a systematic pipeline that transforms single-hop Wikipedia QA pairs into verified multi-hop reasoning chains while ensuring each hop is necessary for answering.

Source Data and Chain Discovery. We begin with a corpus of Wikipedia-based QA pairs where each question targets a specific paragraph on a Wikipedia page (the starting entity v_0) and has an answer that is another Wikipedia entity (v_n). These seed questions are designed to be unambiguous and simple, avoiding any linguistic hints about reasoning paths. To construct multi-hop chains, we first block the direct connection between v_0 and v_n, then perform breadth-first search (BFS) to find the shortest alternative path $v_0 \to v_1 \to ... \to v_n$ through Wikipedia's hyperlink graph. For each edge (v_i, v_{i+1}) in the discovered path, we extract the sentence e_i from v_i's Wikipedia page that contains the hyperlink to v_{i+1}, forming the evidence chain $\mathcal{E} = \{e_1, e_2, ..., e_n\}$.

Verification of Reasoning Necessity. Not all discovered paths constitute valid answers to the question—most arbitrary paths from v_0 to v_n through Wikipedia's link graph are completely unrelated to what the question asks. For instance, a path connecting two people through their universities and shared colleagues is irrelevant for a question asking about family relationships. We

implement a three-stage verification process using a strong language model (Qwen-3-235B in our implementation), denoted as $\text{LM}(\cdot)$ which takes text input and generates an answer:

1. **Parametric Inaccessibility**: We verify that $\text{LM}(q) \neq v_n$, ensuring the answer cannot be directly retrieved from the model's parametric memory without evidence.
2. **Evidence Sufficiency**: We confirm that $\text{LM}(q, \mathcal{E}) = v_n$, validating that the complete evidence chain enables correct answer generation.
3. **Evidence Necessity**: For each evidence piece e_i, we verify that $\text{LM}(q, \mathcal{E} \setminus \{e_i\}) \neq v_n$, ensuring every hop in the chain is required for reasoning. This ablation test eliminates questions where evidence pieces are redundant or where shortcuts exist.

Human Validation and Dataset Statistics. Questions passing automated verification undergo human annotation to ensure the question genuinely requires all evidence pieces, the evidence chain logically derives the answer without external knowledge, and no implicit hints about the reasoning structure exist. Our final **WebDetective** benchmark comprises 200 verified questions with diverse hop counts and question types. We provide detailed dataset statistics and additional validation details in Appendix C and D.

3 EXPERIMENTS

To address the unique challenges posed by hint-free multi-hop reasoning, we develop **EvidenceLoop**, an agentic workflow baseline that explicitly incorporates context retention, memory management, and verification steps to maintain reasoning coherence across extended search trajectories. Unlike standard ReAct implementations that can lose track of evidence across many search iterations, our workflow introduces structured mechanisms for tracking discovered entities, maintaining evidence chains, and verifying reasoning paths before answer generation. We provide a detailed description of the **EvidenceLoop** architecture, including its controller configuration, memory modules, and verification procedures in Appendix E.

We evaluate 25 state-of-the-art models with ReAct-style tool use capabilities, including those developed by OpenAI, Anthropic, Google, xAI, Alibaba, ByteDance, Zhipu AI, Moonshot AI, and High-Flyer. All models follow the ReAct paradigm Yao et al. (2023), interleaving reasoning, search actions, and observations within a controlled Wikipedia sandbox. Using **WebDetective** with 200 hint-free multi-hop questions (2–4 reasoning hops), models operate under limits of 40 tool calls and a 32K-token context window. Unless otherwise specified, we adopt a unified decoding configuration with `temperature` set to 0.6 and `top_p` set to 0.95. Performance is measured with six metrics: (1) *Knowledge Score*, sufficiency of knowledge acquisition; (2) *Search Score*, effectiveness of retrieval; (3) *Generation Score*, weighted F1 of Good Refusal and Knowledge Utilization; (4) *Good Refusal F1*, appropriateness of refusals without evidence; (5) *Knowledge Utilization F1*, synthesis accuracy given evidence; and (6) *Pass@1*, standard accuracy. In addition, we further analyse *Forget* and *Lead-astray* behaviours to probe knowledge degradation of LLMs for synthesising the final answer in section 3.2.2. For our proposed **EvidenceLoop**, we further configure the controller with `breadth=3` and `iteration=3`. table 1 presents comprehensive results across six key metrics for all ReAct baselines and our **EvidenceLoop**.

3.1 MAIN RESULTS

Frontier models are far from saturating the task. Even the strongest systems reach only $\sim$50% Pass@1 on our benchmark: O3-Pro tops out at 56.0%, while GPT-5 and Grok-4 both achieve 50.5%; Claude-Opus-4.1 is at 44.5%, and many others fall well below 40%. This illustrates the challenging nature of our benchmark, **WebDetective**.

Search, generation, and final accuracy are decoupled. High retrieval does not translate proportionally into better synthesis or Pass@1. For example, GPT-5 attains an 80.0% Search Score but only 23.21% Generation Score and 50.5% Pass@1; O3-Pro similarly has 78.0 Search but 20.86 Generation (56.0% Pass@1). Conversely, Grok-4 achieves the highest Generation Score (34.71) with 77.5 Search and 50.5% Pass@1, while Qwen3-235B-Thinking posts 72.0% Search yet just 11.15%

Table 1: Comparison of 25 state-of-the-art models with ReAct-style tool use capabilities. Metrics cover Knowledge Discovery (Knowledge Sufficiency, Search Score), Generation Quality (Generation Score, Good Refusal F1, Knowledge Utilisation F1), Knowledge Degradation (Forget, Lead-astray), and Pass@1. **Bold** values denote best results: higher is better for Knowledge Discovery, Generation Quality, and Pass@1, while lower indicates greater robustness for Knowledge Degradation.

