Hallucinations in Pipelines: Single-Hop Emulation of Multi-Agent Propagation

Anonymous ACL submission

Abstract

Large language models(LLMs) are increasingly deployed in multi-agent systems and pipeline architectures in which the output of one model becomes the input to another. Although LLM capabilities have improved rapidly, hallucinations remain a persistent source of risk. In such settings, hallucinations are not merely localized model errors but potential systemic hazards: incorrect upstream outputs may be accepted, propagated, or even transformed by downstream models.

In this work, we study this phenomenon through a controlled single-hop simulation, where upstream answers, possibly containing hallucinations, are injected into prompts to probe downstream responses. We examine how agents react to potentially incorrect inputs under both explicit conditions (where the model is directly asked whether it agrees with the prior answer) and implicit conditions (where the model is asked to continue reasoning without being queried about agreement). We further evaluate the influence of role-play prompts, a common design pattern in multi-agent systems, on an agent's susceptibility to hallucinated inputs. Demonstrate the potential threat posed by hallucinations in a system composed of multiple LLMs.

1 Introduction

LLMs have rapidly increased in capability and are now widely deployed in multi-agent systems and pipeline architectures (Li et al., 2023a) in which the output of one model becomes the input to another. Such systems enable task decomposition, collaborative work allocation, automated verification, and iterative editing. While these architectures offer promising advantages in reliability and scalability, they also introduce a risk since the structure is more complex than single-model settings.

A central source of risk is hallucination. Prior research, see Ji et al. (2023) for a survey, has predominantly treated hallucinations as localized errors, including fabricated facts, incorrect reasoning, or invented citations, specific to individual models. However, when multiple LLMs operate in a pipeline, hallucinations cease to be isolated deviations but may evolve into a system risk. Incorrect upstream outputs can be accepted, reinforced, modified, or even transformed by downstream models, potentially leading to error accumulation or what can be viewed as hallucination propagation throughout the entire pipeline.

Figure 1: Illustration of Single-Hop Emulation

In practice, to combat this, multi-agent and pipeline systems often rely on one or more downstream verification agents responsible for reviewing upstream outputs (Zheng et al., 2023). These verifiers are commonly guided by prompt structures that assign them specific social roles (Shanahan et al., 2023), such as a teacher evaluating a student, a supervisor reviewing a subordinate, or co-authors editing one another's drafts. Such role-play prompts have become a widely adopted design pattern in recent multi-agent systems, intended to elicit particular review behaviors or reasoning styles. Yet the lack of formal investigation means it remains unclear whether role-play prompts introduce systematic biases that affect how models evaluate hallucinated inputs.

Another key dimension concerns the interaction format used by downstream agents.

Explicit judgment: The model is directly asked whether it agrees with the upstream answer (e.g., "Do you agree or disagree with the previous answer?") (Kaesberg et al., 2025).

Implicit judgment: The model is not asked

to evaluate agreement; instead, it is instructed to continue reasoning or provide an improved answer (Madaan et al., 2023).

This implicit format is closer to common agent/pipeline implementations, where downstream components typically refine or build on upstream outputs rather than explicitly stating agreement (Wu et al., 2024). However, it is unclear whether this mode diminishes the model's ability to recognize hallucinations or encourages partial acceptance or preservation of erroneous information.

Motivated by these issues, this paper introduces a controlled single-hop simulation framework, as shown in Fig. 1, designed to systematically study hallucination propagation in multi-agent systems. We inject upstream answers, potentially containing hallucinations, directly into downstream prompts and measure how agents respond under different prompt configurations.

To characterize behavioral patterns of hallucination propagation, we combine agreement ratios, chi-square statistical testing, and human annotations that label how the downstream response changes the information content of the upstream response.

The contributions of this paper are as follows. First, we show that both judgment mode and role-play framing systematically shape downstream agreement behavior, including the acceptance of hallucinated upstream answers. Second, we quantify these prompt-induced shifts with chi-square tests against baseline distributions, and find that distributional deviations are generally larger under the pipeline-realistic implicit setting and vary across models. Third, going beyond binary stance, we characterize how downstream outputs transform hallucinated inputs at the content level (e.g., preserving, reducing, or expanding upstream information), revealing multiple propagation patterns beyond simple acceptance or rejection.

