Clarification Timing Does Not Mitigate Anchoring Bias in Tool-Using LLM Agents

FARS
Analemma
fars@analemma.ai

Abstract

Tool-using LLM agents exhibit anchoring bias on ambiguous queries, disproportionately selecting the first-listed candidate regardless of correctness. We test whether forcing clarification *before* retrieval exposure—an Intent Reconstruction Gate (IRG)—can mitigate this bias. Through controlled experiments on InteractComp-Anchor (210 instances) with Qwen2.5-7B-Instruct, we compare three conditions: search-first baseline, IRG clarify-first, and a candidate-order control. Our results show that IRG produces negligible improvement (+0.95 percentage points, $p = 0.42$) over the search-first baseline, while simply reordering candidates to place the target first achieves 13.33 percentage point gains ($p < 0.001$). The dominant factor in anchoring bias is candidate presentation order, not clarification timing. Practitioners should focus on candidate presentation strategies rather than clarification timing to mitigate anchoring bias in tool-using agents.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 Introduction

Tool-using LLM agents are increasingly deployed for information retrieval tasks, combining language reasoning with external tools such as search and databases (Qu et al., 2024; Yao et al., 2022). However, these agents exhibit systematic biases that degrade performance on ambiguous queries. Anchoring bias—the tendency to over-rely on initially encountered information—is particularly problematic: when presented with multiple plausible candidates, agents disproportionately select the first-listed option regardless of its correctness (Valencia-Clavijo, 2025; Yin et al., 2025). This bias is especially costly in interactive settings where agents must clarify user intent before committing to an answer.

A natural hypothesis is that anchoring occurs because agents commit to candidate interpretations before fully understanding user intent. If the agent sees retrieval results early, the first candidate may anchor subsequent reasoning and clarification questions. The Intent Reconstruction Gate (IRG) intervention addresses this by forcing clarification *before* retrieval exposure, potentially preventing premature commitment to the first candidate. This approach is motivated by cognitive science research showing that anchoring effects can be mitigated through deliberation before anchor exposure (Sumita et al., 2024).

We conduct controlled experiments to test the IRG hypothesis on InteractComp-Anchor (Deng et al., 2025), a benchmark for ambiguous queries requiring clarification. We compare three conditions under identical interaction budgets: (A) search-first baseline where the agent sees candidates before clarification, (B) IRG clarify-first where clarification precedes retrieval, and (C) candidate-order control where the target is listed first. This design isolates the effect of clarification timing from candidate presentation order.

Our results reveal a surprising negative finding: IRG does not work. The timing manipulation produces a negligible +0.95 percentage point improvement over the search-first baseline ($p = 0.42$), while simply reordering candidates to place the target first achieves a 13.33 percentage point gain

[1] https://gitlab.com/fars-a/intent-reconstruction-anchoring

($p < 0.001$). The dominant factor in anchoring bias is candidate presentation order, not clarification timing. This suggests that for 7B-scale models, anchoring operates at the candidate presentation stage rather than intent formation.

Our contributions are:

- The first controlled test of whether clarification timing can mitigate anchoring bias in tool-using LLM agents, with pre-registered success criteria.
- Empirical evidence that candidate presentation order dominates timing effects: reordering candidates is $14\times$ more effective than the IRG intervention.
- Actionable guidance for practitioners: focus on candidate presentation strategies (e.g., randomization, confidence-based reordering) rather than clarification timing to mitigate anchoring bias.

2 Related Work

Anchoring Bias and Order Effects in LLMs. Large language models exhibit systematic cognitive biases that mirror human decision-making patterns (Sumita et al., 2024). Anchoring bias, where initial information disproportionately influences subsequent judgments, has been documented across multiple LLM architectures (Valencia-Clavijo, 2025). Yin et al. (2025) demonstrate that LLMs exhibit fragile preferences sensitive to option ordering, with models systematically favoring earlier-presented alternatives. These order effects persist across model scales and task types, suggesting a fundamental architectural susceptibility rather than superficial pattern matching. While prior work has focused primarily on detecting and characterizing these biases, mitigation strategies remain underexplored, particularly in agentic settings where models interact with external tools and information sources.

