The Web Tool Trap: Understanding and Mitigating Over-Reliance in Browsing Agents

Extended Abstract

Jiawei Guo
BUPT
Beijing, China
gjw0613work@gmail.com

Hongjie Nie
USTC
Hefei, China
george-nie@mail.ustc.edu.cn

Qianbo Zang
Uni.lu
Ville de Luxembourg, Luxembourg
zangq@proton.me

Shu Yang
KAUST
Thuwal, Saudi Arabia
shu.yang@kaust.edu.sa

Shuodi Liu
BUPT
Beijing, China
liushuodi@bupt.edu.cn

Yiwei Ru
CASIA
Beijing, China
yiwei.ru@cripac.ia.ac.cn

Liuyu Xiang
BUPT
Beijing, China
xiangly@bupt.edu.cn

Di Wang
KAUST
Thuwal, Saudi Arabia
di.wang@kaust.edu.sa

Zhaofeng He
BUPT
Beijing, China
zhaofenghe@bupt.edu.cn

ABSTRACT

Large Language Model (LLM) agents that browse the web are increasingly important, but their effectiveness is hindered by imperfect integration of internal knowledge and external tools. We introduce BrowseBench and present the first systematic investigation into over-reliance patterns of browsing agents. Through controlled experiments, we identify three distinct failure modes: (1) Excessive Conservatism—unnecessary tool invocations for known information; (2) Over-trust in Web Sources—uncritical acceptance of retrieved content; and (3) Planning Deficiency—lack of query decomposition strategies. To address these, we propose three mitigation strategies: Direct Preference Optimization (DPO), Attention Refinement (AR), and Hierarchical Query Decomposition (HQD). Experiments demonstrate that our interventions significantly reduce over-reliance and enhance performance.

KEYWORDS

Large Language Model, Browsing Agent, Benchmark

ACM Reference Format:
Jiawei Guo, Hongjie Nie, Qianbo Zang, Shu Yang, Shuodi Liu, Yiwei Ru, Liuyu Xiang, Di Wang, and Zhaofeng He. 2026. The Web Tool Trap: Understanding and Mitigating Over-Reliance in Browsing Agents: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/HZER2072

1 INTRODUCTION

Large Language Models (LLMs) have evolved from conversational chatbots to sophisticated Artificial Intelligence (AI) agents capable of interacting with external tools [4, 9, 12]. Among these tools, web search has emerged as particularly prominent, enabling LLMs to overcome knowledge cutoff limitations and access up-to-date information. This evolution has led to systems like GPT-4 [1] with browsing capabilities and Claude[1] with web search integration, marking a significant paradigm shift in how AI systems retrieve and process information.

This tool dependency bears striking parallels to cognitive offloading in humans, where individuals strategically transfer cognitive burdens to external resources [8]. Just as humans can become over-reliant on external aids, LLM agents may develop analogous patterns of excessive dependence on web tools.

However, current evaluation benchmarks for tool-augmented LLMs exhibit critical limitations in addressing this phenomenon. Existing frameworks such as ToolBench [6], WebGPT [5], API-Bank [2], and τ-bench [11] predominantly focus on correctness metrics—whether the model successfully uses tools to arrive at accurate answers—while neglecting whether tool invocation was necessary in the first place. This myopic focus on accuracy overlooks the fundamental question of *when* tools should be used versus when parametric knowledge suffices. Moreover, while previous research has separately addressed over-reliance on the call process [10] and call output [3], no unified strategy exists to mitigate both facets simultaneously.

To address these gaps, we propose the first systematic investigation into web tool over-reliance patterns. We create **BrowseBench**, a dataset of 1,500 realistic information-seeking scenarios to rigorously evaluate how browsing agents interact with web search tools. Our study reveals three failure modes: (1) models exhibit **excessive conservatism** by performing unnecessary searches due to knowledge boundary ambiguity and epistemic uncertainty; (2) models **overtrust web sources** by uncritically accepting retrieved content, with attention mechanisms prioritizing linguistic coherence over factual accuracy; and (3) models **lack strategic planning**, either

[1] https://www.anthropic.com/news/claude-3-7-sonnet

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/HZER2072

Table 1: Performance metrics across all domains. Avg.: Overall Average. Our evaluation is conducted under a greedy decoding and zero-shot setting. For all models, we use the same minimalist prompt for comparison fairness.

Model	Avg. USR	Avg. FIIR	Avg. TDD
Claude-3.7-Sonnet	24.8	28.0	2.4
Grok-4	28.8	34.2	2.9
Gemini-2.5-Pro	29.3	34.9	3.0
Gemini-2.5-Flash	30.2	36.1	2.9
GPT-4o	32.5	37.9	3.1
DeepSeek-R1	37.0	42.1	2.8
Kimi-K2	34.1	42.4	3.5
GPT-o4-mini	36.5	45.0	3.2
Qwen2.5-72B-Instruct	47.3	58.7	4.2
Llama3.1-70B-Instruct	52.9	67.7	4.7

Table 2: Mitigation Strategies' Effect on Qwen-2.5-72B-Instruct. M1: DPO; M2: AR; M3: HQD; M1+M2+M3: Hybrid Mitigation Strategy.

Model	Avg. USR	Avg. FIIR	Avg. TDD
Qwen2.5-72B-Instruct	47.3	58.7	4.2
Δ M1	-7.2	-	-
Δ M2	-	-12.4	-
Δ M3	-	-	-1.4
Δ M1+M2+M3	-6.7	-14.1	-1.3

compressing multi-faceted queries into overloaded single searches or generating redundant parallel searches without systematic decomposition.