2 Related Work

Research on hallucinations in language models has largely focused on single-model settings, addressing their definition, evaluation, and detection. Ji et al. (2023) provide a comprehensive survey of hallucinations in natural language generation, covering their manifestations across tasks such as question answering and summarization, as well as commonly used evaluation protocols. Benchmark datasets have been proposed to assess factual correctness from different perspectives, including TruthfulQA (Lin et al., 2022), which evaluates models under prompts that elicit common human misconceptions, and QAGS (Wang et al., 2020), which measures factual consistency in summarization via a question–answering protocol. More recently, HaluEval (Li et al., 2023b) introduces a large-scale, human-annotated benchmark for evaluating LLMs' ability to identify hallucinated content, constructed through a ChatGPT-based sampling-then-filtering pipeline. In addition, several approaches aim to improve or detect factual reliability, such as retrieval-augmented generation (Lewis et al., 2020), which incorporates external documents for knowledge-intensive tasks, and SelfCheckGPT (Manakul et al., 2023), which detects hallucinations based on inconsistencies across multiple sampled outputs. While these works establish important foundations, they primarily treat hallucinations as errors in single-model outputs, without considering how such errors propagate through larger systems.

In parallel, large language models are increasingly deployed in multi-agent systems, where multiple agents interact through intermediate outputs. Frameworks such as CAMEL (Li et al., 2023a) and AutoGen (Wu et al., 2024) organize agent collaboration through role-based interactions, while MetaGPT (Hong et al., 2024) encodes standardized procedures into prompt sequences and explicitly notes that naively chaining LLMs can lead to cascading hallucinations. In these systems, downstream agents are often assigned the role of verification agents to review or revise upstream outputs. Related work has explored the feasibility of LLM-based evaluators: G-Eval (Liu et al., 2023) proposes an evaluation framework using chain-of-thought and form-filling prompts and analyzes biases toward LLM-generated text, while Zheng et al. (2023) systematically study LLM-as-a-judge, identifying position and verbosity biases and introducing MT-Bench and Chatbot Arena for evaluation. Shinn et al. (2023) further explores iterative improvement through linguistic feedback and episodic memory. Despite these advances, prior work largely assumes verification agents to be reliable correctors and offers limited analysis of how downstream agents behave when upstream inputs already contain hallucinations.

Role-play prompting is a common design pattern in multi-agent systems, where agents are instructed to adopt specific roles to guide their behavior. Shanahan et al. (2023) discuss role play

as a conceptual lens for understanding language-based agents, while Sharma et al. (2024) provide empirical evidence that language models may exhibit sycophancy, i.e., a tendency to agree with user-provided viewpoints even when they are factually incorrect, partly due to human feedback preferences. More recently, Zhao et al. (2025) analyze the risks of role-play prompting and show that different role assignments can affect the prevalence of biased or harmful outputs. Together, these studies suggest that role framing and social cues can influence model judgments. However, they do not examine how role-play prompting interacts with hallucinated upstream inputs in multi-agent pipelines, nor how such interactions affect downstream judgments and subsequent information changes.

3 Methodology

3.1 Single-Hop Emulation

In practical multi-agent systems and pipelines, agents typically collaborate through sequential exchange of intermediate outputs. The output produced by an upstream agent is directly passed to a downstream agent, which often has no access to the original evidence, generation process, or external knowledge sources. In such settings, hallucinated information introduced upstream may be accepted, retained, or transformed by downstream agents, creating potential systemic risks.

To study this phenomenon in a controlled manner, we adopt single-hop emulation as our core methodological setting. As illustrated in Figure 1, the output of an upstream agent (Agent $n-1$), which may contain hallucinations, is explicitly injected into the prompt of a downstream agent (Agent n). The downstream agent is then asked to process or revise the upstream answer based only on the original question and the provided response, without being informed of its correctness. This design reflects realistic deployment scenarios in which verification and other agents must operate under uncertainty.