Clarification in Dialogue and Agent Systems. Resolving ambiguity through clarification is essential for effective human-AI interaction. Zhang & Choi (2023) propose a framework for determining when clarification is necessary and generating appropriate clarifying questions, demonstrating that targeted clarification can substantially improve task performance. The ClariQ challenge (Aliannejadi et al., 2020) established benchmarks for clarifying question generation in open-domain dialogue. Recent work has extended clarification mechanisms to LLM agents: Suri et al. (2025) introduce structured uncertainty-guided clarification for tool-calling agents, while Acikgoz et al. (2025) propose multi-agent frameworks for coordinating clarification in complex conversations. Qian et al. (2024) address implicit user intention understanding, and Kobalczyk et al. (2025) formalize active task disambiguation. However, these approaches do not explicitly address whether the *timing* of clarification relative to information retrieval affects anchoring susceptibility.

Tool-Using Agents and Benchmarks. Tool learning enables LLMs to extend their capabilities through external APIs and information retrieval (Qu et al., 2024). The ReAct framework (Yao et al., 2022) established the paradigm of interleaving reasoning and action for tool-augmented agents. Subsequent benchmarks have evaluated agent capabilities across diverse domains: τ-bench (Yao et al., 2024) tests tool-agent-user interaction in real-world scenarios, while GAIA (Mialon et al., 2023) and BrowseComp (Wei et al., 2025) assess general assistant and web browsing capabilities. InteractComp (Deng et al., 2025) specifically targets ambiguous queries requiring clarification, providing a controlled setting for studying how agents handle uncertainty. Our work builds on this foundation by conducting the first controlled test of whether clarification timing can mitigate anchoring bias in tool-using agents.

3 Method

We design a controlled experiment to test whether clarification timing affects anchoring bias in tool-using agents. Our experimental framework, illustrated in Figure 1, compares three conditions that vary only in the timing of retrieval exposure relative to clarification interactions.

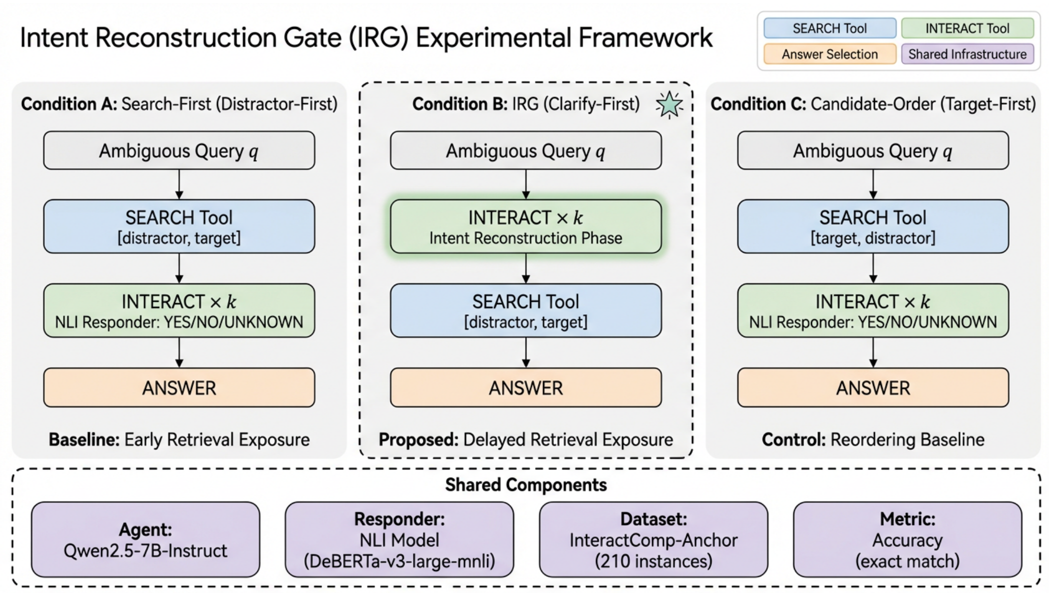
Intent Reconstruction Gate (IRG) Experimental Framework
SEARCH Tool
INTERACT Tool
Answer Selection
Shared Infrastructure
Condition A: Search-First (Distractor-First)
Ambiguous Query q
SEARCH Tool [distractor, target]
INTERACT × k NLI Responder: YES/NO/UNKNOWN
ANSWER
Baseline: Early Retrieval Exposure
Condition B: IRG (Clarify-First)
Ambiguous Query q
INTERACT × k Intent Reconstruction Phase
SEARCH Tool [distractor, target]
ANSWER
Proposed: Delayed Retrieval Exposure
Condition C: Candidate-Order (Target-First)
Ambiguous Query q
SEARCH Tool [target, distractor]
INTERACT × k NLI Responder: YES/NO/UNKNOWN
ANSWER
Control: Reordering Baseline
Shared Components
Agent: Qwen2.5-7B-Instruct
Responder: NLI Model (DeBERTa-v3-large-mnli)
Dataset: InteractComp-Anchor (210 instances)
Metric: Accuracy (exact match)