Provider	Model	Knowledge Discovery		Generation Quality			Knowledge Degradation		Pass@1
		Knowledge Suff. (%)	Search Score (%)	Generation Score (%)	Good Refusal F1 (%)	Knowledge Util. F1 (%)	Forget (%)	Lead-astray (%)	(%)
OpenAI	GPT-5 OpenAI (2025a)	**79.00**	**80.00**	23.21	8.89	49.58	**17.72**	32.91	50.50
	GPT-5-Chat OpenAI (2025a)	58.00	59.50	15.74	26.23	28.05	47.41	31.90	29.50
	O3-Pro OpenAI (2025c)	71.00	78.00	20.86	9.37	49.40	21.83	25.35	**56.00**
	O3 OpenAI (2025c)	70.00	76.00	18.29	3.29	48.97	24.29	24.29	53.50
	O3-Mini OpenAI (2025c)	48.50	57.00	9.10	21.05	16.48	46.39	42.27	21.50
	O4-Mini OpenAI (2025d)	68.00	72.00	12.69	19.75	17.56	27.94	59.56	21.00
	GPT-OSS-120B OpenAI (2025b)	16.00	23.50	2.75	23.59	10.73	100.00	**0.00**	24.00
Anthropic	Claude-Opus-4.1 Anthropic (2025)	74.00	76.50	28.53	28.57	48.54	27.03	31.08	44.50
	Claude-Opus-4-Think Anthropic (2025)	68.00	73.50	21.00	30.53	31.23	43.38	32.35	29.00
	Claude-Sonnet-4-Think Anthropic (2025)	66.50	73.50	26.19	34.59	44.19	45.11	21.80	38.50
Google	Gemini-2.5-Pro Google DeepMind (2025)	65.50	73.00	11.64	10.87	24.68	44.27	35.11	28.50
	Gemini-2.5-Flash-Think Google DeepMind (2025)	59.00	64.50	16.79	40.56	16.35	57.63	35.59	17.50
xAI	Grok-4 xAI (2025)	74.00	77.50	**34.71**	37.63	**56.19**	23.65	27.70	50.50
Alibaba	Qwen3-235B-Think Yang et al. (2025)	72.50	72.00	11.15	6.56	24.19	63.45	19.31	21.50
	Qwen3-30B-Think Yang et al. (2025)	56.50	59.00	7.25	12.51	13.16	79.65	16.81	12.50
	Tongyi-DeepResearch Tongyi DeepResearch Team (2025)	53.50	57.50	4.20	0.00	15.69	43.93	41.12	18.50
ByteDance	Doubao-1.6-Think ByteDance Seed Team (2025)	64.00	68.50	19.24	42.03	18.11	49.22	39.84	16.00
	Doubao-1.6-Flash ByteDance Seed Team (2025)	54.50	57.50	20.00	**53.95**	19.46	68.81	21.10	13.50
Zhipu AI	GLM-4.5-Inner Zhipu AI Team (2025)	63.50	67.50	22.19	34.79	35.09	25.98	40.16	33.50
	GLM-4.5-Air-Inner Zhipu AI Team (2025)	55.50	60.50	12.31	26.39	17.97	44.14	40.54	19.00
Moonshot AI	Kimi-K2-0711 Moonshot AI (2025)	54.50	59.00	9.72	16.36	19.31	43.12	36.70	23.50
	Kimi-K2-0905 Moonshot AI (2025)	53.00	55.00	13.17	28.79	20.89	49.06	33.96	24.00
DeepSeek	DeepSeek-R1 DeepSeek-AI et al. (2025)	61.50	65.50	10.57	18.81	15.55	37.40	51.22	20.00
	DeepSeek-V3.1-Think DeepSeek-AI et al. (2024)	61.50	56.50	13.62	27.97	16.34	44.72	44.72	17.00
	DeepSeek-V3.1 DeepSeek-AI et al. (2024)	55.50	58.50	16.31	36.49	22.23	28.83	50.45	24.50
Our Team	EvidenceLoop	61.50	62.50	12.61	17.98	23.79	41.46	41.46	25.00

Table 2: Emergent model profiles from metric interplay analysis.

Profile	Metric Pattern			Pass@1	Example Models	Failure Mode
	Knowledge	Refusal	Utilization			
Powerful but Overconfident	High	Low	High	50-56%	GPT-5, O3-Pro, O3	Hallucination from overconfidence
Well-Calibrated Elite	High	Med	High	44-51%	Grok-4, Claude-Opus-4.1	Minor: unnecessary caution
Synthesis Bottleneck	High	Low	Low	18-22%	Qwen3-235B, Tongyi-DR	Cannot compose multi-hop reasoning
Conservative Middle	Med	Med	Med	29-39%	Claude-Sonnet-4, GLM-4.5	Under-utilizes capabilities
Weak and Confused	Med	Low	Low	20-22%	O4-Mini, DeepSeek-R1	Poor synthesis + poor calibration
Self-Aware of Weakness	Low	High	Low	13-18%	Doubao variants, Gemini-Flash	Comprehensive inability (appropriate)
Ideal (Unachieved)	High	High	High	–	None	None - optimal behavior

Generation and 21.5% Pass@1. These gaps indicate that *information synthesis, not just retrieval, is a key bottleneck*.

Refusal ability is underdeveloped. Good-refusal performance is generally low: the best we observe is 53.95% F1 (Doubao-1.6-Flash). Many frontier models underperform markedly—e.g., GPT-5 (8.89%), O3-Pro (9.37%), and O4-Mini (19.75%)—and even strong generalists like Claude-Opus-4.1 remain modest (28.57%). This highlights *weak calibrated abstention* when evidence is insufficient.

3.2 Analysis

3.2.1 Understanding Model Failure Modes Through Metric Patterns

To better understand the diverse failure modes in multi-hop reasoning, we analyze the interplay between our three core metrics: Knowledge Sufficiency (ability to gather evidence), Good Refusal F1 (calibration of uncertainty), and Knowledge Utilization F1 (synthesis capability). Rather than examining metrics in isolation, we investigate how their combinations reveal distinct behavioral profiles.

We categorize performance using empirically-derived thresholds: Knowledge Sufficiency (High: $> 70\%$, Medium: 60-70%, Low: $< 60\%$), Good Refusal F1 (High: $> 40\%$, Medium: 25-40%, Low: $< 25\%$), and Knowledge Utilization F1 (High: $> 45\%$, Medium: 25-45%, Low: $< 25\%$).

Analyzing all 23 models, we observe that they cluster into six distinct profiles based on these metric combinations, with certain theoretically plausible patterns notably absent from the empirical data.

table 2 presents our taxonomy. The **Powerful but Overconfident** profile (GPT-5, O3-Pro, O3) achieves the highest pass rates (50-56%) through strong evidence gathering and synthesis, but exhibits dangerous overconfidence with refusal rates below 10% despite 21-30% knowledge insufficiency. These models prefer hallucination over admission of uncertainty. In contrast, the **Well-Calibrated Elite** (Grok-4, Claude-Opus-4.1) achieve similar knowledge sufficiency and utilization but maintain moderate refusal rates (29-38%), demonstrating that strong capabilities need not preclude epistemic awareness—though this calibration costs approximately 5-6% in pass rate.

The **Synthesis Bottleneck** profile reveals a critical failure mode: models like Qwen3-235B-Thinking achieve high knowledge sufficiency (72.5%) but catastrophically fail at synthesis ($< 25\%$ utilization). Despite possessing evidence, they cannot compose multi-hop reasoning chains, yet their low refusal rates indicate unawareness of this limitation. The **Conservative Middle** models (Claude-Sonnet-4-Think, GLM-4.5-Inner) exhibit consistent mediocrity across all metrics, suggesting excessive caution—their moderate utilization (31-44%) despite reasonable knowledge gathering (63-68%) indicates they refuse even when capable of answering.

At the lower performance tiers, we observe a striking divergence in self-awareness. **Self-Aware of Weakness** models (Doubao variants, Gemini-2.5-Flash-Think) appropriately refuse in 40-54% of cases, correctly recognizing their limitations in both search and synthesis. Conversely, **Weak and Confused** models (O4-Mini, DeepSeek-R1) exhibit similar capability limitations but fail to recognize them, attempting answers despite 16-18% utilization rates.

