INTERACTCOMP: EVALUATING SEARCH AGENTS WITH AMBIGUOUS QUERIES

Mingyi Deng[1*], **Lijun Huang**[2*], **Yani Fan**[2], **Jiayi Zhang** [2†], **Fashen Ren**[2], **Jinyi Bai**[3], **Fuzhen Yang**[3], **Dayi Miao**[2], **Zhaoyang Yu**[1], **Yifan Wu**[2], **Yanfei Zhang**[1], **Fengwei Teng**[1], **Yingjia Wan**[1,4], **Song Hu**[1], **Yude Li**[1], **Xin Jin**[1], **Conghao Hu**[1], **Haoyu Li**[1], **Qirui Fu**[1], **Tai Zhong**[5], **Xinyu Wang**[6], **Xiangru Tang**[7], **Nan Tang**[2], **Chenglin Wu**[1], **Yuyu Luo**[2]

[1]DeepWisdom [2]The Hong Kong University of Science and Technology (Guangzhou)
[3]Renmin University of China [4]University of California, Los Angeles
[5]Agent Universe [6]McGill University [7]Yale University

ABSTRACT

Language agents have demonstrated remarkable potential in web search and information retrieval. However, these search agents assume user queries are complete and unambiguous, an assumption that diverges from reality where users begin with incomplete queries requiring clarification through interaction. Yet most agents lack interactive mechanisms during the search process, and existing benchmarks cannot assess this capability. To address this gap, we introduce INTERACTCOMP, a benchmark designed to evaluate whether search agents can recognize query ambiguity and actively interact to resolve it during search. Following the principle of *easy to verify, interact to disambiguate*, we construct 210 expert-curated questions across 9 domains through a target-distractor methodology that creates genuine ambiguity resolvable only through interaction. Evaluation of 17 models reveals striking failure: the best model achieves only 13.73% accuracy despite 71.50% with complete context, exposing systematic overconfidence rather than reasoning deficits. Forced interaction produces dramatic gains, demonstrating latent capability current strategies fail to engage. Longitudinal analysis shows interaction capabilities stagnated over 15 months while search performance improved sevenfold, revealing a critical blind spot. This stagnation, coupled with the immediate feedback inherent to search tasks, makes INTERACTCOMP a valuable resource for both evaluating and training interaction capabilities in search agents. The code is available at https://github.com/FoundationAgents/InteractComp.

1 INTRODUCTION

Language agents have demonstrated remarkable potential across diverse domains, including code generation (Zhang et al., 2025; Hong et al., 2024b) , data analysis (Hong et al., 2024a; Li et al., 2025b;a), information retrieval (Geng et al., 2025; Song et al., 2025), and decision-making (Liu et al., 2025a; Liang et al., 2025). A notable trend is the rapid development of search agents (OpenAI, 2025d; Google, 2025b), which can handle complex user queries and gather information across the internet by performing search, browse, and reasoning actions (Mialon et al., 2023; Wei et al., 2025).

However, these advanced search agents assume user queries are complete and unambiguous. In practice, users begin with incomplete queries admitting multiple plausible interpretations, and only through interaction can the true intent be identified. Yet most search agents lack interactive mechanisms during search. Commercial agents (OpenAI, 2025d) engage in a single clarification, with no further interaction once search begins. When faced with ambiguity, agents confidently commit to assumed queries, leading to incorrect answers and wasted computational resources.

Existing benchmarks cannot assess this capability. Search benchmarks like GAIA (Mialon et al., 2023) and BrowseComp (Wei et al., 2025) provide all necessary resources upfront, enabling agents

*These authors contributed equally to this work.
†Corresponding author: Jiayi Zhang(jzhang361@connect.hkust-gz.edu.cn)

