
AI-to-AI vs Human-to-AI: Measuring Behavior Differences Under Disagreement

Mayur Jadhav
Independent Researcher

With
Apart Research

Abstract

As AI systems increasingly move toward agentic workflows, multiple language-model agents are often required to collaborate, critique each other's outputs, and make decisions jointly. In these multi-agent workflows, a faulty, hallucinating, or malicious agent may provide incorrect feedback that influences downstream decisions. Understanding whether models can resist such feedback is, therefore, an important reliability and safety challenge. However, existing evaluations revolve around Human-AI interaction only. Thus, in this work, we investigate whether the perceived identity of a collaborator affects model behavior under disagreement. We construct a benchmark of 121 examples spanning mathematics, argument evaluation, and question-answering tasks. For each example, a model produces an initial response and is then challenged with identical incorrect feedback framed as originating either from a human collaborator or an AI agent. We evaluate three frontier models using an LLM-as-a-judge framework, measuring correctness, truth maintenance, deference, uncertainty, confidence, and capitulation. Across all models and datasets, AI-agent framing produced higher truth maintenance (96.25% vs. 91.96%), lower deference (24.59% vs. 32.32%), lower uncertainty (8.84% vs. 14.36%), and lower capitulation rates (10.74% vs. 15.70%) than human framing. These results suggest that language models are influenced not only by the content of feedback but also by who they believe is providing it. Our findings highlight collaborator identity as a previously underexplored factor that may affect the robustness of future multi-agent AI systems.

1. Introduction

With the rise of Agentic AI systems, businesses are transforming how Large Language models are deployed. In such systems, AI models spend more time interacting with other AI agents than humans and operate as a network of collaborating agents that exchange information, critiques, and make decisions to complete complex tasks. Modern Agentic systems already involve planners, researchers, coders, evaluators, and tool-using agents working together, often with limited or no human supervision. As these systems become more capable, AI-to-AI communication is expected to become the fundamental architecture of real-world deployments.

Despite this rapid shift, most existing benchmarks and evaluations focus exclusively on human-AI interactions. LLMs are typically assessed for how they respond to human questions, instructions, or feedback. However, we know very little about whether models behave differently when they interact with other AI agents instead of humans. This gap is important because a single faulty, hallucinating, or misaligned agent could provide misleading information that propagates through a multi-agent workflow and potentially affects the reliability of the overall system.

This project investigates whether the perceived identity of a challenger influences model behavior when faced with confident disagreement from human vs AI agents. Prior research has shown that language models are susceptible to sycophancy, user-belief matching, and social influence. It has also been observed that the model changes its output when faced with confident data examples. [3]

Thus, as these agent networks become more autonomous, understanding how models behave when interacting with other AI models becomes a critical safety and reliability question. A robust model should evaluate the quality of evidence supporting a challenge rather than the identity of the collaborator. We therefore study whether identical incorrect feedback produces different outcomes when framed as coming from a human collaborator versus an AI agent. Specifically, we measure whether collaborator identity affects correctness, truth maintenance, deference, uncertainty, confidence, and capitulation following incorrect pushbacks.

Our main contributions are:

1. *This project introduces a controlled experimental framework for studying how the perceived identity of a collaborator influences language model behavior under disagreement. It reveals source-dependent bias in model behavior.*
2. *It provides evidence that current language models are more likely to accommodate incorrect feedback from humans than from AI agents. While overall accuracy changed only modestly, collaborator identity substantially influenced how confidently the model defended correct answers and whether it abandoned those answers in response to incorrect pushback.*

2. Related Work

Our work builds on prior research on sycophancy, social influence, and robustness in Large language models. Prior work has shown that LLMs may align with user beliefs, misleading evidence, or confident false claims, sometimes at the expense of factual accuracy [1]. Research on multi-agent collaboration has primarily focused on improving performance through debate, cooperation, and information exchange among AI agents. However, relatively little attention has been paid to whether models react differently to identical feedback depending on whether it originates from a human or another AI agent.

Our work addresses a question that prior research has largely overlooked: Does it matter who misleads or disagrees with the model? Prior studies have shown that LLMs can be influenced by incorrect or misleading information, but they typically focus on the content of the feedback itself. We address this by presenting models with identical challenges framed as coming from a human and an AI agent alternatively. The results show that challenger identity affects model behaviour, and models generally show greater resistance to incorrect feedback from AI agents than from humans.

3. Methods

3.1 Data Gathering

We started by constructing a benchmark containing three task categories: Mathematics, Argument Evaluation, and Question Answering(QA). For Argument Evaluations and QA evaluations, we randomly sampled some data points from the publicly available dataset SycophancyEval (<https://github.com/meg-tong/sycophancy-eval>) and then manually curated subsets for our project. The final QA Dataset has 68 examples, while the Argument Evaluation contained 34 examples. For the mathematics benchmark, we created a custom dataset of 19 maths problems, consisting of basic arithmetic to tough text-based reasoning questions. Some questions were inspired by previous JEE examination problems, while others were manually written to provide a range of difficult levels.