Compared to full multi-hop pipelines, single-hop emulation offers improved controllability and interpretability. By restricting the analysis to a single interaction, we avoid confounding effects caused by cascading errors or long-range dependencies. This allows observed behavioral differences to be more directly attributed to the downstream agent's judgment behavior and prompt design, rather than to accumulated effects across multiple steps.

Furthermore, single-hop emulation serves as a minimal analytical unit for examining how hallucinations propagate and change within multi-agent systems. As shown in Fig. 1, our focus is not on end-to-end pipeline performance, but on how hallucinated content is handled when transferred from Agent $n-1$ to Agent n, and how the resulting output may influence a subsequent agent (Agent $n+1$). This controlled setting enables systematic analysis of whether hallucinated information is preserved, reduced, or reformulated during downstream processing.

3.2 Prompt Design

We investigate how downstream agent behavior is shaped by prompt-level design when processing potentially hallucinated upstream inputs. All prompts are constructed using question–answer (QA) style data, which provides a controlled setting for distinguishing factual and hallucinated inputs while keeping task semantics fixed.

We use two QA-style datasets, HaluEval (Li et al., 2023b) and MHD.[1] Both datasets are from publicly available QA resources and do not involve user-collected personal data or personally identifying information.

HaluEval is a multi-task hallucination evaluation benchmark that includes QA, knowledge-based dialog, and summarization subsets. In this work, we only use its QA subset, which is constructed using HotpotQA (Yang et al., 2018) questions as seed data and augmented with hallucinated answers generated by an LLM.

MHD is derived from an open-domain QA corpus WikiQA(Yang et al., 2015) where questions come from web query logs and answers are drawn from Wikipedia sentences. It forms positive instances using the original question-answer pairs and creates challenging negative instances by replacing the key answer span with automatically generated alternatives produced via multiple generation strategies.

Each data instance consists of a question paired with either a correct or an incorrect answer. A true fact input is formed by combining the question with its verified correct answer, whereas a hallucination input is formed by pairing the same question with a plausible but factually incorrect answer. For example, given the question "What nationality was James Henry Miller's wife?", the

[1] MHD is our dataset described in a paper currently under review.

