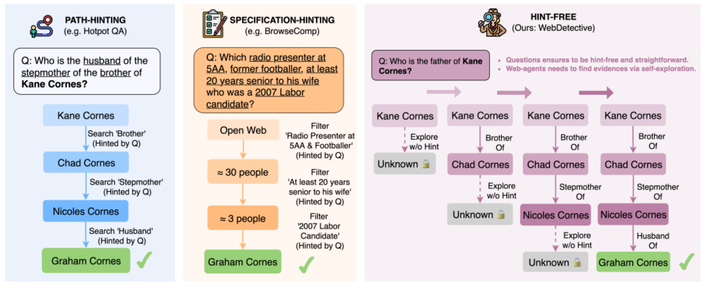
Demystifying Deep Search: A Holistic Evaluation With Hint-Free Multi-Hop Questions And Factorised Metrics

M. Song, R. Liu, X. Wang, Y. Jiang, P. Xie, F. Huang, J. Zhou, D. Herremans, S. Poria

Why Benchmarks Fail



Hidden Shortcut: Many *multi-hop* benchmarks leak the solution path to models

- *Path-Hinting* narrates the reasoning sequence inside the question
- *Specification-Hinting* fingerprints the answer with distinctive attributes

Evaluation Gap: A single pass rate cannot distinguish poor search from weak synthesis or bad refusal calibration

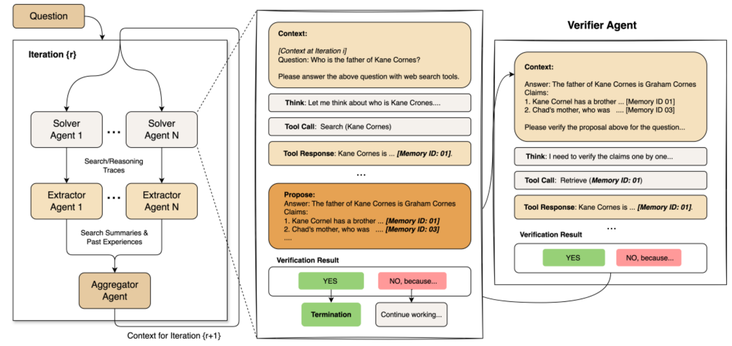
Benchmark Design

(1) Build *hint-free* questions where only the user need is stated clearly
(2) Mask the Wikipedia sandbox so each next entity must be reached through prior evidence
(3) Validate necessity and sufficiency with LM checks plus human review

Dataset: 200 questions, mean 2.85 hops, with broad domain coverage

Thanks for your attention!
Scan me for details →

EvidenceLoop: Agentic Framework



Workflow: Parallel solvers agents explore, an extractor distills evidence, and an aggregator refines context across rounds

Evidence Memory: Every retrieved fact gets an Evidence ID so agents reason over compact summaries with traceable retrieval

Verification: *Atomic* claims must cite evidence and pass entailment checks before answering or refusing

Factorized Metrics

Aspect	Standard ReAct	ReAct + Self-Reflection	EvidenceLoop (Ours)
Controller Loop	Single thought-tool-observation	Same + occasional reflection	Two-phase: exploration then verification
Memory	Flat dialogue history	Same + reflection messages	Structured buffer: (entity, snippet, source) tuples
Verification	Often none	Critiques not tied to evidence	Explicit loop using only evidence buffer
Breadth & Iterations	Single trajectory	Single trajectory	Parallel trajectories + explore-aggregate cycles
Refusal Behaviour	Implicit, rarely triggered	Loosely specified	Explicitly tied to verification: incomplete → refuse

Knowledge Discovery:

- *Knowledge Sufficiency* measures whether all needed evidence was acquired
- *Search Score* rewards efficient search plus valid parametric use

Generation Quality:

- *Good Refusal* measures abstaining when evidence is insufficient
- *Knowledge Utilization* measures correct answering when evidence is sufficient

Diagnostic Insight: *Forget* and *Lead-astray* separate failure to use found evidence from distraction by noisy trajectories

Main Results

Profile	Metric Pattern: Knowledge	Refusal	Utilization	Pass@1	Example Models	Failure Mode
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

Hard Benchmark: Best Pass@1 is only 56.0% under a mean 2.85 hops, showing hint-free path discovery remains *unsolved*

Key Decoupling: Top models often score around 72-80% in Search Score yet remain far lower in Generation Score

Weak Refusal: Strong systems usually *guess* rather than abstain when evidence is missing

Failure Profiles

Provider	Model	Knowledge Discovery: Knowledge Suff. (%)	Search Score (%)	Generation Quality: Generation Score (%)	Good Refusal F1 (%)	Knowledge Util. F1 (%)	Knowledge Degradation: Forget (%)	Lead-astray (%)	Pass@1 (%)
OpenAI	GPT-OSS-120B OpenAI (2025b)	16.00	23.50	2.75	23.59	10.73	100.00	**0.00**	24.00
	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
	o3 OpenAI (2025c)	70.00	76.00	18.29	3.29	48.97	24.29	24.29	53.50
	o3-Pro OpenAI (2025c)	71.00	78.00	20.86	9.37	49.40	21.83	25.35	**56.00**
	GPT-5-Chat OpenAI (2025a)	58.00	59.50	15.74	26.23	28.05	47.41	31.90	29.50
	GPT-5 OpenAI (2025a)	**79.00**	**80.00**	23.21	8.89	49.58	**17.72**	32.91	50.50
Anthropic	Claude-Sonnet-4-Think Anthropic (2025)	66.50	73.50	26.19	34.59	44.19	45.11	21.80	38.50
	Claude-Opus-4-Think Anthropic (2025)	68.00	73.50	21.00	30.53	31.23	43.38	32.35	29.00
	Claude-Opus-4.1 Anthropic (2025)	74.00	76.50	28.53	28.57	48.54	27.03	31.08	44.50
Google	Gemini-2.5-Flash-Think Google DeepMind (2025)	59.00	64.50	16.79	40.56	16.35	57.63	35.59	17.50
	Gemini-2.5-Pro Google DeepMind (2025)	65.50	73.00	11.64	10.87	24.68	44.27	35.11	28.50
xAI	Grok-4 xAI (2025)	74.00	77.50	**34.71**	37.63	**56.19**	23.65	27.70	50.50
Alibaba	Qwen3-30B-Think Yang et al. (2025)	56.50	59.00	7.25	12.51	13.16	79.65	16.81	12.50
	Qwen3-235B-Think Yang et al. (2025)	72.50	72.00	11.15	6.56	24.19	63.45	19.31	21.50
	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-Flash ByteDance Seed Team (2025)	54.50	57.50	20.00	**53.95**	19.46	68.81	21.10	13.50
	Doubao-1.6-Think ByteDance Seed Team (2025)	64.00	68.50	19.24	42.03	18.11	49.22	39.84	16.00
Zhipu AI	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
	GLM-4.5-Inner Zhipu AI Team (2025)	63.50	67.50	22.19	34.79	35.09	25.98	40.16	33.50
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 DeepSeek-AI et al. (2024)	61.50	56.50	13.62	27.97	16.34	44.72	44.72	17.00
	DeepSeek-V3.1-Terminus 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

Behavior Clusters: Models separate into overconfident elites, synthesis-bottleneck systems, conservative middling agents, and self-aware weaker models

Dominant Failure: Synthesis is the main bottleneck; many agents find enough evidence but still fail to compose grounded answers

Additional Findings: *Forget* exceeds *Lead-astray,* and test-time scaling soon plateaus