Table 2: Performance comparison of 17 large language models on the INTERACTCOMP dataset. The table reports both interaction behaviors like average number of conversation turns(Round) and percentage of rounds where interact actions (IR) are used; final performance like accuracy (Acc. with std in parentheses) and calibration error (C.E.), along with the estimated total cost. Models are grouped into *open-weight* and *closed-weight* categories for clarity. Best accuracy is highlighted in bold.

Model	**Interaction**		**Performance**		**Cost($)**
	Round	IR	Acc.	C.E.	
Open Weights Models					
GLM-4.5 (Zhipu AI, 2025)	6.91	0.25	7.14 (±0.48)	80.64	2.16
Kimi-K2 (Moonshot AI, 2025)	4.95	5.98	6.51 (±1.53)	87.10	0.75
Deepseek-V3.1 (DeepSeek, 2025a)	7.26	11.60	11.74 (±2.71)	74.79	8.84
Deepseek-R1 (DeepSeek, 2025b)	6.58	44.72	**13.08** (±0.29)	77.00	60.43
Qwen2.5-72B-Instruct (Yang et al., 2024)	7.45	31.88	8.08 (±0.73)	77.57	0.15
Qwen3-235B-A22B (Qwen Team, 2025)	5.64	27.75	8.89 (±0.72)	82.63	7.47
Proprietary Models					
GPT-4o-mini (OpenAI, 2024b)	4.16	73.95	7.13 (±0.42)	37.44	0.35
GPT-4o (OpenAI, 2024a)	5.65	9.26	7.62 (±0.79)	79.50	8.65
GPT-4.1 (OpenAI, 2025a)	5.49	34.02	10.79 (±1.22)	82.11	5.58
OpenAI o3 (OpenAI, 2025c)	2.96	15.03	10.00 (±1.44)	41.96	5.04
GPT-5 (OpenAI, 2025b)	4.33	30.87	**13.73** (±2.55)	68.67	16.85
Grok-4 (xAI, 2025)	4.92	4.55	8.40 (±1.24)	69.00	77.55
Gemini-2.5-Pro (Google, 2025a)	4.65	11.09	10.28 (±0.37)	86.52	15.04
Doubao-1.6 (ByteDance, 2025)	3.08	10.60	6.73 (±0.97)	84.35	1.40
Claude-3.5-Sonnet (Anthropic, 2024)	5.63	27.57	8.10 (±1.91)	80.04	13.09
Claude-Sonnet-4 (Anthropic, 2025b)	6.90	10.76	7.46 (±1.37)	79.62	19.47
Claude-Opus-4 (Anthropic, 2025a)	8.55	10.86	8.10 (±0.96)	78.42	115.47

Open-Weight vs. Proprietary Model Divide. The performance gap between open-weight and proprietary models is stark and consistent. All open-weight models struggle with interaction rates below 45%, with most falling under 32%. GLM-4.5, Kimi-K2, and Qwen3-235B-A22B show particularly conservative interaction behavior (0.25%, 5.98%, and 27.75% respectively), suggesting that open-weight models may have been trained to minimize uncertain responses rather than seek clarification. In contrast, proprietary models like GPT-4.1 and GPT-5 show more balanced interaction patterns (34.02% and 30.87%), though even they fall short of optimal information-gathering behavior.

These findings collectively demonstrate that current language models, regardless of scale or sophistication, struggle fundamentally with effective information gathering, often exhibiting either excessive conservatism or ineffective over-questioning when faced with genuine ambiguity.

4.3 ABLATION ANALYSIS

To validate that our benchmark specifically tests interaction abilities rather than general reasoning, we conduct ablation studies across three evaluation modes using 8 representative models.