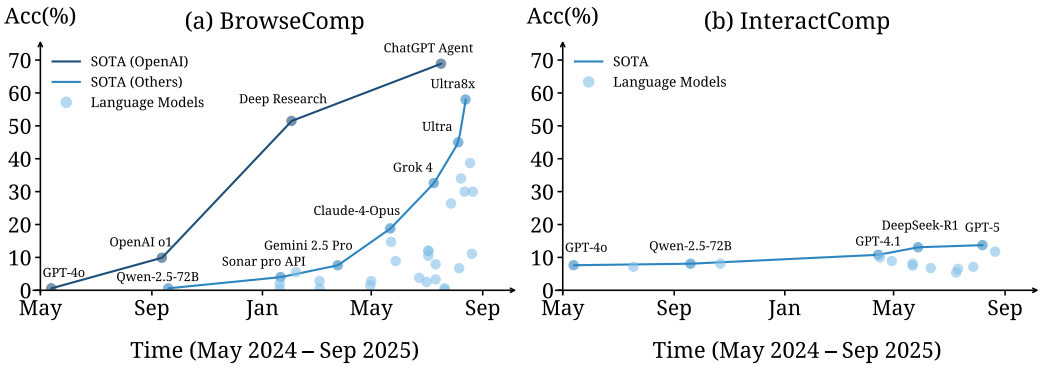

Figure 1: Despite rapid progress on complete search queries (BrowseComp: seven-fold over 15 months), agent performance on ambiguous, interaction-dependent queries (InteractComp) has stagnated around 6-14%. This growing disparity reveals a critical blind spot in agent development.

to proceed without clarifying ambiguous intent. Interaction benchmarks like IN3 (Qian et al., 2024) and Tau-Bench (Yao et al., 2024) focus on general conversation but lack grounding in verifiable search tasks. Neither addresses the question: *Can agents recognize query ambiguity and actively interact to gather disambiguating information during search?* Without proper assessment of this capability, we cannot determine whether recent advances in search agents translate to handling real-world scenarios where user intent must be uncovered rather than assumed.

Motivated by this gap, we introduce INTERACTCOMP, a benchmark designed to evaluate whether search agents can recognize ambiguity and actively interact to resolve it. Our design follows a core principle: **easy to verify, interact to disambiguate**. Questions have short, verifiable answers (1-2 words) that are answerable with enough context, yet require interaction to obtain specific details needed for disambiguation. We achieve this through a target-distractor design: questions use only shared attributes of a lesser-known target and a popular alternative, creating genuine ambiguity that search alone cannot resolve. Agents must interact with simulated users to uncover distinctive attributes not given in the initial query. INTERACTCOMP contains 210 expert-curated questions across 9 domains in both English and Chinese, validated to ensure interaction is necessary and answers are verifiable.

Systematic evaluation of 17 models confirms our design principle and reveals a striking failure pattern. When provided complete disambiguating context, models achieve strong performance with the best reaching 71.50% accuracy, validating questions are answerable once information is complete. However, even the best model achieves only 13.73% in the full interaction setting, with most models in single digits. This 5× performance gap exposes the core problem: models fail not due to search and reasoning deficits, but systematic overconfidence that prevents them from engaging in interaction despite having access to it.

Scaling experiments confirm this diagnosis. Simply increasing interaction opportunities from 5 to 20 rounds yields minimal improvement (from 14% to 20%), as models barely increase their questioning behavior. In contrast, forcing models to interact before answering produces dramatic gains (from 14% to 40%). Longitudinally, as shown in Figure 1, interaction capabilities have shown almost no improvement across all models over 15 months, while BrowseComp performance improved seven-fold during the same period. This stagnation is striking given our forced interaction experiments demonstrate the capability is latent rather than absent and readily improvable. This finding, combined with the clean reward signals from search outcomes, makes INTERACTCOMP well-suited for RLVR approaches to improve model interaction with humans.

Our contributions are threefold. (1) We introduce INTERACTCOMP, a benchmark evaluating interaction capabilities in search scenarios, with clean reward signals enabling future training approaches. (2) We provide diagnostic evidence across 17 models that interaction failure stems from systematic overconfidence rather than capability deficits. (3) We demonstrate through longitudinal analysis that interaction represents a critical blind spot in agent development, with INTERACTCOMP providing a foundation for addressing this neglected dimension.