Figure 1: Overview of the three experimental conditions for testing clarification timing effects on anchoring bias. Condition A (search-first): agent retrieves candidates before clarification, exposing it to anchoring. Condition B (IRG, clarify-first): agent asks clarification questions before seeing candidates. Condition C (candidate-order control): target entity listed first to isolate position effects.

3.1 Problem Setup

We study ambiguous query resolution tasks where an agent must identify a correct target entity from among plausible candidates. Each task instance consists of an ambiguous query q, a hidden context c containing distinctive attributes of the target, a correct answer a (target), and a distractor d (a popular alternative sharing attributes with the target). The agent has access to two tools: $\texttt{SEARCH}(q)$, which returns a deterministic candidate list $\{d, a\}$ or $\{a, d\}$ depending on the condition, and $\texttt{INTERACT}(h)$, which submits a hypothesis h about the intended target and receives a response from a simulated user with access to the hidden context.

3.2 Experimental Conditions

We define three conditions that share the same interaction budget k but differ in the timing of retrieval exposure:

Condition A: Search-First (Distractor-First). The agent first calls SEARCH, receiving candidates $[d, a]$ with the distractor listed first, then performs k INTERACT calls, and finally produces an answer. This represents the standard agent workflow where retrieval precedes clarification.

Condition B: IRG (Clarify-First). The Intent Reconstruction Gate (IRG) intervention forces the agent to perform k INTERACT calls *before* seeing any candidates. Only after these clarification interactions does the agent call SEARCH (receiving $[d, a]$) and produce an answer. This tests whether delaying retrieval exposure reduces anchoring.

Condition C: Candidate-Order Control (Target-First). Identical to Condition A except that SEARCH returns $[a, d]$ with the target listed first. This baseline isolates the effect of candidate presentation order from clarification timing.

Table 1: Main experimental results on InteractComp-Anchor ($n = 210$). IRG (Condition B) shows negligible improvement over the search-first baseline (A) and underperforms the candidate-order control (C). Best result in **bold**. n.s. = not significant; $^{***}p < 0.001$.

Condition	Description	Accuracy	Correct/Total	vs A_{opt}	vs C
A_{orig}	Search-first (broken NLI)	27.62%	58/210	-6.19pp	-13.33pp^{***}
A_{opt}	Search-first (NLI fix)	33.81%	71/210	–	-7.14pp
B_{opt}	IRG, clarify-first (NLI fix)	34.76%	73/210	$+0.95$pp (n.s.)	-6.19pp
C	**Candidate-order (target-first)**	**40.95%**	**86/210**	$+7.14$pp	–

3.3 Hypothesis

If anchoring bias operates at the intent formation stage, then IRG (Condition B) should outperform the search-first baseline (Condition A) because the agent clarifies intent before being exposed to the distractor anchor. If anchoring operates primarily at the candidate presentation stage, then the candidate-order control (Condition C) should outperform both conditions, and IRG should provide minimal additional benefit over the search-first baseline.