In response, we develop three mitigation strategies: Direct Preference Optimization (DPO) [7] to learn nuanced search decision boundaries, Attention Refinement (AR) to improve focus on relevant retrieved content, and Hierarchical Query Decomposition (HQD) to enhance multi-round tool coordination. Our experiments demonstrate that these interventions significantly reduce over-reliance behaviors, with important implications for deploying tool-augmented LLMs in real-world applications.

2 BROWSEBENCH

We develop BrowseBench comprising 1,500 queries equally distributed across five domains: Culture & Society, Science & Technology, Biology & Medicine, Environment, and Finance (300 each). Each query is annotated with 3-6 keywords and validated against expert decision paths.

Metrics. We introduce three novel metrics:

(1) Unnecessary Search Rate (USR) measures redundant searches when internal knowledge suffices;

$$\text{USR} = \frac{N_{\text{unnecessary}}}{N_{\text{total}}} \times 100\% \tag{1}$$

(2) False Information Incorporation Rate (FIIR) quantifies uncritical adoption of misleading content;

$$\text{FIIR} = \frac{N_{\text{false\_incorporated}}}{N_{\text{false\_retrieved}}} \times 100\% \tag{2}$$

(3) Task Decomposition Deviation (TDD) evaluates query planning against expert demonstrations.

$$\text{TDD} = \left|L_{\text{agent}} - L_{\text{expert}}\right| = \left|\sum_{i=1}^{T} \mathbb{I}[\text{step}_i] - L_{\text{expert}}\right| \tag{3}$$

3 EXPERIMENTAL RESULTS

As shown in Table 1, experimental result reveals three failure modes:

Excessive Conservatism. Models perform unnecessary searches due to: (a) knowledge boundary ambiguity—failing to distinguish stable theoretical foundations from evolving data; (b) contextual misdirection—temporal markers triggering retrieval for axiomatic facts; (c) cognitive triggering—epistemic qualifiers prompting defensive searches.

Over-trust in Web Sources. The Transformer attention mechanism prioritizes linguistic coherence over factual accuracy, mistaking fluency for correctness. As a result, LLMs uncritically accept web-retrieved information, with FIIR ranging from 28.0% (Claude-3.7-Sonnet) to 67.7% (Llama3.1-70B).

Lack Strategic Planning. Models lack strategic planning capabilities, exhibiting: (a) query compression, collapsing multi-faceted requirements into semantically overloaded single searches; (b) redundant parallel searches, generating overlapping queries without systematic decomposition. Task Decomposition Deviation (TDD) scores reveal systematic deficits: Claude-3.7-Sonnet (2.4) vs. Llama3.1-70B (4.7).

4 MITIGATION STRATEGIES

As shown in Table 2:

DPO. We construct preference pairs where responses with and without tools are compared. If both yield identical answers, we penalize tool use; otherwise, we reward it. Training Qwen-2.5-72B-Instruct on 10K pairs reduces USR by 7.2% on average.

AR. A query-aware mechanism dynamically adjusts attention weights: content matching search keywords receives high attention, while peripheral content requires exact constraint matching. This filters tangentially related information, reducing FIIR by 12.4% on average.

HQD. We decompose complex queries into directed acyclic graphs of sub-queries, implementing progressive acquisition that respects information dependencies. This improves multi-round coordination, reducing TDD by 1.4 on average.

Hybrid Mitigation Strategy. When combined, the strategies show synergistic effects on FIIR (14.1% reduction), as HQD's focused sub-queries enable more precise attention filtering. However, slight trade-offs occur in USR and TDD, as enhanced precision requirements make the model marginally more conservative.

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Chen Chen, Xinlong Hao, Weiwen Liu, Xu Huang, Xingshan Zeng, Shuai Yu, Dexun Li, Shuai Wang, Weinan Gan, Yuefeng Huang, et al. 2025. ACEBench: Who Wins the Match Point in Tool Usage? *arXiv preprint arXiv:2501.12851* (2025).

[3] Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024. Enhancing Noise Robustness of Retrieval-Augmented Language Models with Adaptive Adversarial Training. arXiv:2405.20978 [cs.AI] https://arxiv.org/abs/2405.20978

[4] Xinzhe Li. 2025. A review of prominent paradigms for llm-based agents: Tool use, planning (including rag), and feedback learning. In *Proceedings of the 31st International Conference on Computational Linguistics*. 9760–9779.

[5] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2022. Webgpt: Browser-assisted question-answering with human feedback, 2022. *URL https://arxiv. org/abs/2112.09332* (2022).

[6] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. 2024. ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs. In *The Twelfth International Conference on Learning Representations*.

[7] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems* 36 (2023), 53728–53741.

[8] Evan F Risko and Sam J Gilbert. 2016. Cognitive offloading. *Trends in cognitive sciences* 20, 9 (2016), 676–688.

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

[10] Hongshen Xu, Zihan Wang, Zichen Zhu, Lei Pan, Xingyu Chen, Lu Chen, and Kai Yu. 2025. Alignment for efficient tool calling of large language models. *arXiv preprint arXiv:2503.06708* (2025).

[11] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik R Narasimhan. 2025. τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. In *The Thirteenth International Conference on Learning Representations*.

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