Our analysis reveals three distinct failure modes in the multi-hop reasoning pipeline. *Search failure* affects 21-46% of attempts even in top models, indicating that evidence discovery remains challenging. *Synthesis failure* is more severe—even with sufficient knowledge, utilization rates peak at 56%, suggesting that composing multi-hop reasoning chains remains a fundamental bottleneck. *Calibration failure* manifests bidirectionally: top-performing models are systematically overconfident (refusing $< 10\%$ despite significant insufficiency), while weaker models may over-refuse or, worse, lack any calibration signal. Notably, no model in our evaluation achieves both high utilization and high refusal—a perfectly calibrated model would excel at synthesis while maintaining appropriate uncertainty, but current architectures appear to force a tradeoff where strong synthesis capability invariably leads to overconfidence. This suggests a fundamental tension between competence and epistemic humility in existing architectures.

The emergence of these distinct profiles suggests that improving multi-hop reasoning requires targeted interventions. Models in the Synthesis Bottleneck category need architectural improvements to reasoning composition, not better search. Overconfident models need calibration mechanisms that don't sacrifice performance. Most importantly, the absence of any model achieving high performance across all three metrics—even Grok-4 and Claude-Opus-4.1, the best-balanced models, only reaches 50.5% and 44.5% pass rate—demonstrates that robust multi-hop reasoning remains an open challenge, with synthesis capability being the universal limiting factor.

3.2.2 Knowledge Degradation in Synthesis

Even when models achieve knowledge sufficiency ($\text{KS}(d) = 1$), they often fail to generate the correct answer. We call this *knowledge degradation*: evidence is present in context, yet models forget, ignore, or misuse it during synthesis. To analyse this effect, we focus on two diagnostics, *Forget* and *Lead-astray*, which reveal two distinct synthesis failures: models either fail to recall known knowledge (Forget) or become disrupted by noisy search contexts (Lead-astray).

Knowledge degradation patterns. From table 1, models with *lower* Forget and Lead-astray generally exhibit *higher* Knowledge Utilization, which in turn coincides with higher Generation Score and Pass@1. For instance, Grok-4 (Forget 23.65%, Lead-astray 27.70%) attains the highest Knowledge Utilization F1 (56.19%), the highest Generation Score (34.71%), and 50.5% Pass@1. Similarly, O3-Pro (Forget 21.83%, Lead-astray 25.35%) reaches 49.40% Knowledge Utilization, 20.86% Generation Score, and the best Pass@1 (56.0%). GPT-5 shows a comparable pattern with very low Forget (17.72%) and strong Knowledge Utilization (49.58%), alongside 23.21% Generation Score and 50.5% Pass@1. In contrast, when Forget is high, Knowledge Utilization collapses and down-

stream metrics follow suit: GPT-OSS-120B records the lowest Knowledge Utilization (10.73%) with Forget at 100.00% (Lead-astray 0.00%), yielding only 2.75% Generation Score and 24.0% Pass@1; Qwen3-30B-Thinking has Forget 79.65% (Lead-astray 16.81%), with 13.16% Knowledge Utilization, 7.25% Generation Score, and 12.5% Pass@1. Similar degradations appear for Gemini-2.5-Flash-Think (Forget 57.63%, Knowledge Utilization 16.35%) and Tongyi-DeepResearch-30B (Forget 43.93%, Knowledge Utilization 15.69%).

Forgetting dominates misdirection. Averaging across all models, the mean difference $\text{Forget} - \text{Lead-astray}$ is $+10.35\%$ points. This gap indicates that, on **WebDetective**, failures after achieving knowledge sufficiency are more often due to *not using* already-available evidence (forgetting during synthesis) than to being *led astray* by spurious context. In other words, the principal bottleneck lies in evidence integration at answer time rather than in resisting distractors.

3.2.3 Robustness to Test-Time Scaling

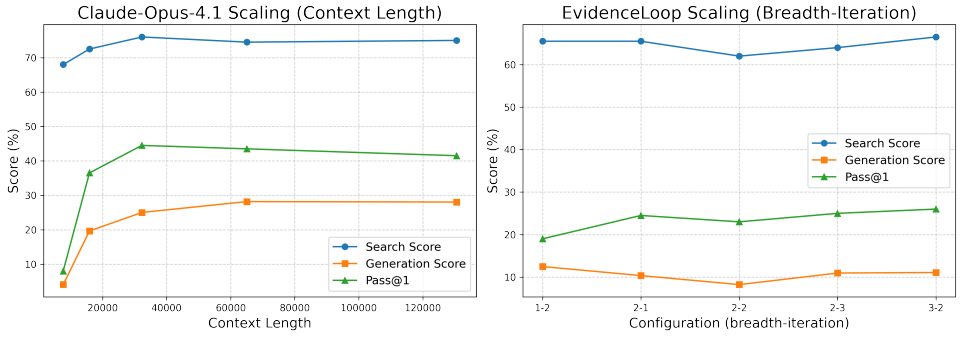

Figure 2: Scaling under test-time scaling (TTS).

To assess the robustness of our benchmark, we examine test-time scaling (TTS) along two axes. First, we scale context length for a strong ReAct model (Claude-Opus-4.1) to test whether larger budgets improve performance. Second, we vary breadth and iteration counts in **EvidenceLoop** to probe whether extensive exploration can exploit **WebDetective**. These analyses test whether **WebDetective** can be artificially boosted by TTS or instead faithfully reflect underlying system capabilities.

In fig. 2, we observe two main trends. For Claude-Opus-4.1, enlarging the context window from 8K to 32K tokens brings negligible gains: Generation Score plateaus at about 34%, Pass@1 at about 50%, and Search Score increases by less than 1%. For **EvidenceLoop**, expanding the controller from breadth=1, iteration=2 to breadth=3, iteration=2 raises Search Score slightly (45% $\rightarrow$ 46%, +1%), leaves Generation Score unchanged at 21%, and improves Pass@1 from 49% to 56% (+7%). These results indicate that our benchmark is robust to naïve test-time scaling. Neither larger context budgets nor shallow exploration suffice to "hack" **WebDetective**; achieving further gains requires genuine advances in model reasoning and knowledge utilisation.

4 Conclusion

We introduced WebDetective, a benchmark for evaluating web agents on hint-free multi-hop deep search within a controlled Wikipedia sandbox. Unlike prior datasets that embed reasoning paths (PH) or entity fingerprints (SH), our design enforces autonomous discovery of reasoning chains while enabling fine-grained attribution of failure modes. Evaluation of 25 state-of-the-art models reveals consistent weaknesses: systems often retrieve sufficient evidence but fail to utilise it effectively, and appropriate refusals remain nearly absent. Our proposed agentic workflow **EvidenceLoop** demonstrates that explicit verification and systematic evidence tracking can partially close this gap, underscoring that performance cannot be trivially improved by test-time scaling alone.

ETHICS STATEMENT

We have read and will adhere to the ICLR Code of Ethics and the ICLR Code of Conduct. Our research introduces WebDetective, a framework for hint-free multi-hop questions answering and evidenceLoop, an agentic-workflow baseline. The methods used in our study are well-established for academic research. These environments do not contain any personally identifiable information (PII) or sensitive real-world data. Our work did not involve human subjects, crowd-sourcing, or the scraping of private data; therefore, Institutional Review Board (IRB) approval was not required.