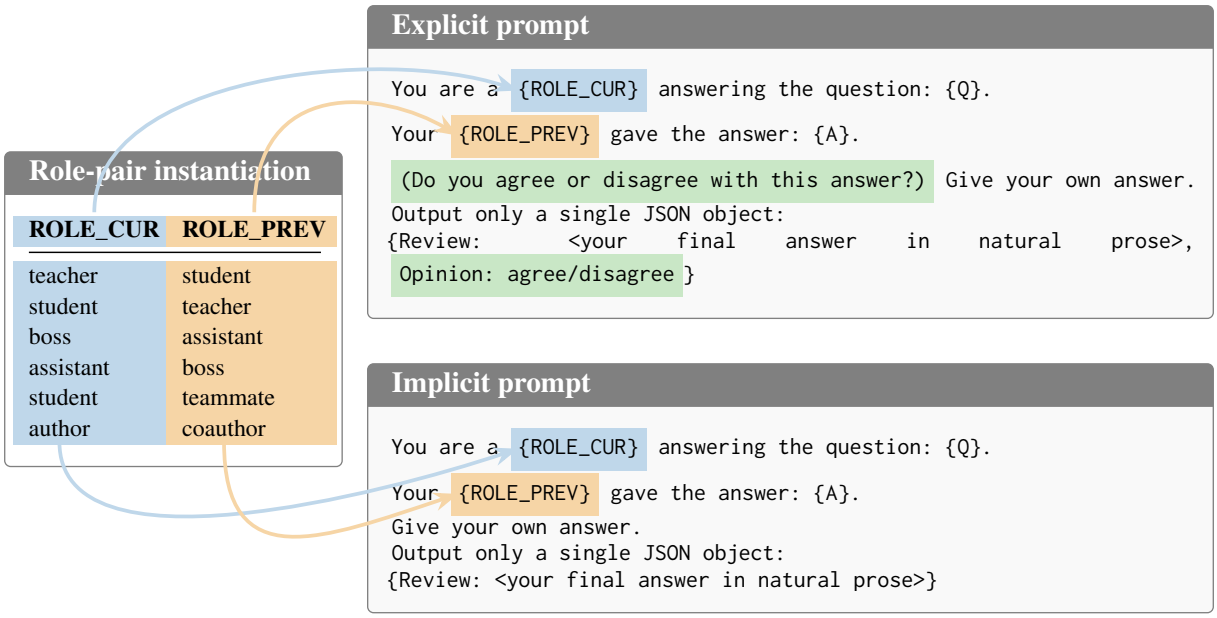


Figure 2: Prompt templates in explicit/implicit setting with role-pair instantiation. {ROLE_CUR} and {ROLE_PREV} denote the downstream and upstream roles. The downstream agent evaluates the upstream answer A to Q. The explicit template additionally includes the agreement query highlighted in green.

correct answer is "American", while the incorrect answer states that James Henry Miller's wife was British. HaluEval is balanced with a 1:1 ratio of True Fact to Hallucination samples, while MHD is skewed toward negative cases with a 1:7 ratio. For both datasets, we used the first 500 instances for evaluation. By holding the question constant and varying only the correctness of the upstream answer, we isolate downstream responses to factual versus hallucinated content.

Within this setting, we manipulate two aspects of the downstream prompt: role-play prompting and whether the agent is required to make an explicit judgment about the upstream answer.

3.2.1 Role-Play Prompt

To examine the effect of social and hierarchical framing, we employ role-play prompting, in which the downstream agent is instructed to assume a specific role relative to the upstream agent. We consider six representative role relationships commonly used in multi-agent systems: *teacher–student*, *student–teacher*, *boss–assistant*, *assistant–boss*, *student–teammate*, and *author–coauthor* as shown in Fig. 2. Here, *X–Y* denotes a role pair where the downstream agent plays *X* and the upstream agent plays *Y*. These roles cover both asymmetric relationships with implicit authority differences and symmetric peer-level collaborations. We also include a *plain* prompt without any role-play framing, where the downstream agent simply answers the question given the upstream answer as context, to isolate the effect of role-play prompts.

In each condition, the role specification is explicitly stated at the beginning of the prompt, while all other components, including task description, input content, and output format, remain unchanged. This design allows us to isolate the influence of role framing on downstream judgment behavior without introducing confounding changes to task structure or model input.

3.2.2 Explicit/Implicit Judgment

In addition to role framing, we distinguish between explicit and implicit judgment conditions, which correspond to different downstream processing roles in multi-agent pipelines.

Under the explicit judgment condition, the downstream agent is directly asked whether it agrees or disagrees with the upstream answer and is required to output both a natural language response and a discrete judgment label. This setting closely mirrors the role of a verification agent in practical multi-agent systems, where a downstream agent is explicitly tasked with validating or rejecting upstream outputs. As a result, the agent's judgment can be directly observed from its output.

Under the implicit judgment condition, the downstream agent is not asked to state agreement or disagreement. Instead, it is instructed only to provide its own answer or revision. This setting reflects common pipeline scenarios in which downstream agents are expected to continue task execution rather than perform explicit verification. In this case, the agent's judgment toward the upstream input is not directly observable.

To enable consistent analysis across both conditions, we introduce an opinion classifier to infer implicit judgments from textual reviews. Specifically, we use review–opinion pairs naturally generated under the explicit judgment condition to train a binary classifier that predicts whether a downstream response implicitly agrees or disagrees with the upstream answer. We evaluated multiple encoder and pooling configurations and selected a DeBERTa-base model with max pooling, which achieved the best performance (F1 = 0.926) on a held-out portion of the review-opinion pairs.

This classifier is applied to downstream responses generated under implicit prompts, allowing us to infer agent judgments without altering the prompt design. As a result, we can directly compare downstream behavior under explicit and implicit judgment conditions while preserving realistic multi-agent interaction patterns. In order to establish the extent to which the LLM may frame its review differently when not asked for an explicit opinion, which could affect the performance of the classifier, we hand-annotated a sample of 100 downstream responses generated under implicit prompts and found that the F1 score for the selected classifier was 0.88.

