Breaking the Speech-Act Contract: Uptake Failures and Void Commitments in Large Language Models

Anonymous ACL submission

Abstract

Large language models do not just answer questions in dialogue; they are also expected to make the right conversational move. When a user apologizes, thanks the assistant, corrects an earlier claim, expresses frustration, or sets a preference, the reply carries a pragmatic obligation. We call failures to meet that obligation *uptake failures*, and model statements that imply unsupported memory or future behavior *void commitments*. We introduce a 300-item benchmark of user-assistant exchanges across six speech-act families, three difficulty tiers, and both synthetic and adapted-from-real-conversation items, with strong annotation agreement ($\kappa_{\text{family}} = 1.00$, $\kappa_{\text{act}} = 0.978$). We also present SpeakFix, a lightweight post-generation repair layer that detects likely contract violations and fixes them with a prepend-not-replace strategy. Across GPT-4o, Claude Sonnet 4.6, and GPT-5.2, baseline uptake accuracy ranges from 71.0% to 77.3%, while void commitment rates range from 8% to 9%. SpeakFix raises uptake accuracy to 87.0–93.3% and reduces void commitments to 0%, with paired sign-test improvements of $p \leq 2.9 \times 10^{-11}$ for every model. The rule-based uptake evaluator reaches 0.869 precision/recall/F1 and $\kappa = 0.593$ on a held-out 90-response validation set. In a five-reviewer manual evaluation of 54 repaired failure cases (270 total judgments), 80.0% of repairs are rated better than the original, 15.9% equivalent, and 4.1% worse; helpfulness and naturalness are preserved in 98.1% and 96.3% of judgments, respectively. Taken together, the results suggest that speech-act contract violations are common enough to measure and often reducible with lightweight repair.

1 Introduction

Modern language models are good at answering questions, writing code, and following explicit instructions. But conversation is not just information transfer. As speech-act theory reminds us (Austin, 1962; Searle, 1969), utterances do things: they apologize, thank, offer, correct, complain, and commit. Those moves create local expectations. An apology invites reassurance; a correction invites acknowledgment; a complaint invites repair.

We test whether current Large Language Models (LLMs) reliably honor those expectations. We focus on two recurring failures. First, the model often makes the wrong next move: it gives advice instead of accepting an apology, explains instead of acknowledging a correction, or moves straight to task content without addressing the user's frustration. We call these *uptake failures*. Second, models sometimes produce commitment-shaped language that overstates what the deployment can actually do, especially around memory and future consistency. We call these *void commitments*. The problem is easy to see in short examples:

> **Apology:** *"Sorry, I pasted the wrong file earlier."*
> **LLM:** *"Sure! To help you with that, I will need you to paste the correct file."* *(uptake failure)*

> **Preference:** *"Please keep your answers under 100 words."*
> **LLM:** *"I'll remember that preference for future sessions."* *(void commitment)*

> **Frustration:** *"You're still not following my instructions!"*
> **LLM:** *"Here is the corrected version."* *(uptake failure)*

These failures matter because they make assistants feel subtly off even when the content is fluent. They are also easy to miss in benchmarks that focus on factual accuracy or generic helpfulness (Ouyang et al., 2022; Bai et al., 2022). To study them directly, we build a 300-item benchmark covering six speech-act families, three difficulty tiers, and

both synthetic and adapted-from-real-conversation prompts. Across three strong chat models, baseline uptake accuracy under our final evaluator ranges from 71.0–77.3%, and void commitments appear in 8–9% of responses.

Contributions. This paper makes three contributions:

1. **Framework and benchmark.** We define *speech-act contracts* – the pairing of a user speech-act family with its expected response act and introduce a 300-item benchmark spanning across six families (apology, thanks, offer, correction, frustration, preference), three difficulty tiers, and both synthetic and adapted-from-real-conversation items. Two annotators achieve $\kappa = 1.00$ on family labels and $\kappa = 0.978$ on expected response-act labels across a 54-item agreement sample.

2. **Mitigation pipeline.** We present a lightweight, model-agnostic pipeline that detects likely uptake failures and void commitments in a draft response and repairs them with a prepend-not-replace strategy that preserves the original response content while restoring the missing conversational move.

3. **Empirical evaluation.** Across three frontier LLMs, GPT-4o, Claude Sonnet 4.6, and GPT-5.2, the pipeline raises uptake accuracy from 71.0–77.3% to 87.0–93.3% and reduces void commitments from 8–9% to 0%. We report evaluator validation, confidence intervals, paired significance tests, and family-/difficulty-level analyses showing where the gains concentrate.