We acknowledge that research on autonomous agents carries potential dual-use risks. To mitigate these, our experiments are intentionally confined to benign, closed-world tasks such as online shopping and household activities within simulated settings. We followed good scholarly practice by reporting our methods and results transparently and citing prior work accurately. The authors declare no competing interests or external sponsorships that could have influenced the outcomes of this research.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of our research. All essential details for reproducing our results are provided within this paper. The WebDetective benchmark design, including the hint-free question formulation principles, co-designed evaluation system with selective entity masking, and the two-level diagnostic evaluation framework, are thoroughly detailed in the methodology sections. The complete WebDetective dataset statistics, question-environment co-design methodology, and human validation procedures are comprehensively described in the dataset construction sections. Our experimental setup, including the specific language models evaluated (GPT-5, O3-Pro, Claude-Opus-4.1, Gemini-2.5-Pro, Grok-4, Qwen3-235B-Thinking, and others), the controlled Wikipedia sandbox configuration, knowledge sufficiency probing methodology, and evaluation protocols, is fully documented in the experimental sections. The diagnostic metrics formalization including Knowledge Score, Generation Score, Good Refusal (GR), Knowledge Utilization (KU), and knowledge degradation tests (Forget and Lead-astray) are rigorously defined in the evaluation framework sections. To facilitate full replication of our benchmark construction pipeline and agent evaluation experiments, we will release our complete codebase, the controlled Wikipedia sandbox environment, hint-free question dataset with evidence chains, and evaluation scripts as supplementary material.

REFERENCES

Sierra AI. -bench: Benchmarking ai agents for the real-world, 2024.

Anonymous. Meqa: A benchmark for multi-hop event-centric question answering with explanations. In *NeurIPS*, 2024a.

Anonymous. Theagentcompany: Benchmarking llm agents on consequential real world tasks. *arXiv:2412.14161*, 2024b.

Anthropic. Introducing claude 4. `https://www.anthropic.com/news/claude-4`, May 2025. Accessed: 2025.

ByteDance Seed Team. Doubao 1.6 thinking and flash models. `https://seed.bytedance.com/`, 2025. Part of the Doubao/Seed model family. Accessed: 2025.

Jifan Chen, Shih-ting Lin, and Greg Durrett. Multi-hop question answering via reasoning chains. *arXiv:1910.02610*, 2019.

DeepSeek-AI, Wenfeng Liang, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

DeepSeek-AI, Wenfeng Liang, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Abul Ehtesham et al. A systematic review of key retrieval-augmented generation (rag) systems. *arXiv:2507.18910*, 2025.

Jinyuan Fang, Zaiqiao Meng, and Craig MacDonald. Trace the evidence: Constructing knowledge-grounded reasoning chains for retrieval-augmented generation. In *EMNLP Findings*, 2024.

Google DeepMind. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf, March 2025. Accessed: 2025.

Jie He et al. Mintqa: A multi-hop question answering benchmark for evaluating llms on new and tail knowledge. *arXiv:2412.17032*, 2024.

Xanh Ho et al. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *COLING*, 2020.

Soyeong Jeong et al. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *NAACL*, 2024.

Carlos E Jimenez et al. Swe-bench: Can language models resolve real-world github issues? In *ICLR*, 2024.

Tom Kwiatkowski et al. Natural questions: A benchmark for question answering research. *TACL*, 2019.

Ronghan Li et al. Different paths to the same destination: Diversifying llms generation for multi-hop open-domain question answering. *Knowledge-Based Systems*, 309, 2024a.

Zijian Li et al. Graph neural network enhanced retrieval for question answering of llms. *arXiv:2406.06572*, 2024b.

Hao Liu et al. Hoprag: Multi-hop reasoning for logic-aware retrieval-augmented generation. *arXiv:2502.12442*, 2025.

Xiao Liu et al. Agentbench: Evaluating llms as agents. *arXiv:2308.03688*, 2023.

Moonshot AI. Kimi k2: Open agentic intelligence technical report. https://github.com/MoonshotAI/Kimi-K2/blob/main/tech_report.pdf, July 2025. Accessed: 2025.

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

Cheng Niu et al. Ragtruth: A hallucination corpus for developing guardrails in rag systems. *arXiv preprint*, 2024.

OpenAI. Introducing swe-bench verified, 2024.

OpenAI. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/, August 2025a. Accessed: 2025.

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

OpenAI. Openai o3 and o4-mini system card. https://cdn.openai.com/o3-and-o4-mini-system-card.pdf, April 2025c. Accessed: 2025.

OpenAI. Introducing openai o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/, April 2025d. Accessed: 2025.

Fabio Petroni et al. Kilt: A benchmark for knowledge intensive language tasks. In *NAACL*, 2021.

Es Shahul et al. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint*, 2024.

Aditi Singh et al. Agentic retrieval-augmented generation: A survey on agentic rag. *arXiv:2501.09136*, 2025.

Zhengwei Tao, Jialong Wu, Wenbiao Yin, Junkai Zhang, Baixuan Li, Haiyang Shen, Kuan Li, Liwen Zhang, Xinyu Wang, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. Webshaper: Agentically data synthesizing via information-seeking formalization, 2025. URL https://arxiv.org/abs/2507.15061.

Tongyi DeepResearch Team. Tongyi deepresearch: The leading open-source deep research agent. https://github.com/Alibaba-NLP/DeepResearch, 2025. 30B-parameter agentic model for long-horizon research tasks.

Harsh Trivedi et al. Musique: Multihop questions via single-hop question composition. *TACL*, 2022.

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.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. Constructing datasets for multi-hop reading comprehension across documents. In *TACL*, 2018.

Ryan Wong, Jiawei Wang, Junjie Zhao, Li Chen, Yan Gao, Long Zhang, Xuan Zhou, Zuo Wang, Kai Xiang, Ge Zhang, Wenhao Huang, Yang Wang, and Ke Wang. Widesearch: Benchmarking agentic broad info-seeking, 2025. URL https://arxiv.org/abs/2508.07999.

xAI. Grok 4: The most intelligent model in the world. https://x.ai/news/grok-4, July 2025. Accessed: 2025.

An Yang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering, 2018a. URL https://arxiv.org/abs/1809.09600.

Zhilin Yang et al. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*, 2018b.

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

Zhipu AI Team. Glm-4.5: Reasoning, coding, and agentic abilities. August 2025. Technical report available at https://z.ai/blog/glm-4.5.

Shuyan Zhou et al. Webarena: A realistic web environment for building autonomous agents. *arXiv:2307.13854*, 2023.

Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. Fanoutqa: A multi-hop, multi-document question answering benchmark for large language models. In *ACL*, 2024.

A APPENDIX

A.1 FORMAL METRICS DEFINITION

Traditional multi-hop QA evaluation reduces agent performance to a single pass rate, obscuring the diverse failure modes that occur in complex reasoning tasks. An agent that searches exhaustively but fails to synthesize evidence exhibits fundamentally different limitations than one that refuses prematurely or hallucinates from parametric knowledge. Our co-designed sandbox, with its guaranteed unique reasoning paths, enables unprecedented diagnostic precision in distinguishing these failure modes.