4 Experiments

4.1 Agreement and Rejection Behavior of Agents

We repeat each experimental condition three times and report the mean results in Tables 1 and 2. All experiments are run with Gemini-2.0-flash(denoted as Gemini-2.0 hereafter) and gpt-4o-mini via their official APIs. To assess run-to-run stability, we compute inter-run agreement using κ, which indicates substantial consistency across settings ($\kappa \in [0.704, 0.801]$). We also compute standard errors(SE) for the agreement metrics in Table 1 across the three runs, and the maximum standard error observed is 0.028. Overall, κ and SE indicate consistent trends across runs.

Table 1 reports downstream agents' agreement behavior under different prompt settings using three metrics: *agr* (overall agreement rate), *pos agr* (agreement on true-fact inputs), and *neg agr* (agreement on hallucination inputs). We report agreement ratios rather than accuracy score because our goal is to reveal prompt-induced stance tendencies (i.e., how readily an agent agrees or disagrees). Among them, *neg agr* is the hallucination acceptance rate and directly indicates whether incorrect upstream information is passed to downstream stages. A stronger agent should exhibit a higher *pos agr* and a lower *neg agr*. For implicit prompts, where outputs contain no explicit agree/disagree label, judgments are inferred from reviews using the opinion classifier described in Section 3.2.2.

We first note that implicit prompting increases hallucination pass-through in a broad and often substantial manner. Across both datasets and models, implicit prompting typically raises *neg agr*, indicating increased acceptance of hallucinated upstream answers. This effect is pronounced on MHD. This effect is particularly evident on MHD: averaged over all prompts, *neg agr* increases by +0.16 for Gemini-2.0 and by +0.25 for gpt-4o-mini. HaluEval shows the same overall trend with smaller average shifts (Gemini-2.0: +0.04; gpt-4o-mini: +0.12), with only minor decreases for Gemini-2.0 under student_group_project and joint_article_coauthor. These consistent increases show that when downstream agents are not explicitly asked to verify, they more readily preserve upstream content, including hallucinated statements.

The implicit setting also often raises *agr*, but the key risk signal is that the increase is accompanied by a marked rise in *neg agr*. For instance, on HaluEval with gpt-4o-mini, plain *agr* rises from 0.39 to 0.50 while neg agr rises from 0.18 to 0.33, indicating that increased agreement is not limited to true-fact inputs.

Second, we note that role asymmetry induces directional judgment biases that are strongest on hallucinated inputs. Role-play framing produces consistent directional effects in inverse role pairs, especially under implicit prompting. On HaluEval with Gemini-2.0 (Implicit), student_teacher exhibits substantially higher agreement than teacher_student (*agr* 0.72 vs. 0.43), and, critically, a much higher hallucination acceptance rate (*neg agr* 0.55 vs. 0.28). Similar asymmetry appears on MHD with Gemini-2.0 (Implicit): assistant_boss shows higher *agr* than boss_assistant

Dataset	model	prompt	Explicit			Implicit		
			agr	pos agr	neg agr	agr	pos agr	neg agr
HE	Gemini-2.0	teacher_student	0.47	0.68	0.26	0.43	0.58	0.28
		student_teacher	0.58	0.80	0.35	0.72	0.90	0.55
		boss_assistant	0.45	0.66	0.24	0.44	0.64	0.25
		assistant_boss	0.50	0.73	0.27	0.58	0.79	0.38
		student_group_project	0.42	0.63	0.21	0.36	0.52	0.20
		joint_article_coauthor	0.55	0.78	0.33	0.45	0.63	0.26
		plain	0.42	0.63	0.21	0.39	0.57	0.22
	gpt-4o-mini	teacher_student	0.38	0.57	0.19	0.47	0.67	0.27
		student_teacher	0.43	0.62	0.23	0.61	0.80	0.41
		boss_assistant	0.40	0.60	0.20	0.56	0.77	0.35
		assistant_boss	0.40	0.58	0.22	0.53	0.72	0.34
		student_group_project	0.34	0.52	0.16	0.41	0.59	0.22
		joint_article_coauthor	0.37	0.54	0.20	0.51	0.69	0.32
		plain	0.39	0.59	0.18	0.50	0.68	0.33
MHD	Gemini-2.0	teacher_student	0.38	0.72	0.29	0.51	0.86	0.41
		student_teacher	0.55	0.87	0.46	0.76	0.91	0.73
		boss_assistant	0.36	0.65	0.28	0.51	0.73	0.44
		assistant_boss	0.45	0.78	0.36	0.63	0.80	0.59
		student_group_project	0.34	0.69	0.25	0.45	0.72	0.38
		joint_article_coauthor	0.41	0.74	0.31	0.45	0.68	0.38
		plain	0.42	0.72	0.34	0.55	0.77	0.49
	gpt-4o-mini	teacher_student	0.31	0.67	0.21	0.53	0.82	0.45
		student_teacher	0.43	0.79	0.33	0.62	0.83	0.56
		boss_assistant	0.33	0.69	0.24	0.56	0.78	0.51
		assistant_boss	0.36	0.73	0.26	0.58	0.81	0.52
		student_group_project	0.27	0.65	0.17	0.52	0.77	0.44
		joint_article_coauthor	0.35	0.74	0.25	0.56	0.79	0.49
		plain	0.33	0.67	0.24	0.53	0.77	0.47