2 Background and Related Work

Speech-act theory. The philosophical foundation for our work lies in classical pragmatics. Austin (1962) established that utterances do things beyond conveying propositional content, introducing the framework of illocutionary acts: statements, questions, promises, apologies, and so on. Searle (1969) systematized this taxonomy and defined the felicity conditions that determine whether a speech act is properly performed. Grice (1975) added the cooperative principle: speakers and listeners tacitly commit to relevance, quality, and manner, and violating these maxims creates implicature. Levinson (1983) synthesized this tradition into a comprehensive account of conversation as a cooperative achievement governed by mutual obligations. Our notion of a *speech-act contract* directly inherits from this tradition: a user's conversational move creates a local felicity condition on the assistant's next turn, and failing that condition is a pragmatic violation regardless of the informativeness of the reply.

Dialogue act classification in natural language processing (NLP). Dialogue act classification has been a long-standing NLP task, with the Switchboard Dialogue Act corpus (Jurafsky et al., 1997) establishing an influential annotation scheme grounded in speech-act theory. Recent work directly probes how well large language models handle this task. Qamar et al. (2025) show that LLMs frequently fail at fundamental sub-tasks underlying dialogue act prediction, including turn management and communicative function identification, often performing no better than simple rule-based baselines despite their general linguistic competence. Maitra et al. (2025) find that LLM-generated responses exhibit systematically different dialogue act distributions from human conversation, reflecting gaps in learned conversational norms. At a broader pragmatic level, Settaluri et al. (2024) evaluate LLM competence across implicature, presupposition, and reference tasks in the Pragmatics Understanding Benchmark (PUB), revealing significant gaps relative to human performance. Our work differs from this tradition in that we do not classify the user's act in isolation; we evaluate whether the model's response performs the required next act given the user's move - a joint, relational property that prior classification benchmarks do not directly measure.

LLM alignment and social competence. More recent LLM work has emphasized instruction following, helpfulness, and human preference alignment (Ouyang et al., 2022; Bai et al., 2022; Wei et al., 2022). Other studies examine prosocial behavior and social reasoning in multi-turn settings (Sap et al., 2022; Kim et al., 2022; Zhou et al., 2024). These lines of work reveal that model responses can be fluent and factually accurate while still failing at the level of appropriate social action. The notion of a void commitment connects to work on long-term conversational memory in LLM agents (Maharana et al., 2024; Bae et al., 2022), but targets a simpler and more auditable failure mode:

language that implies memory or future behavioral guarantees that the evaluated deployment does not actually support.

LLM evaluation and the present work. For response-quality validation, we use LLM-as-judge methods (Zheng et al., 2023; Liu et al., 2023) cautiously and only on failure cases, given known judge biases (Shi et al., 2024). To our knowledge, prior work has not combined uptake failures and unsupported conversational commitments in a single benchmark-and-mitigation framework. Existing alignment benchmarks typically measure task completion, instruction adherence, or safety rather than the local pragmatic move the model makes in response to a specific user act. Our contribution is to isolate and operationalize this narrower question, produce an annotated benchmark with strong inter-annotator agreement, and show that lightweight post-generation repair substantially closes the gap.

3 The Speech-Act Contract Framework

We use the term *speech-act contract* to refer to the local pragmatic obligation created by a user's conversational move. In our setting, a contract has two parts: a triggering user act and a normatively expected next response act. The key question is not whether the assistant's reply is fluent or helpful in a broad sense, but whether it performs the right move at that point in the conversation.

Definition 1 (Speech-Act Contract). *A speech-act contract is a triple $\mathcal{C} = (\mathcal{F}, \mathcal{R}, \Phi)$, where $\mathcal{F} = \{f_1, \ldots, f_k\}$ is a finite set of user speech-act families, $\mathcal{R} = \{r_1, \ldots, r_m\}$ is a finite set of response acts, and $\Phi : \mathcal{F} \to 2^{\mathcal{R}} \setminus \{\emptyset\}$ maps each family to its set of acceptable response acts. A model response $\hat{r}$ satisfies the contract for family f if and only if $\hat{r} \in \Phi(f)$.*

We study six user speech-act families that recur frequently in assistant interaction and that admit relatively clear next-move expectations: **apology**, **thanks**, **offer**, **correction**, **frustration**, and **preference**. Each family is paired with a primary expected response act: *acceptance_reassurance* for apologies, *acknowledgment* for thanks, *accept_or_decline_offer* for offers, *acknowledge_update* for corrections, *repair_acknowledgment* for frustration expressions, and *bounded_acknowledgment* for preference-setting turns. Some benchmark items additionally permit a small set of acceptable alternative acts when the user turn is intentionally mixed or ambiguous.