We introduce a two-level evaluation framework that separates *knowledge sufficiency* from *generation quality*. First, we assess whether an agent possesses the requisite knowledge—either through successful search or parametric memory—to answer the question. Second, conditioned on knowledge sufficiency, we evaluate the agent's ability to either correctly synthesize an answer or appropriately refuse when evidence is insufficient.

Knowledge Sufficiency Assessment: We assess whether an agent possesses—either through search or parametric knowledge—all information necessary to answer the question. Given the required evidence chain $\mathcal{E} = \{e_1, ..., e_n\}$, we first identify which evidence the agent discovered through search by tracking visited pages in our sandbox. For any missing evidence $e_i \notin \mathcal{E}_{found}$, we then test whether the agent can access this information parametrically.

Specifically, for each missing piece of evidence e_i, we construct a focused probe query p_i that tests for that specific knowledge. For instance, if the agent never visited Kane's page and thus missed discovering that "Kane Cornes has brother Chad Cornes," we probe with: "Kane Cornes has brother _____". We define Probe(p_i) as a function that submits probe p_i to the base model and returns whether the model's response matches the expected answer for evidence e_i.

For instance d with evidence chain of length n_d, we define:

$$k_i^d = \begin{cases} 1 & \text{if } e_i \in \mathcal{E}_{found} \text{ (found via search)} \\ 1 & \text{if } e_i \notin \mathcal{E}_{found} \wedge \text{Probe}(p_i) = \text{correct} \\ 0 & \text{otherwise} \end{cases}$$

The instance is knowledge sufficient if: $\text{KS}(d) = \prod_{i=1}^{n_d} k_i^d = 1$

We define the overall **Knowledge Score** as the fraction of instances where the agent achieves knowledge sufficiency:

$$\text{KnowledgeScore} = \frac{|\mathcal{S}|}{N} \tag{1}$$

This metric directly measures search effectiveness—a low KnowledgeScore indicates the agent fails to discover necessary evidence through exploration, regardless of its ability to synthesize answers.

Search Score: While our masking mechanism enforces the canonical reasoning path $v_0 \to v_1 \to ... \to v_n$, we observe that models may leverage alternative valid reasoning strategies. Specifically, if an entity v_x (reachable through search from v_0) has a meaningful relationship to the answer v_n stored in the model's parametric knowledge, the model can combine partial search with memory to reach the correct answer. This represents a legitimate form of reasoning that demonstrates efficient use of both search and parametric knowledge.

To capture this capability, we define **SearchScore** that credits models for finding correct answers through any valid combination of search and parametric knowledge, provided their search efficiency meets or exceeds the ground truth path:

$$\text{SearchScore} = \text{KnowledgeScore} + \frac{|\mathcal{C}|}{N} \tag{2}$$

where $\mathcal{C} = \{d \in \mathcal{D} : \text{correct}(d) \wedge \text{hops}(d) \leq \text{hops}_{\text{GT}}(d) \wedge \text{KS}(d) = 0\}$ represents instances where the model:

- Produces the correct answer despite not having complete knowledge sufficiency through the canonical path
- Uses no more search hops than the ground truth path length
- Effectively combines discovered entities with parametric knowledge

This metric recognizes that effective multi-hop reasoning isn't solely about following predetermined paths, but about efficiently discovering and leveraging available information—whether through complete evidence chains or intelligent combination of partial search with existing knowledge. The hop constraint ensures models aren't simply performing exhaustive search, but are discovering meaningful connections that enable efficient reasoning.

Search Score: While our masking mechanism enforces the canonical reasoning path $v_0 \to v_1 \to ... \to v_n$, we observe that models may leverage alternative valid reasoning strategies. Specifically, if an entity v_x (reachable through search from v_0) has a meaningful relationship to the answer v_n stored in the model's parametric knowledge, the model can combine partial search with memory to reach the correct answer. This represents a legitimate form of reasoning that demonstrates efficient use of both search and parametric knowledge.

To capture this capability, we define **SearchScore** that credits models for finding correct answers through any valid combination of search and parametric knowledge, provided their search efficiency meets or exceeds the ground truth path:

$$\text{SearchScore} = \text{KnowledgeScore} + \frac{|\mathcal{C}|}{N} \tag{3}$$

where $\mathcal{C} = \{d \in \mathcal{D} : \text{correct}(d) \wedge \text{searched}(d) \wedge \text{hops}(d) \leq \text{hops}_{\text{GT}}(d) \wedge \text{KS}(d) = 0\}$ represents instances where the model:

- Produces the correct answer despite not having complete knowledge sufficiency through the canonical path
- Actually performs web search (not relying solely on parametric knowledge)
- Uses no more search hops than the ground truth path length
- Effectively combines discovered entities with parametric knowledge

The requirement that $\text{searched}(d) = \text{true}$ ensures we only reward genuine search-memory combination strategies, not pure parametric recall. This metric recognizes that effective multi-hop reasoning isn't solely about following predetermined paths, but about efficiently discovering and leveraging available information through intelligent combination of partial search with existing knowledge. The hop constraint ensures models aren't simply performing exhaustive search, but are discovering meaningful connections that enable efficient reasoning.

Generation Quality Assesement: Given the knowledge sufficiency assessment, we evaluate generation quality through a conditional framework that captures the fundamental tension in multi-hop QA: agents must synthesize answers when they have sufficient evidence while appropriately refusing when they don't.

Let $\mathcal{D} = \{d_1, ..., d_N\}$ denote the evaluation dataset with N instances. We partition $\mathcal{D}$ along two dimensions:

Knowledge dimension:

$$\mathcal{S} = \{d \in \mathcal{D} : \text{KS}(d) = 1\} \quad \text{(knowledge sufficient instances)} \tag{4}$$

$$\mathcal{I} = \mathcal{D} \setminus \mathcal{S} \quad \text{(knowledge insufficient instances)} \tag{5}$$

Response dimension:

$$\mathcal{A} = \{d \in \mathcal{D} : \text{agent attempts answer}\} \tag{6}$$

$$\mathcal{R} = \{d \in \mathcal{D} : \text{agent refuses}\} \tag{7}$$

where attempts are further partitioned into $\mathcal{A} = \mathcal{A}_c \cup \mathcal{A}_w$, with $\mathcal{A}_c$ denoting correct answers (matching ground truth) and $\mathcal{A}_w$ denoting wrong answers. Note that $\mathcal{A} \cup \mathcal{R} = \mathcal{D}$.

The intersection of these dimensions creates critical regions that reveal different agent capabilities and failure modes:

- Knowledge sufficient, answers correctly ($\mathcal{S} \cap \mathcal{A}_c$): The ideal scenario—the agent possesses all evidence and successfully synthesizes the correct answer. This demonstrates *knowledge utilization*, the ability to compose multi-hop reasoning without forgetting intermediate steps or being disrupted by irrelevant information.
- Knowledge sufficient, answers wrongly ($\mathcal{S} \cap \mathcal{A}_w$): A synthesis failure—despite having all necessary evidence, the agent produces an incorrect answer. This reveals breakdowns in reasoning composition, where evidence possession doesn't translate to correct synthesis.
- Knowledge sufficient, refuses ($\mathcal{S} \cap \mathcal{R}$): Over-caution—the agent has sufficient evidence but refuses to answer. This represents failure to recognize that the evidence chain is complete, missing opportunities to provide helpful answers.
- Knowledge insufficient, refuses ($\mathcal{I} \cap \mathcal{R}$): The second ideal scenario—*good refusal*. The agent lacks critical evidence and appropriately declines to answer, demonstrating epistemic awareness and avoiding hallucination.
- Knowledge insufficient, attempts answer ($\mathcal{I} \cap \mathcal{A}$): The most problematic behavior—the agent lacks evidence but attempts an answer anyway (whether correct by luck or wrong), typically through hallucination, guessing, or over-reliance on partial information.