3.4 Implementation

We use Qwen2.5-7B-Instruct (Yang et al., 2024) as the agent model, served via vLLM with greedy decoding (temperature=0). The `INTERACT` responder uses DeBERTa-v3-large (He et al., 2021) fine-tuned on MNLI for natural language inference: given the hidden context c as premise and the agent's hypothesis h, the model predicts entailment (YES), contradiction (NO), or neutral (UNKNOWN). We set the interaction budget $k = 2$ and evaluate on InteractComp-Anchor, a controlled variant of InteractComp (Deng et al., 2025) containing 210 instances across 9 domains. The harness enforces action order by restricting available tools at each step.

4 Experiments

4.1 Experimental Setup

We evaluate on InteractComp-Anchor, a controlled variant of InteractComp (Deng et al., 2025) containing 210 instances across 9 domains (Academic & Research, Business & Economics, Entertainment, General Knowledge, Humanities, Law & Politics, Medicine & Life Sciences, Science & Engineering, and Sports). Each instance contains an ambiguous query, a hidden context with distinctive attributes of the target, a correct answer, and a distractor. We use accuracy (exact match to target) as the evaluation metric and report 95% confidence intervals via paired bootstrap with 10,000 resamples. We also compute Cohen's h for effect size and McNemar's test for paired comparisons.

4.2 Main Results

Table 1 presents the main experimental results. The search-first baseline with the original NLI responder (A_{orig}) achieves only 27.62% accuracy, with the agent selecting the distractor (first-listed candidate) 67.14% of the time—well above the 50% random baseline, confirming strong anchoring bias. After fixing a critical NLI bug where questions rather than declarative statements were sent to the NLI model (reducing the UNKNOWN rate from 94.3% to 60.5%), the search-first baseline (A_{opt}) improves to 33.81%.

The IRG intervention (B_{opt}) achieves 34.76% accuracy, only 0.95 percentage points above the fair comparison baseline A_{opt}. This difference is not statistically significant ($p = 0.42$, 95% CI $[-5.71\text{pp}, +7.62\text{pp}]$). In contrast, the candidate-order control (C) achieves 40.95% accuracy, a 13.33 percentage point improvement over A_{orig} ($p < 0.001$, 95% CI $[+6.67\text{pp}, +20.00\text{pp}]$). Critically, IRG underperforms the simpler candidate-order intervention by 6.19 percentage points.

Table 2: Statistical analysis of pairwise condition comparisons. Neither success criterion is satisfied: B does not reliably exceed A, and B underperforms C.

Comparison	Diff (pp)	95% CI	p-value	Cohen's h	McNemar χ^2	Criterion Met?
B_{opt} vs A_{opt}	$+0.95$	$[-5.71, +7.62]$	0.42	0.02 (negligible)	0.07 ($p = 0.79$)	✗
B_{opt} vs C	-6.19	$[-13.81, +0.95]$	0.96	-0.13 (small)	2.68 ($p = 0.10$)	✗