This setup yields two failure types. **Uptake failure** occurs when the assistant does not perform an acceptable next move for the user's speech act. We measure this at scale using uptake accuracy:

Definition 2 (Uptake Accuracy). *Given a benchmark $B = \{(u_i, \mathcal{C}_i)\}_{i=1}^{N}$ of user turns paired with their contracts, and a model M that produces response $\hat{r}_i = M(u_i)$ for each turn, the uptake accuracy of M on B is*

$$\mathrm{UA}(M, B) \;=\; \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\big[\hat{r}_i \in \Phi(f_i)\big], \quad (1)$$

where f_i is the speech-act family of turn u_i and $\Phi(f_i)$ is its set of acceptable response acts.

Void commitment occurs when the assistant implies a capability that the evaluated deployment does not support, especially around durable memory, persistent preferences, or future behavior guarantees beyond the current interaction.

Definition 3 (Void Commitment). *A response $\hat{r}$ contains a void commitment if it asserts or implies a state that the deployment cannot guarantee; specifically, any claim about durable memory, persistent user preferences, or future behavior beyond the current interaction (e.g., "I'll remember this for future sessions"). A statement explicitly bounded to the current interaction (e.g., "I'll keep this concise in this chat") is not a void commitment.*

The framework is intentionally local. We do not claim that every conversation has one universally correct continuation. We target a narrow, high-impact slice of dialogue behavior that can be annotated reliably and evaluated at scale. Our benchmark therefore should not be read as a full model of dialogue competence; it is a targeted probe of one recurring class of conversational obligations.

4 Benchmark Construction

The benchmark contains 300 user turns with gold speech-act metadata. Each item includes the user turn, a speech-act family label, a primary expected response act, acceptable alternative acts when needed, a difficulty label, and a source-type label. Our goal is not maximal scale but a controlled core benchmark that supports comparative evaluation across families and conditions while remaining small enough for careful item design, agreement analysis, and targeted error inspection.

Benchmark Slice	Count	Percent
Total items	300	100.0
Per family	50 each	16.7 each
Easy	105	35.0
Near-miss	101	33.7
Mixed-act	94	31.3
Synthetic	147	49.0
Adapted from real conversation	153	51.0

Table 1: Composition of the speech-act contract benchmark.

Coverage. Table 1 summarizes benchmark composition. The benchmark is balanced by family, with 50 items each for apology, thanks, offer, correction, frustration, and preference. It spans three difficulty tiers: 105 easy items, 101 near-miss items, and 94 mixed-act items. Easy items contain a single salient move. Near-miss items tempt a plausible but pragmatically wrong response. Mixed-act items layer multiple pressures into one turn. The benchmark design follows three principles. First, balance across families prevents a single frequent act from dominating the overall metric. Second, the difficulty split distinguishes straightforward uptake from cases in which a fluent but pragmatically misaligned answer is easy to produce. Third, the source split separates controlled coverage from messier, more realistic phrasing. The resulting dataset is large enough to expose stable cross-model differences at the family and difficulty level, while still supporting careful curation and agreement analysis.

Source design. We mix 147 synthetic items with 153 adapted-from-real-conversation items. The former give us coverage and balance; the latter introduce messier phrasing and more realistic pragmatic interference. The adapted items are manually rewritten from publicly available interaction patterns rather than copied from model generations, so the benchmark is not constructed from the same outputs later used in evaluation.

Inter-annotator agreement. We assessed agreement on a 54-item sample covering all six families using Cohen's κ (Cohen, 1960; Artstein and Poesio, 2008). Agreement on speech-act family labels is perfect ($\kappa_{\text{family}} = 1.00$), and agreement on expected response-act labels is near-perfect ($\kappa_{\text{act}} = 0.978$; 98.1% accuracy). This does not mean every item has a single philosophically correct continuation, but it does suggest that our operational label space is stable enough for comparative evaluation.

5 Mitigation Pipeline

SpeakFix is intentionally lightweight. It operates on a completed draft response and asks whether that draft violates the local speech-act contract. The goal is minimal intervention: fix the pragmatic failure while keeping as much of the original content as possible.