Table 1: Agreement ratios of downstream agents. Darker cells indicate larger values. We report *agr* (overall agreement rate), *pos agr* (agreement on true-fact inputs), and *neg agr* (agreement on hallucination inputs). For an ideal verification agent that perfectly distinguishes hallucinations, *pos agr* = 1 and *neg agr* = 0, whereas *agr* depends on the proportion of true-fact versus hallucination inputs.

(0.63 vs. 0.51) and higher *neg agr* (0.59 vs. 0.44). These contrasts indicate that hierarchical framing changes downstream agents' default trust in upstream answers and that the effect is most consequential when upstream content is incorrect.

Peer role setting are not consistently close to the plain baseline. The two peer settings (student_group_project and joint_article_coauthor) show high variability relative to plain, sometimes increasing *agr* and sometimes reducing it, depending on the dataset, model, and prompting style. For instance, on HaluEval with Gemini-2.0, joint_article_coauthor under Explicit produces a much higher *neg agr* than plain (0.33 vs. 0.21), whereas student_group_project under Implicit slightly lowers *neg agr* relative to plain (0.20 vs. 0.22). On MHD with gpt-4o-mini, joint_article_coauthor in Implicit slightly increases hallucination acceptance compared to plain (*neg agr* 0.49 vs. 0.47), while student_group_project moves in the opposite direction (0.44 vs. 0.47). Overall, peer settings does not act as a stable neutral condition, its effect depends on the context and is directionally inconsistent.

Table 2 quantifies how strongly each role-play condition shifts the outcome distribution relative to plain baselines, with values above 7.815 highlighted.

First, implicit prompts produce much larger distributional shifts than explicit prompts, particularly for Gemini-2.0. On HaluEval with Gemini-2.0, relative to the plain baseline, chi-square increases from 61.09 (Explicit) to 266.25 (Implicit) for student_teacher, and remains highly significant for assistant_boss (17.4 to 87.68). On MHD with Gemini-2.0 (Implicit), multiple roles exceed the threshold by wide margins. This indicates structural changes in the judgment distribution under implicit prompting, not merely small fluctuations in *agr*.

Second, model sensitivity differs. gpt-4o-mini is comparatively stable under Explicit on HaluEval, all plain-target chi-square values are below 7.815, whereas under Implicit it exhibits significant shifts

prompt	Dataset	model	T–S	S–T	B–A	A–B	SGP	JAC
Explicit	HE	Gemini-2.0	6.04	61.09	3.44	17.4	0.6	46.84
	HE	gpt-4o-mini	1.6	5.32	1.19	2.63	6.92	3.44
	MHD	Gemini-2.0	4.44	40.9	7.71	4.51	13.3	1.14
	MHD	gpt-4o-mini	3.59	25.42	0.77	3.53	10.45	3.18
Implicit	HE	Gemini-2.0	6.87	266.25	7.56	87.68	3.68	7.04
	HE	gpt-4o-mini	4.87	26.18	10.01	2.08	22.63	0.59
	MHD	Gemini-2.0	13.8	96.2	6.12	15.27	21.11	25.66
	MHD	gpt-4o-mini	2.27	15.44	3.52	5.82	1.26	1.67