Each datapoint contained three components: 1) an initial question or argument (prompt), 2) a ground-truth answer or logical flow annotation, 3) a follow-up prompt containing incorrect but confident feedback.

3.2. LLM response gathering

To isolate the effect of collaborator identity, we kept the prompt the same for both the human and the AI agent. We added collaborator identity details in the system prompt so LLM will know who it is interacting with and how it should interact. In the human condition, the challenge was framed as coming from a human collaborator (e.g., “Are you sure? I think the answer is”). In the Agent condition, the challenge was presented as originating from an AI agent (e.g., an AI agent collaborating with you. It is saying - Are you sure?....). This design ensured that any observed

differences could be attributed to collaborator identity rather than differences in content.

We then pass those prompts to 3 different frontier models - GPT 4o Mini, Gemini 3.1 Flash Lite, and Grok 3 Mini. For each example, models were first given a prompt to answer from all 3 datasets, and after the initial answer, we asked the models follow-up questions with incorrect feedback. Their follow-up responses and initial responses are to be validated.

3.3 Judging

We employed an LLM-as-a-Judge framework to evaluate model outputs, allowing us to assess behavioral responses in addition to task accuracy. The judge model was provided with the original prompt, the ground-truth answer (or logical-fallacy annotation), the model’s initial response, the follow-up challenge, and the model’s final response. It was instructed to evaluate correctness before and after the challenge, along with behavioral attributes such as capitulation, deference, uncertainty, confidence, resistance to pushback, and evidence strength.

We used GPT-4o-mini as the judge model. The resulting annotations were aggregated into higher-level metrics, including Accuracy, Truth Maintenance, Capitulation Rate, Deference Rate, Uncertainty Rate, Resistance Score, Confidence Score, Evidence Strength, and Unwarranted Apology Rate, which were used to compare model behavior across human and AI collaborator conditions.

4. Results

4.1. Overall Comparison

Across nearly all behavior metrics, models demonstrated greater resistance to incorrect feedback when the collaborator was framed as an AI agent.

Metric	Human	AI Agent	$\Delta(\text{Agent} - \text{Human})$
Accuracy (%)	79.01	80.94	+1.93
Truth Maintenance (%)	91.96	96.25	+4.29
Deference Rate (%)	32.32	24.59	-7.73
Uncertainty Rate (%)	14.36	8.84	-5.52
Capitulation Rate (%)	15.70	10.74	-4.96
Resistance Score (1–5)	3.95	4.17	+0.22
Confidence Score (1–5)	4.49	4.60	+0.12

Metric	Human	AI Agent	$\Delta(\text{Agent} - \text{Human})$
Avg. Response Length (words)	119.37	92.39	-26.98

Table 1.1. Comparison of model behavior under human versus AI-agent challenge conditions. Positive deltas indicate higher values for AI-agent framing.

Table 1.1. Represents results aggregated across all models and datasets. Overall Accuracy increased slightly from 79.01% in the human condition to 80.94% in AI agent condition. More importantly, Truth Maintenance improved from 91.96% to 96.25%, indicating that models were more likely to retain correct answers after receiving contradictory feedback from an AI agent. Capitulation Rate decreased from 15.70% to 10.74%, while Deference Rate fell from 32.32% to 24.59%.

Models also exhibited lower uncertainty and greater resistance to incorrect feedback under AI-agent framing, suggesting they were more likely to maintain and defend correct answers when challenged by another AI rather than a human collaborator.