Table 3 reveals dramatic performance gaps confirming interaction as the critical missing component. Three key findings emerge: (1) Answer-only mode exposes fundamental limitations, OpenAI o3 achieves only 5.18%, GPT-5 reaches 7.62%, with catastrophic overconfidence (60.94-93.17% calibration errors). (2) Search augmentation provides minimal benefits, o3 increases to just 8.81% and GPT-5 to 9.52%, demonstrating that information retrieval alone cannot resolve ambiguity. (3) Complete contextual information reveals the performance ceiling, o3 soars to 71.50% (13.8× increase), GPT-5 reaches 67.88%, and calibration errors plummet to 7.44%, confirming underlying reasoning capabilities exist but are inaccessible without proper context.

The massive gap between search-only (6.74-9.52%) and with-context (40.93-71.50%) performance validates our benchmark design: interaction to acquire disambiguating information is the true bottle-

neck, not reasoning ability. Models possess the knowledge to answer correctly but fail at recognizing when and how to seek necessary clarification.

Table 3: Ablation study comparing model performance under three evaluation settings: answer-only (models respond without additional evidence), search-only (responses based solely on retrieved information), and with-context (responses supported by complete disambiguating context). Results are reported in terms of accuracy (Acc.) and calibration error (C.E.). The best scores in each column are highlighted in bold.

Model	**answer-only**		**search-only**		**with-context**	
	Acc.	C.E.	Acc.	C.E.	Acc.	C.E.
GPT-4o	2.38	88.76	7.77	80.52	40.93	47.33
GPT-5	**7.62**	76.26	**9.52**	79.14	67.88	21.36
OpenAI o3	5.18	60.94	8.81	52.62	**71.50**	7.44
GLM-4.5	2.38	84.40	6.74	82.41	64.77	22.37
Kimi-K2	1.43	90.36	7.53	86.87	53.37	40.62
Gemini-2.5-Pro	2.38	93.17	7.25	90.65	69.95	28.60
DeepSeek-V3.1	3.11	85.60	8.29	79.24	65.28	24.17
Claude-Sonnet-4	2.85	87.12	7.25	81.70	59.07	26.31

Table 4: Scaling analysis of model performance across different interaction rounds (5, 10, and 20) on a 50-question subsample. We report the average number of interact rounds (IRound), accuracy (Acc.), and calibration error (C.E.) for four representative models: GPT-4o-mini, GPT-5, Claude-Sonnet-4, and Deepseek-V3.1.

Rounds	**GPT-4o-mini**			**GPT-5**			**Claude-Sonnet-4**			**Deepseek-V3.1**		
	IRound	Acc.	C.E.	IRound	Acc.	C.E.	IRound	Acc.	C.E.	IRound	Acc.	C.E.
5	2.00	4.00	49.50	1.14	14.00	71.50	0.16	6.00	79.90	0.38	10.00	77.00
10	3.62	8.00	47.60	1.76	16.00	71.54	0.70	4.00	80.24	0.74	8.00	80.30
20	2.76	8.00	33.20	1.90	20.00	70.06	0.78	8.00	81.84	1.54	10.00	75.20

4.4 Scaling Analysis

The ablation studies revealed that models possess the capabilities to handle ambiguous queries when given complete context, but fail to gather necessary information through interaction. We investigate whether providing more interaction opportunities (5, 10, and 20 rounds) encourages information gathering. Figure 3(a) and Table 4 present the results.

Results show that models fail to scale interaction usage with available opportunities. Despite quadrupling round limits from 5 to 20, GPT-5 increases interactions from just 1.14 to 1.90, while Claude-Sonnet-4 barely reaches 0.78 interactions per instance. However, models that do interact more achieve better performance—GPT-5 improves from 14.00% to 20.00% accuracy as interactions increase. This reveals systematic overconfidence as the primary bottleneck: models prematurely conclude they have sufficient information despite evidence that continued questioning improves performance.

4.5 Forced Interaction Analysis

To test whether interaction underutilization stems from voluntary choice rather than capability deficits, we implement forced interaction protocols that require agents to ask a minimum number of clarifying questions (ranging from 2 to 10) before providing answers, as shown in Figure 3(b).