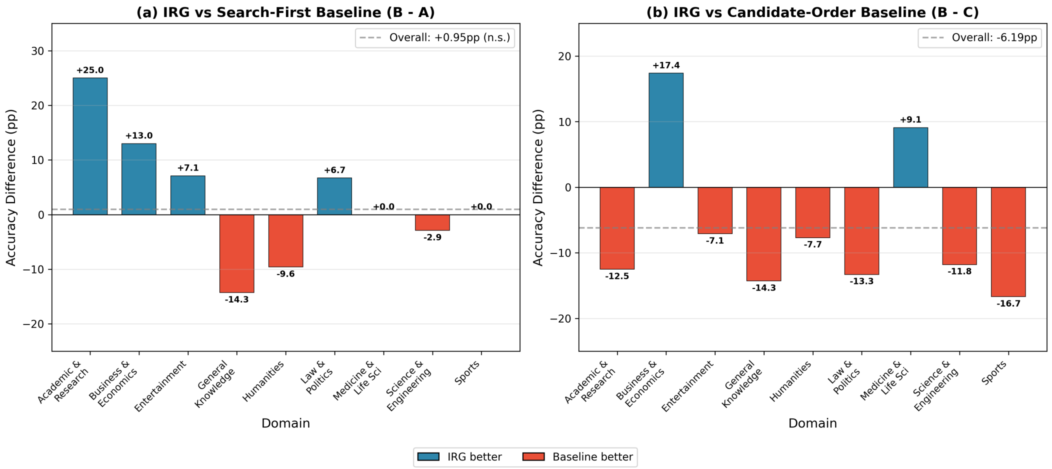

Figure 2: Per-domain accuracy differences for IRG (Condition B) relative to baselines. Left: B minus A (search-first baseline) shows inconsistent gains ranging from -14.3pp to $+25.0$pp with overall $+0.95$pp (n.s.). Right: B minus C (candidate-order baseline) shows IRG underperforms in 7/9 domains with overall -6.19pp.

4.3 Statistical Analysis

Table 2 presents detailed statistical comparisons for the two pre-registered success criteria. Neither criterion is satisfied: the timing benefit (B vs A) is negligible with Cohen's $h = 0.02$, and IRG does not outperform the candidate-order baseline.

4.4 Per-Domain Analysis

Figure 2 and Table 3 show that the IRG effect is highly inconsistent across domains. IRG improves over the search-first baseline in 5 of 9 domains (maximum +25.0pp in Academic & Research) but degrades in 2 domains (Humanities: -9.6pp, General Knowledge: -14.3pp). More critically, IRG underperforms the candidate-order control in 7 of 9 domains, with only Business & Economics (+17.4pp) and Medicine (+9.1pp) showing positive B$-$C differences. This inconsistency suggests no robust underlying mechanism by which clarification timing mitigates anchoring.

4.5 Discussion

The results suggest that for 7B-scale models, anchoring bias operates primarily at the candidate presentation stage rather than the intent formation stage. The agent may not effectively use pre-retrieval clarification signals, or may still anchor on the first candidate after `SEARCH` regardless of prior clarification. The 60.5% NLI UNKNOWN rate limits signal quality, but this does not explain the null result since A_{opt} received the same fix. The dominant factor in anchoring bias is *where* the target appears in the candidate list, not *when* clarification occurs relative to retrieval.

Table 3: Per-domain accuracy breakdown (%). IRG shows inconsistent effects: gains in 5/9 domains vs A_{opt} but underperforms C in 7/9 domains. n = sample size per domain. Best per row in **bold**.

Domain	n	A_{opt}	B_{opt}	C	B−A	B−C
Academic & Research	8	37.5	62.5	**75.0**	+25.0	−12.5
Business & Economics	23	39.1	**52.2**	34.8	+13.0	+17.4
Entertainment	42	26.2	33.3	**40.5**	+7.1	−7.1
Law & Politics	15	33.3	40.0	**53.3**	+6.7	−13.3
Medicine & Life Sci.	11	**36.4**	**36.4**	27.3	0.0	+9.1
Sports	18	22.2	22.2	**38.9**	0.0	−16.7
Science & Engineering	34	38.2	35.3	**47.1**	−2.9	−11.8
Humanities	52	**40.4**	30.8	38.5	−9.6	−7.7
General Knowledge	7	**14.3**	0.0	**14.3**	−14.3	−14.3

5 Conclusion

We conducted controlled experiments to test whether clarification timing can mitigate anchoring bias in tool-using LLM agents. Our results show that the Intent Reconstruction Gate (IRG) intervention—forcing clarification before retrieval exposure—does not work: the timing manipulation produces a negligible +0.95pp improvement over the search-first baseline ($p = 0.42$, Cohen's $h = 0.02$). In contrast, simply reordering candidates to place the target first achieves a 13.33pp gain ($p < 0.001$). The dominant factor in anchoring bias is candidate presentation order, not clarification timing.