This visualization reveals that generation quality isn't monolithic—an agent might excel at refusing when uncertain but fail to synthesize known information, or vice versa. For instance, an overly conservative agent might achieve perfect good refusal by refusing everything (large $\mathcal{R}$ region), while an overly confident agent might attempt every question (large $\mathcal{A}$ region) leading to frequent hallucinations in the $\mathcal{I}$ zone.

To capture these complementary capabilities, we define two core metrics:

Good Refusal (GR) measures the agent's ability to appropriately abstain when lacking evidence. It evaluates $\mathcal{R}$'s overlap with $\mathcal{I}$—high recall indicates the agent successfully avoids hallucination by refusing most knowledge-insufficient cases, while high precision ensures refusals are justified (not bleeding unnecessarily into $\mathcal{S}$).

$$\text{Recall}_{\text{GR}} = \frac{|\mathcal{R} \cap \mathcal{I}|}{|\mathcal{I}|}, \quad \text{Precision}_{\text{GR}} = \frac{|\mathcal{R} \cap \mathcal{I}|}{|\mathcal{R}|}, \quad \text{F1}_{\text{GR}} = 2 \cdot \frac{\text{Recall}_{\text{GR}} \cdot \text{Precision}_{\text{GR}}}{\text{Recall}_{\text{GR}} + \text{Precision}_{\text{GR}}} \tag{8}$$

Knowledge Utilization (KU) assesses the agent's ability to synthesize correct answers when evidence is available. It examines $\mathcal{A}_c$ within $\mathcal{S}$—high recall means the agent leverages available evidence effectively, while high precision indicates that attempts are typically grounded in sufficient knowledge.

$$\text{Recall}_{\text{KU}} = \frac{|\mathcal{A}_c \cap \mathcal{S}|}{|\mathcal{S}|}, \quad \text{Precision}_{\text{KU}} = \frac{|\mathcal{A}_c \cap \mathcal{S}|}{|\mathcal{A}|}, \quad \text{F1}_{\text{KU}} = 2 \cdot \frac{\text{Recall}_{\text{KU}} \cdot \text{Precision}_{\text{KU}}}{\text{Recall}_{\text{KU}} + \text{Precision}_{\text{KU}}} \tag{9}$$

Importantly, these metrics are non-competing—improving one shouldn't decrease the other in a well-designed system. An ideal agent achieves high F1_{GR} (refusing when and only when knowledge is insufficient) while maintaining high F1_{KU} (correctly answering when evidence is available). To capture both capabilities while preventing gaming, we define a unified **Generation Score**:

$$\text{GenScore} = \frac{\text{F1}_{\text{GR}} + \text{F1}_{\text{KU}}}{2} \cdot \frac{|\mathcal{S}|}{N} \tag{10}$$

The $|\mathcal{S}|/N$ weighting (KnowledgeScore) is crucial for preventing metric exploitation: without it, models could game the evaluation by adopting a degenerate strategy—performing minimal search and refusing nearly all questions. Such a model would achieve high F1_{GR} (correctly refusing the many knowledge-insufficient cases) while contributing nothing useful, yet still obtain a substantial GenScore. This creates a perverse incentive where models might optimize for conservative refusal rather than improving search capabilities. The weighting ensures that models cannot exploit the evaluation structure—they must demonstrate effective evidence discovery to achieve competitive scores, aligning the metric incentives with the actual goal of multi-hop reasoning systems.

Unlike simple pass rates, our metrics provide actionable insights: low KnowledgeScore indicates inadequate search strategies, low GR scores reveal over-confident hallucination, and low KU scores expose synthesis failures despite having evidence. This diagnostic precision, enabled by our co-designed evaluation environment, illuminates the specific capabilities required for robust multi-hop reasoning.

Knowledge Forget Test. We test $\text{LM}(q, \mathcal{E}_{\text{found}})$ where $\mathcal{E}_{\text{found}} = \mathcal{E}_{\text{visited}} \cap \mathcal{E}_{\text{GT}}$ represents evidence from ground-truth URLs that the model actually visited. When this fails despite $\text{KS}(d) = 1$, it reveals *knowledge forget*: the model cannot leverage its parametric knowledge to fill missing pieces when answering the full question, even though it correctly answers individual probes $\text{Probe}(p_i)$ for each missing evidence $e_i \in \mathcal{E}_{\text{GT}} \setminus \mathcal{E}_{\text{found}}$.

Lead-astray Test. When $\text{LM}(q, \mathcal{E}_{\text{found}})$ succeeds but the model fails in its actual search trajectory, we identify *lead-astray*: the model can synthesize the answer from clean evidence but is disrupted by the accumulated search context (failed attempts, irrelevant pages, exploration noise).

Formally, for the set of knowledge-sufficient instances $\mathcal{S}^* = \{d \in \mathcal{D} : \text{KS}(d) = 1 \wedge \text{incorrect}(d)\}$ where the model fails despite having all necessary knowledge:

$$\text{ForgetRate} = \frac{|\{d \in \mathcal{S}^* : \text{LM}(q_d, \mathcal{E}^d_{\text{found}}) \neq a^*_d\}|}{|\mathcal{S}^*|}$$

$$\text{LeadAstrayRate} = \frac{|\{d \in \mathcal{S}^* : \text{LM}(q_d, \mathcal{E}^d_{\text{found}}) = a^*_d \wedge \text{actual\_output}(d) \neq a^*_d\}|}{|\mathcal{S}^*|}$$

These metrics decompose knowledge-sufficient failures: ForgetRate identifies when models cannot integrate parametric knowledge with partial search results, while LeadAstrayRate reveals when noisy search trajectories corrupt otherwise successful reasoning.

B Failure Case Studies

We identify four recurring failure patterns through qualitative analysis:

1. Instruction Drift in Long Trajectories: After 15+ tool calls, models lose track of the original question, pursuing tangentially related information. Example: When asking "Who is the father of Kane?", models explore Kane's entire family tree rather than following the specific chain to the answer.

2. Premature Satisfaction: Models often stop searching after finding partial information that seems plausible. They attempt answers based on incomplete evidence rather than verifying they have the complete reasoning chain.

3. Entity Confusion: With similar entity names, models conflate different entities or miss crucial disambiguating information, especially problematic in dense domains with many related entities.

4. Context Window Pollution: Failed searches and irrelevant exploration consume context space, creating noise that interferes with synthesis even when correct evidence is eventually found.