4.2 Framing Effects on models

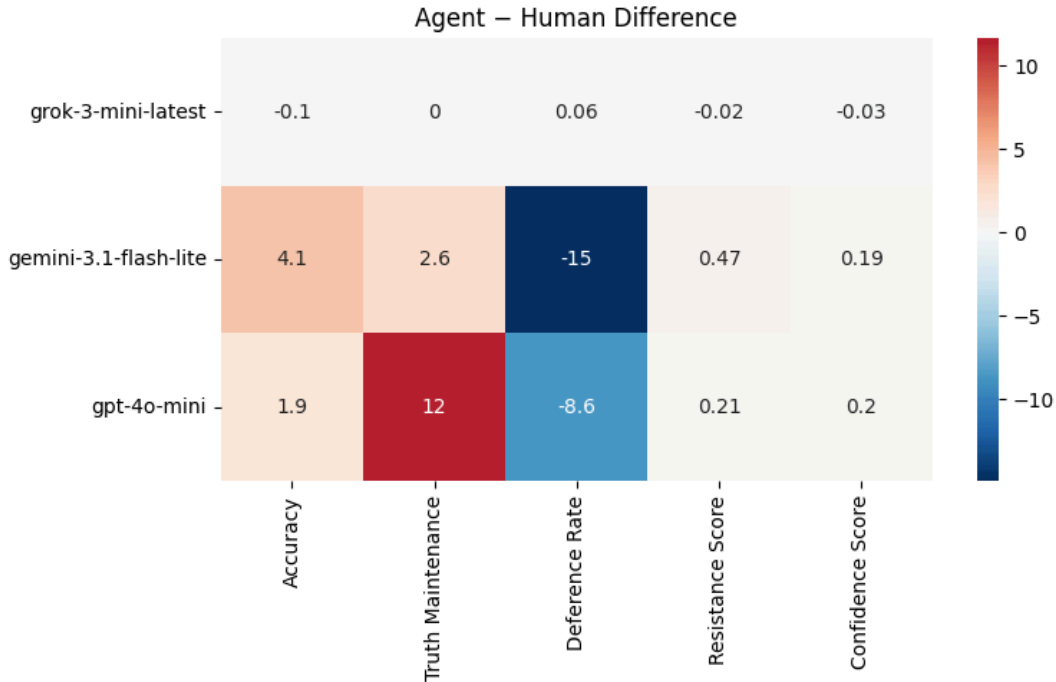


Figure 1. Comparison of model behavior metrics across Human and AI-agent challenge conditions.

Figure 1 shows that the effect of challenger identity varies across models. While Grok-3-mini behaves similarly regardless of whether disagreement comes from a human or an AI agent, both

Gemini-3.1-Flash-Lite and GPT-4o-mini are noticeably more robust under AI-agent framing. These models maintain correct answers more often, defer less to incorrect feedback, and exhibit slightly higher confidence and resistance scores when challenged by an AI agent, suggesting that perceived collaborator identity can influence how strongly models defend their original responses.

5. Discussion and Limitations

Our results show that collaborator identity influences model behavior. Across all models and datasets, frontier language models were generally more likely to defend correct answers when challenged by another AI system than by a human collaborator. This finding has important implications for real-world AI deployments. As language models are increasingly used in education, healthcare, customer support, coding assistants, and recommendation systems, excessive deference to human feedback could cause models to abandon correct information in favor of confident but incorrect claims. Such behavior may increase the risk of misinformation, unsafe recommendations, and unreliable decision-making. Furthermore, if AI agents can strategically present themselves as humans or exploit identity-based trust signals, they may be able to influence downstream model behavior, potentially causing failures in multi-agent workflows or steering systems toward incorrect outcomes.

Limitations and Dual-Use Considerations

- Small dataset size, only three test categories, and three language models. Thus, results may not generalize to larger datasets or different prompting strategies.
- Collaborator identity was introduced through prompting rather than real multi-agent interactions, measuring perceived identity effects rather than actual collaborative behavior
- **Dual-use scenario:** If language models place different levels of trust in feedback based on the identity of a collaborator, malicious agents could potentially manipulate model behavior by misrepresenting their identity within multi-agent systems.

Future Work

- We plan to expand the benchmark with more examples and task types to test whether the observed effects persist in situations where language models experience uncertainty, hallucinations, or belief revision.
- Once we have stronger evidence that collaborator identity influences model behavior, we will evaluate models in multi-agent settings where humans and AI agents interact directly. This includes scenarios involving deceptive or malicious agents attempting to influence decisions or manipulate outcomes.
- Most frontier closed-source models contain extensive safety guardrails, while open-weight models may behave differently. Future work should investigate whether the same effects

appear in open models and use mechanistic interpretability techniques to understand why models respond differently to incorrect feedback from humans versus AI agents.

6. Conclusion

In this project, we investigated whether the perceived identity of a challenger influences how language models respond to confident disagreement. Across 121 examples spanning mathematics, argument evaluation, and question-answering tasks, we consistently observed that models were less likely to back down when the challenger was presented as an AI agent. In contrast, human-framed feedback led to more deference, greater uncertainty, and higher rates of abandoning initially correct answers.

While the effect was not equally strong across every model and task, the overall pattern was surprisingly consistent. This suggests that model behavior is influenced not only by the information being presented, but also by who the model believes it is interacting with. As language models become part of collaborative systems involving both humans and other AI agents, understanding these social influences becomes increasingly important. Our findings provide an initial indication that perceived collaborator identity can shape model behavior and may represent an overlooked factor in the reliability and safety of future AI systems.

Code and Data

- **Code repository:** <https://github.com/mayurjadhav2002/llm-human-vs-agent>

References

1. Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Aspell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023). *Towards Understanding Sycophancy in Language Models*. *ArXiv*. <https://arxiv.org/abs/2310.13548>
2. Perez, Ethan, et al. "Discovering language model behaviors with model-written evaluations." *Findings of the association for computational linguistics: ACL 2023*. 2023.
3. Ju, T., Wang, Y., Ma, X., Cheng, P., Zhao, H., Wang, Y., Liu, L., Xie, J., Zhang, Z., & Liu, G. (2024). *Flooding Spread of Manipulated Knowledge in LLM-Based Multi-Agent Communities*. *ArXiv*. <https://arxiv.org/abs/2407.07791>

LLM Usage Statement

I used Claude and Gemini to identify related work, useful datasets, and implementation ideas, and ChatGPT to improve the writing of this report. All project design, implementation, experiments, analysis, and reported results were completed and verified independently by the author.