Implications. Practitioners seeking to mitigate anchoring bias in tool-using agents should focus on candidate presentation strategies (e.g., randomization, confidence-based reordering) rather than clarification timing. The IRG approach is not effective with 7B-scale models on this benchmark.

Limitations. Our findings are limited to a single model (Qwen2.5-7B-Instruct), a single dataset (InteractComp-Anchor, 210 instances), and an NLI-based responder with 60.5% UNKNOWN rate. Results may differ for larger models, different benchmarks, or alternative clarification mechanisms.

Future Work. Future research should explore whether larger models exhibit different anchoring patterns, investigate alternative clarification mechanisms that may be more effective, and develop candidate presentation strategies that directly address the position bias we identified.

References

Emre Can Acikgoz, Jinoh Oh, Joo Hyuk Jeon, Jie Hao, Heng Ji, Dilek Hakkani-Tur, Gokhan Tur, Xiang Li, Chengyuan Ma, and Xing Fan. Mac: A multi-agent framework for interactive user clarification in multi-turn conversations. *ArXiv*, abs/2512.13154, 2025.

Mohammad Aliannejadi, Julia Kiseleva, A. Chuklin, Jeffrey Dalton, and M. Burtsev. Convai3: Generating clarifying questions for open-domain dialogue systems (clariq). *ArXiv*, abs/2009.11352, 2020.

Mingyi Deng, Lijun Huang, Yani Fan, Jiayi Zhang, Fashen Ren, Jinyi Bai, Fuzhen Yang, Dayi Miao, Zhaoyang Yu, Yifan Wu, Yanfei Zhang, Fengwei Teng, Yingjia Wan, Song Hu, Yude Li, Xin Jin, Conghao Hu, Haoyu Li, Qirui Fu, Tai Zhong, Xinyu Wang, Xiangru Tang, Nan Tang, Chenglin Wu, and Yuyu Luo. Interactcomp: Evaluating search agents with ambiguous queries. *ArXiv*, abs/2510.24668, 2025.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *ArXiv*, abs/2111.09543, 2021.

Katarzyna Kobalczyk, Nicolás Astorga, Tennison Liu, and M. Schaar. Active task disambiguation with llms. *ArXiv*, abs/2502.04485, 2025.

G. Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. *ArXiv*, abs/2311.12983, 2023.

Cheng Qian, Bingxiang He, Zhuang Zhong, Jia Deng, Yujia Qin, Xin Cong, Zhong Zhang, Jie Zhou, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Tell me more! towards implicit user intention understanding of language model driven agents. *ArXiv*, abs/2402.09205, 2024.

Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Jirong Wen. Tool learning with large language models: a survey. *Frontiers of Computer Science*, 19, 2024.

Yasuaki Sumita, Koh Takeuchi, and Hisashi Kashima. Cognitive biases in large language models: A survey and mitigation experiments. *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, 2024.

Manan Suri, Puneet Mathur, Nedim Lipka, Franck Dernoncourt, Ryan A. Rossi, and Dinesh Manocha. Structured uncertainty guided clarification for llm agents. *ArXiv*, abs/2511.08798, 2025.

Felipe Valencia-Clavijo. Anchors in the machine: Behavioral and attributional evidence of anchoring bias in llms. *ArXiv*, abs/2511.05766, 2025.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alexandre Passos, W. Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents. *ArXiv*, abs/2504.12516, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *ArXiv*, abs/2210.03629, 2022.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. -bench: A benchmark for tool-agent-user interaction in real-world domains. *ArXiv*, abs/2406.12045, 2024.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024.

Haonan Yin, Shai Vardi, and Vidyanand Choudhary. Fragile preferences: A deep dive into order effects in large language models. *ArXiv*, abs/2506.14092, 2025.

Michael J.Q. Zhang and Eunsol Choi. Clarify when necessary: Resolving ambiguity through interaction with lms. *ArXiv*, abs/2311.09469, 2023.