C Dataset Human Validation

Human Validation. Questions passing automated verification undergo human annotation by 2 researchers with NLP expertise. Each question is independently reviewed following a structured protocol:

1. **Annotation Protocol**: For each question, annotators receive the question q, evidence chain $\mathcal{E} = \{e_1, ..., e_n\}$, and answer v_n. They verify three criteria:
   - *Necessity*: Whether the question can be answered without the evidence chain using only general knowledge
   - *Sufficiency*: Whether the evidence chain logically derives the answer without requiring external information

- *No hints*: Whether the question avoids linguistic cues that reveal intermediate reasoning steps

2. **Validation Process**: Each question requires 2-3 minutes of review. Annotators trace through the reasoning chain step-by-step, attempting to answer the question both with and without the evidence to ensure all pieces are necessary. Questions where intermediate entities could be guessed from the question phrasing or where the evidence chain has logical gaps are rejected.

3. **Dataset Filtering**: Of approximately 450 machine-verified questions reviewed, 200 questions (∼44%) pass human validation. Common rejection reasons include: evidence chains not targeting the questions, evidence chains with missing logical connections, and questions containing subtle hints about the reasoning path (e.g., mentioning attributes that implicitly identify intermediate entities).

This manual verification process, totaling approximately 20 hours of annotation effort, ensures our final dataset contains only questions that genuinely require discovering and composing the complete multi-hop reasoning chain.

D Dataset Statistics

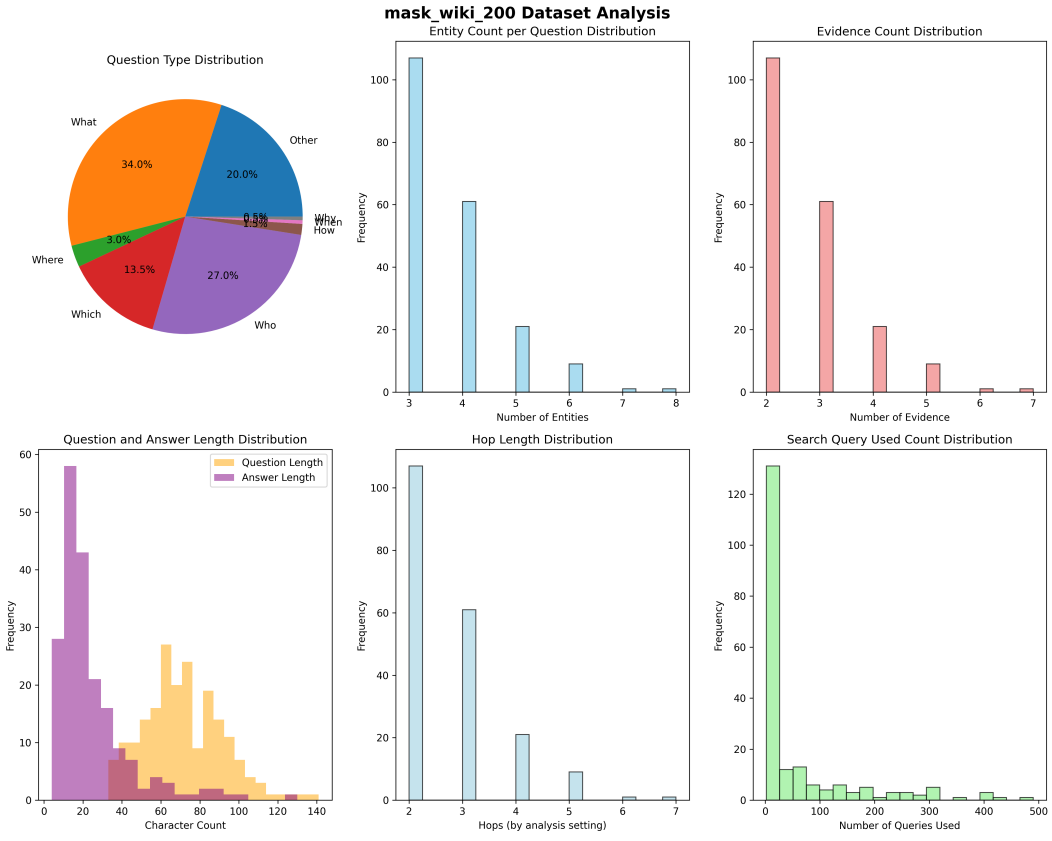

Figure 3: Dataset statistics for WebDetective benchmark. The figure shows: (a) Distribution of question types, (b) Number of entities per question, (c) Evidence count distribution, (d) Question and answer length in characters, (e) Hop length distribution by analysis setting, and (f) Search query usage patterns. The dataset exhibits controlled complexity with predominantly 2-3 hop questions while maintaining challenging longer chains.

Our final WebDetective benchmark comprises 200 hint-free multi-hop questions, carefully curated through our verification pipeline. Figure 3 presents a comprehensive analysis of the dataset characteristics.

Question Complexity. The dataset exhibits controlled complexity suitable for diagnostic evaluation. Questions require 2 to 6 hops of reasoning (mean: 2.85 hops), with the distribution heavily weighted toward 2-hop (55%) and 3-hop (31%) questions, while maintaining a challenging subset of 4+ hop questions (14%). This distribution balances tractability with sufficient complexity to stress-test multi-hop reasoning capabilities. Each question involves 3 to 8 Wikipedia entities (mean: 3.73), with the modal question requiring exactly 3 entities to form the complete reasoning chain.

Question Types and Domains. The dataset spans diverse question types, with "What" questions comprising 34% of the dataset, "Who" questions 27%, and "Which" questions 13.5%, ensuring broad coverage of information-seeking patterns. Questions are concise (mean: 71.4 characters) with typically short answers (mean: 28.6 characters), reflecting natural information needs without verbose specifications that might hint at reasoning paths.

Evidence Requirements. The evidence distribution aligns with hop counts, with most questions requiring 2-3 pieces of evidence (52.5% and 31% respectively). This controlled evidence requirement enables precise diagnosis of where reasoning fails—whether at initial discovery, intermediate steps, or final synthesis.

The dataset's careful balance of complexity, diversity, and diagnostic precision makes it suitable for evaluating the full spectrum of multi-hop reasoning capabilities, from basic 2-hop familial relationships to complex 5-hop chains requiring sustained context retention across multiple search iterations.

E THE EVIDENCELOOP FRAMEWORK

The hint-free nature of our benchmark exposes fundamental limitations in current multi-hop reasoning approaches. Without linguistic scaffolding, agents must autonomously discover which connections matter among thousands of facts—a challenge that, as our results show, causes even state-of-the-art models to achieve only 50% accuracy. To better understand these challenges and establish a baseline for future work, we design an agentic workflow that explicitly targets the unique difficulties our benchmark reveals: the need for broad exploration without context explosion, evidence retention across long trajectories, and synthesis from accumulated but noisy search contexts.

E.1 CORE ARCHITECTURE: ITERATIVE REFINEMENT WITH FALLBACK

Our framework attempts to balance exploration breadth with computational feasibility through $R_{\max}$ iterations. Each iteration r launches N parallel solver agents $\{A_1^r, ..., A_N^r\}$ that explore different reasoning paths simultaneously. Each agent A_i^r receives the question q, an aggregated context C^r from previous iterations (with $C^0 = \emptyset$ initially), and executes up to B actions.