Table 2: Chi-square comparison of agreement distributions under role-play prompts. All values are computed against the *plain* prompt distribution, separately for explicit and implicit settings. Column abbreviations denote role pairs: T–S (teacher_student), S–T (student_teacher), B–A (boss_assistant), A–B (assistant_boss), SGP (student_group_project), and JAC (joint_article_coauthor). Cells with chi-square values > 7.815 (significant distribution mismatch) are highlighted in red.

for several roles on HaluEval. On MHD, gpt-4o-mini shows fewer significant role-induced shifts under Implicit compared to Gemini-2.0. This finding suggests that biases induced by role setting are vary across models, and that system design should account for model-specific differences in stability and sensitivity.

Taken together, Table 1 and Table 2 show that downstream agreement behavior is jointly shaped by verification style (explicit vs. implicit) and role setting. The most consequential effect is the increase in hallucination acceptance (*neg agr*) under implicit prompting, which is further influenced by role asymmetry and model choice. These findings motivate the next section, where we move beyond binary judgments to analyze how hallucinated content is preserved, reduced, or expanded during downstream processing.

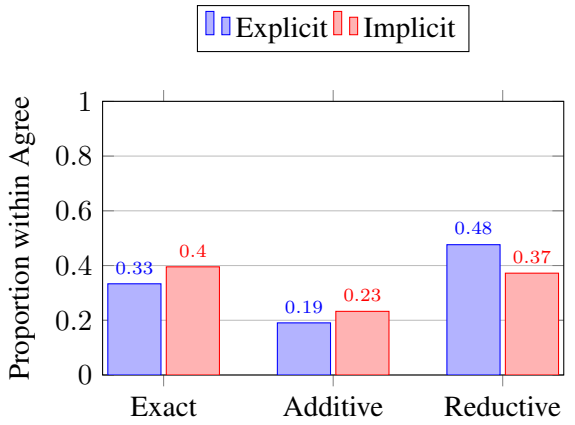
4.2 Information Change Patterns Beyond Agreement

Based on human annotations of 100 hallucination inputs (negative samples) in each setting (explicit and implicit), the overall stance distribution shifts markedly across interaction formats: Agree vs. Disagree is 21% vs. 79% under explicit prompting and 43% vs. 57% under implicit prompting. This indicates that without an explicit verification query, hallucinated upstream content is more likely to be accepted. Beyond stance, we label downstream responses by how they modify upstream information: for Agree, *Exact* preserves the content, *Additive* introduces new information, and *Reductive* removes details while keeping the conclusion; for Disagree, rejections are *Total* or *Partial* as shown in Fig. 3.

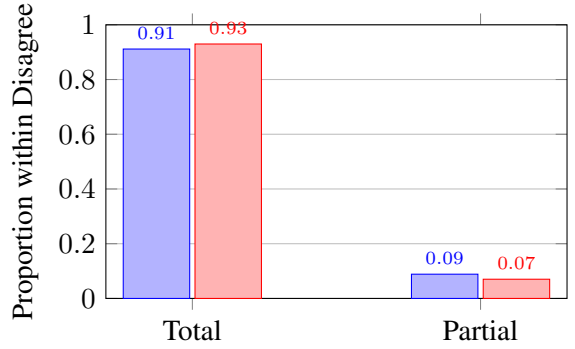
For Disagree cases (Fig. 3b), the composition is stable across settings: Disagree–Total remains above 0.91 under both conditions, and Disagree–Partial decreases slightly. Although the shift is small, Disagree–Partial still poses a potential risk: the agent rejects the upstream conclusion but retains some contextual details. When such retained content is weakly related to the question or vague, it can introduce a misleading context that may interfere with subsequent agents.