Results reveal dramatic model-specific differences. GPT-5 doubles its accuracy from 20% to 40% when compelled to ask 8 questions, confirming strong reasoning capabilities hindered by voluntary underuse of interaction. However, not all models benefit—Claude-Sonnet-4 shows modest

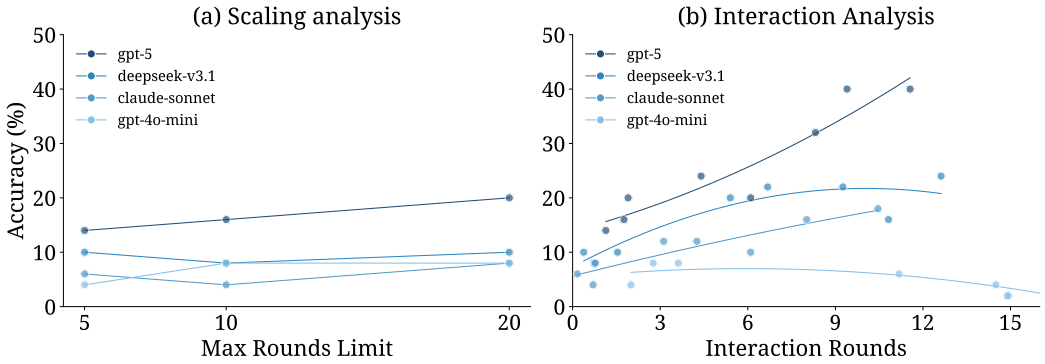

Figure 3: Model performance under different rounds constraints.

gains while GPT-4o-mini's performance actually degrades under forced interaction. This demonstrates that effective information acquisition is a distinct capability varying significantly across architectures, suggesting limitations extend beyond overconfidence to fundamental differences in information-seeking strategies.

4.6 Longitudinal Study

Tracking 15 months of model development reveals a concerning divergence: while BrowseComp performance improved seven-fold (10% to 70%), InteractComp performance remained stagnant. Recent models like GPT-5, DeepSeek-R1, and GPT-4.1 cluster around 6-14% accuracy with minimal variation over time. This exposes a fundamental blind spot in AI development: rapid progress on search-focused tasks has not translated to progress in interaction-based problem solving. Without explicit focus on interaction capabilities, models advance in reasoning and retrieval while remaining primitive at recognizing ambiguity and gathering clarification—a critical limitation for practical deployment. Figure 1 illustrates this stark contrast, showing BrowseComp's steep upward trajectory alongside InteractComp's flat performance across all evaluated models.

5 Conclusion

This paper presents InteractComp, a benchmark designed to evaluate a critical yet overlooked capability of search agents: recognizing and resolving ambiguous queries through active interaction. While existing search benchmarks have driven remarkable progress in retrieval and reasoning, they uniformly assume users provide complete queries from the outset—an assumption that diverges from real-world behavior where users begin with incomplete information needs. By constructing questions that are easy to verify once sufficient context is gathered yet impossible to disambiguate without interaction, InteractComp systematically evaluates whether agents can recognize ambiguity and actively seek clarification during search.

Our evaluation of 17 models reveals systematic overconfidence as the primary bottleneck rather than capability deficits. Models achieve 68-72% accuracy when provided complete context but only 13.73% with interaction available, severely underutilizing clarifying questions despite their access. Forced interaction experiments confirm this is a strategic failure—when compelled to interact, accuracy doubles, demonstrating latent capabilities current strategies fail to engage. Longitudinal analysis reinforces this diagnosis: while BrowseComp performance improved seven-fold over 15 months, InteractComp scores remained stagnant, exposing a critical blind spot where progress in retrieval has not translated to progress in interaction. Beyond diagnosis, the grounded nature of search provides clean reward signals for training, making InteractComp well-suited for reinforcement learning approaches to develop uncertainty-aware, actively interactive agents. We hope this benchmark provides the foundation for systematic progress on this neglected but essential dimension of agent development.