2 RELATED WORK

Search Benchmarks and Agents. Recent benchmarks evaluate search agents along two dimensions. Web-scale search benchmarks like BrowseComp (Wei et al., 2025) assess information gathering across the entire web with complete queries, spawning variants for Chinese (Zhou et al., 2025a), multimodal content (Li et al., 2025d), and enhanced questions (Chen et al., 2025). Tool-augmented benchmarks like GAIA (Mialon et al., 2023) and WebWatcher (Geng et al., 2025) additionally require agents to handle multimedia and perform computations. These benchmarks have motivated diverse agent designs. Reinforcement learning approaches like R1-Searcher (Song et al., 2025) and Search-R1 (Jin et al., 2025) learn integrated search-reasoning patterns, while data synthesis methods like WebSailor (Li et al., 2025c) and WebExplorer (Liu et al., 2025b) enhance long-horizon capabilities. Additionally, both manually designed and self-designed search agents (Zhang et al., 2025; Zeng et al., 2025; Teng et al., 2025) have achieved strong performance through careful workflow engineering.

Interaction Benchmarks and Agents. Complementary to search benchmarks, recent work evaluates agents' interaction capabilities. SWEET-RL (Zhou et al., 2025b) proposes ColBench for multi-turn collaborative reasoning with RL-based credit assignment across turns. UserBench (Qian et al., 2025a) and UserRL (Qian et al., 2025b) create gym environments for training agents on user-centric tasks where goals are underspecified and preferences emerge incrementally. IN3 (Qian et al., 2024) and Tau-Bench (Yao et al., 2024) evaluate implicit intention understanding and tool-agent-user interaction respectively. These benchmarks collectively reveal that current models struggle with proactive clarification and user alignment—for instance, agents uncover fewer than 30% of user preferences through active questioning in UserBench.

However, these benchmarks primarily focus on general conversational settings or tool-use scenarios, lacking grounding in search tasks where intent errors lead to objectively wrong retrieval results. INTERACTCOMP differs by evaluating interaction capabilities specifically in search scenarios, where ambiguous queries must be resolved through clarification before effective retrieval can occur, and where search outcomes provide natural reward signals for training interaction strategies.

3 THE INTERACTCOMP BENCHMARK

Table 1: A task instance from INTERACTCOMP. Tasks in INTERACTCOMP comprise an ambiguous query, the simulated user's context, and a concise answer.

Question: Which team-based striking sport features two sides alternating offense and defense, where individuals sequentially hit a high-speed projectile and teammates coordinate to intercept it in the air? Outcomes depend on whether the projectile is intercepted or lands within the valid playing field. Defense relies on wide positioning and collaboration, all offensive players take turns striking, flight speeds often exceed 100 mph, protective gear is required due to impact risk, and the sport is governed by long-standing associations or leagues.	**Context:** Struck object is a plastic puck, resembling an ice hockey puck. Striking method uses a whip-like swing: the hitter lashes the puck with a long wooden rod. Defenders wield wooden boards, swinging them to block the puck in mid-air. Field is a giant fan shape, about 300 meters long with a 10–12 degree angle. Defensive teams deploy 18–20 players spread across the field to form a defensive line. Scoring is based on distance and landing point: offensive points depend on how far the puck travels and whether it touches the ground.
Distractor: *BaseBall*	**Answer:** *Hornussen*

The INTERACTCOMP dataset was constructed entirely by human annotators with the assistance of search tools and language models. While BrowseComp (Wei et al., 2025) evaluates complex search and reasoning with complete initial information, INTERACTCOMP evaluates whether agents can recognize ambiguity and actively gather necessary context through interaction during the search process. Our core design principle follows "**Easy to verify, Interact to disambiguate**": questions have concise answers that are straightforward to verify once found, yet remain ambiguous without interaction to uncover distinguishing details. This section describes the task structure (§3.1), construction methodology (§3.2), and dataset statistics (§3.3).