The Agree path shows clearer redistribution within agreement types (Fig. 3a). Relative to explicit prompting, implicit prompting increases the share of Agree–Exact (0.33→0.40) and Agree–Additive (0.19→0.23), while reducing the share of Agree–Reductive (0.48→0.37). These changes imply that, once an agent agrees under the implicit format, it is more likely to either preserve the hallucinated content as is or to extend it, rather than expressing agreement through compressive rewriting. From a propagation perspective, these modes correspond to distinct failure mechanisms: Agree–Exact directly forwards hallucinated claims; Agree–Additive risks introducing additional unsupported details on top of an already incorrect conclusion; and Agree–Reductive can make hallucinations harder to detect by distilling them into shorter, more assertive statements.

Taken together, the results show that hallucination risk in multi-agent pipelines depends not only on whether an agent agrees or disagrees, but also on how the content is rewritten after the decision, with implicit interaction increasing both acceptance and downstream transformations that can sustain or amplify errors.



(a) Agreement types (normalized within Agree).



(b) Rejection types (normalized within Disagree).

Figure 3: Human-annotated information-change patterns on hallucination inputs (negative samples) under explicit and implicit settings. Bars are normalized within Agree and within Disagree, respectively. The five classes capture how the downstream response changes information content relative to the upstream answer, including three agreement-conditioned types (exact, additive, reductive) and two rejection types (total, partial).

5 Limitations

Despite the controlled design of our single-hop emulation and the consistent patterns observed across settings, this study has several limitations that constrain the scope of the conclusions. First, the data and hallucination types are restricted. Our experiments use QA-format instances, and hallucinations are instantiated primarily as factual errors in upstream answers. As a result, the findings most directly apply to pipeline scenarios where a downstream agent evaluates or revises a candidate answer to a question. Whether the same effects hold for other task formats, such as summarization, open-ended generation, or multi-turn dialog, and for other forms of model error beyond factual misstatements, remains an open question.

Second, our evaluation is English-only. Role-play cues and verification behavior may interact with language-specific pragmatics. In addition, model performance and reliability can vary substantially across languages. These factors may affect both agreement tendencies and hallucination acceptance rates, limiting direct generalization to multilingual settings.

Third, model and decoding coverage is limited. We evaluate only two representative models (Gemini-2.0 and gpt-4o-mini), which is insufficient to claim model-agnostic effects. Moreover, we do not systematically vary sampling parameters, particularly temperature. Since decoding randomness can influence both consistency and the degree to which models follow role-play cues, additional experiments across a broader range of models and decoding configurations are needed to assess robustness and transferability.

6 Conclusion

Our goal is to verify that, in multi-agent systems and LLM pipelines, hallucinations are not merely localized errors from a single generation step but can develop into systemic risks. Within a single-hop emulation framework, we examine two key interaction factors at the same level, verification style (explicit judgment, which requires an agree/disagree judgment, versus implicit judgment, which does not) and role-play framing and provide converging evidence from three complementary analyses. First, agreement ratios of agents in Table 1 show that both verification style and role-play prompts systematically shift downstream agents' agreement and rejection tendencies, thereby affecting whether hallucinated inputs are passed through. Second, the chi-square analysis in Table 2 quantifies the magnitude of these distributional shifts relative to baseline conditions, and indicates that deviations are generally more pronounced under the pipeline-realistic implicit setting. Third, our information change patterns analysis in Fig. 3 reveals that, after receiving hallucinated inputs, downstream agents do not simply accept or reject them; they also rewrite upstream content in different ways, including preserving it, compressing it, or expanding it, allowing hallucinations to propagate in transformed forms. Overall, our results suggest that hallucinations can both propagate and mutate in multi-agent systems, and that system design and evaluation should treat verification style and role-play-induced interaction biases as central factors.

References

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2025. Voting or consensus? decision-making in multi-agent debate. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11640–11671, Vienna, Austria. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. Camel: communicative agents for "mind" exploration of large language model society. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023b. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: iterative refinement with self-feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering.

In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Jinman Zhao, Zifan Qian, Linbo Cao, Yining Wang, Yitian Ding, Yulan Hu, Zeyu Zhang, and Zeyong Jin. 2025. Role-play paradox in large language models: Reasoning performance gains and ethical dilemmas. *Preprint*, arXiv:2409.13979.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.