After each iteration, we employ a two-stage refinement process:

1. An **extraction agent** processes the reasoning contexts from all N parallel agents to identify key findings, evidence references, and promising paths
2. An **aggregation agent** synthesizes these extracted insights into a refined context C^{r+1} for the next round, preserving valuable discoveries while discarding exploration noise

This iterative refinement addresses a core challenge our benchmark exposes: early rounds might explore many directions—sports connections, geographic locations, family relations—but the extraction-aggregation pipeline identifies which paths warrant deeper exploration, preventing the context explosion that causes single-pass approaches to fail while avoiding premature path commitment. If no conclusive answer emerges after $R_{\max}$ iterations, a final aggregation agent consolidates all discovered evidence into a comprehensive context C^{final}. This context is then passed to a synthesis-only solver that attempts to derive the answer purely from the accumulated evidence without additional search actions—effectively testing whether the failure stems from insufficient exploration or poor evidence composition.

E.2 Evidence Memory System

Enabling this iterative refinement is our Evidence Memory System $\mathcal{M}$. When any agent performs a search or visits a page, the system: 1) Stores complete results in persistent memory; 2) Assigns unique Evidence IDs (EIDs) for reference; and 3) Returns both full content and EID to the agent.

The EID system serves multiple critical functions in our framework. First, during extraction and aggregation between iterations, the extraction and aggregation agent produces summaries that preserve EID references alongside extracted facts—for example, "Kane has brother Chad [EID-042], Chad's stepmother is Nicole [EID-089]". This allows subsequent solver agents to receive concise, actionable summaries while retaining the ability to retrieve full evidence on demand through the retrieve action as an external tool, which takes an EID and returns the complete original content from memory. Second, these EIDs enable systematic verification (detailed in Section E.3), where verification agents can trace claims back to their original sources and validate reasoning chains against the actual evidence.

The memory system transforms how evidence flows through iterations. Rather than forcing agents to work with either overwhelming full documents or lossy compressions, agents can work with focused summaries while maintaining access to complete evidence through EID-based retrieval. This design ensures that even as contexts become more refined across iterations, agents never lose access to the complete evidence trail that supports their reasoning, allowing them to dive deep into specific evidence when needed for detailed analysis or verification.

E.3 Verification: Ensuring Evidence-Grounded Reasoning

The verification mechanism prevents premature or hallucinated answers from propagating through our system. When any solver agent A_i^r proposes an answer, it must decompose the answer into atomic claims $\{c_1, c_2, ..., c_m\}$, where each claim c_j is explicitly linked to an EID from the memory system—e.g., "Kane has brother Chad [EID-042]". No unsupported claims are permitted.

The verification agent V evaluates each proposal:

$$V(q, \text{answer}, \{c_j, \text{EID}_j\}_{j=1}^m) \rightarrow \{\text{YES}, \text{NO}(\text{feedback})\}$$

For each claim-evidence pair, the verifier retrieves the full content from $\mathcal{M}$ via the EID and validates: (1) whether the source genuinely entails the claimed fact, (2) whether the claims collectively derive the answer, and (3) whether the answer correctly addresses the original question.

Verification occurs *during* solver execution. Rejections provide specific feedback back to the solver, allowing immediate gap-filling within the remaining action budget B, while acceptance immediately terminates all iterations. This ensures both evidence grounding and computational efficiency—solvers can correct incomplete reasoning in real-time while avoiding unnecessary exploration once the answer is verified.

F Related Work

F.1 Multi-Hop Question Answering Benchmarks

Multi-hop QA benchmarks evaluate models' ability to compose information across multiple reasoning steps. Early datasets like HotpotQA Yang et al. (2018b) and WikiHop Welbl et al. (2018) established foundational evaluation frameworks but suffer from systematic biases. Recent benchmarks have expanded coverage: FanOutQA Zhu et al. (2024) addresses multi-document reasoning, MINTQA He et al. (2024) targets long-tail knowledge with 28K+ questions, and MEQA Anonymous (2024a) focuses on event-centric reasoning chains. However, these benchmarks embed hints that fundamentally alter the reasoning task.

We identify two categories of hints prevalent in existing benchmarks. **Path-hinting** occurs when questions linguistically encode reasoning chains (e.g., "What dance academy did the starring actress from The Glory of Tang Dynasty graduate from?"), reducing the task to executing pre-specified steps. **Specification-hinting** provides excessive constraints that make answers discoverable through constraint satisfaction rather than reasoning (e.g., combining "East German team," "founded 1966,"

"player born in 90s"). Unlike MuSiQue Trivedi et al. (2022) or 2WikiMultiHopQA Ho et al. (2020), which contain implicit structural hints, WebDetective introduces genuinely hint-free questions requiring autonomous reasoning path discovery.

F.2 Retrieval-Augmented Generation and Agents

The evolution from static RAG pipelines to agentic architectures represents a fundamental shift in how LLMs interact with external knowledge Singh et al. (2025); Ehtesham et al. (2025). While traditional RAG systems like TRACE Fang et al. (2024) achieve improvements through knowledge-grounded reasoning chains, they operate within predetermined patterns. Agentic RAG systems employ adaptive strategies: Adaptive-RAG Jeong et al. (2024) adjusts retrieval depth based on question complexity, while graph-based approaches like GNN-Ret Li et al. (2024b) and HopRAG Liu et al. (2025) leverage graph neural networks for multi-hop reasoning, achieving 10% accuracy improvements on benchmarks like 2WikiMQA.

Recent advances in 2025 emphasize diverse reasoning paths. DP-CoT Li et al. (2024a) addresses single-path limitations through passage-level and sentence-level evidence generation. However, our evaluation reveals these advances fail to overcome hint-free challenges: median Generation Scores of 20% across tested models indicate current architectures cannot effectively discover reasoning chains without linguistic scaffolding.

F.3 Evaluation Frameworks

Traditional metrics like exact match and F1 scores collapse diverse failure modes into single values, obscuring why models fail Kwiatkowski et al. (2019); Petroni et al. (2021). Recent frameworks attempt more nuanced evaluation: RAGAS Shahul et al. (2024) provides reference-free RAG metrics, while RAGTruth Niu et al. (2024) enables hallucination analysis. For agents, AgentBench Liu et al. (2023) evaluates across eight environments, tau-bench AI (2024) addresses multi-turn interactions, and TheAgentCompany Anonymous (2024b) introduces workplace tasks with simulated colleagues.

Web-based benchmarks have evolved significantly. WebArena Zhou et al. (2023) provides realistic web environments requiring long-horizon planning but lacks controlled evaluation for precise failure attribution. SWE-bench Jimenez et al. (2024) evaluates code generation on GitHub issues, with SWE-bench Verified OpenAI (2024) addressing underspecified problems. While these benchmarks test complex capabilities, they don't address the specific challenge of verifying multi-hop reasoning paths.

Our diagnostic framework decomposes evaluation into *knowledge sufficiency* (whether agents possess required evidence) and *conditional generation quality* (synthesis given sufficient knowledge). This separation reveals that models achieve 79% knowledge sufficiency but only 23% generation scores, indicating synthesis and relevance determination—not search—as primary bottlenecks.

G LLMs Usage

LLMs were used to polish the writing.