Step 1: user-act classification. The pipeline first predicts the triggering speech-act family using deterministic lexical cues. This prediction maps directly to the expected response act.

Step 2: response-act and void-commitment detection. The draft response is then classified for its observed response act using a second deterministic matcher. Void commitments are detected separately through a small lexicon of unbounded memory and persistence phrases, including *I'll remember*, *I've made a note*, *for future sessions*, *going forward*, and *from now on*. Statements explicitly bounded to the current interaction (e.g., *in this chat* or *in this conversation*) are not flagged.

Step 3: minimal repair. If the response already satisfies the expected act and contains no void commitment, the pipeline leaves it unchanged. Otherwise, repair proceeds in two stages: strip or bound unsupported commitments, then prepend a short uptake phrase only if the response still misses the expected move. We use this prepend-not-replace strategy because earlier, heavier rewrites improved automatic scores but made responses sound less natural. The same repair rules, prefixes, and commitment-bounding policy are applied to all three evaluated models; we do not tune the system separately per model.

6 Experimental Setup

We evaluate three deployed chat models: GPT-4o (`gpt-4o`), Claude Sonnet 4.6 (`claude-sonnet-4-6`), and GPT-5.2 (`gpt-5.2`). Each benchmark item is presented as a single-turn user message, and the model is asked to respond naturally in one short reply. This single-turn setup matches the local nature of the contract definition: we ask whether the immediate next move is appropriate given the user's turn, not whether a longer multi-turn exchange eventually repairs the interaction.

Automatic benchmark evaluation. We classify the observed response act and mark the response as correct if it falls within the item's set of acceptable acts. We report overall, family-level, difficulty-level, and source-level uptake accuracy, along with raw void-commitment rate.

Mitigation evaluation. We then pass each raw output through SpeakFix and score the repaired response with the same evaluator. Our main automatic outcomes are changes in uptake accuracy and void-commitment rate. Because the same evaluator is used before and after repair, these automatic gains should be interpreted alongside the separate human evaluation rather than in isolation.

Manual evaluation on failure cases. Automatic gains do not tell us whether repaired responses still read well to humans. We therefore manually evaluate only repaired failure cases: responses that originally missed uptake or contained a void commitment and were materially changed by the pipeline. We sample 54 such cases, stratified across models and speech-act families; this is a different 54-item set from the evaluator-validation sample. We then ask five reviewers to compare the original and repaired responses using the same rubric: *better/same/worse*, *helpfulness preserved*, and *naturalness preserved*. We report pooled judgment rates and mean pairwise agreement.

Agreement and classifier checks. Because the automatic metrics rely on rule-based recognition, we also validate the evaluator on a manually annotated 90-item response sample. It achieves 0.869 precision, recall, and F1 for binary uptake correctness, with 0.822 accuracy and Cohen's $\kappa = 0.593$ (Table 3; Figure 1).

7 Results and Analysis

We report three results: baseline contract-violation rates, automatic gains from SpeakFix, and a five-reviewer human evaluation on failure cases.

7.1 Baseline Performance

Under the final evaluation setup, the three models achieve baseline uptake accuracies between 71.0% and 77.3%. GPT-5.2 performs best overall (77.3%), followed by GPT-4o (73.0%) and Claude Sonnet 4.6 (71.0%). Void commitments are more stable across models: 8.0% for GPT-5.2 and 9.0% for both GPT-4o and Claude Sonnet 4.6. This consistency suggests that unsupported commitment language is not an idiosyncratic artifact of a single provider, even though uptake quality varies substantially by model.

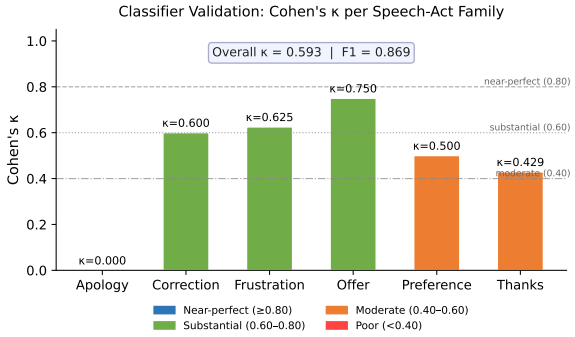

Figure 1: Classifier validation Cohen's κ by speech-act family. The apology family reports $\kappa = 0.0$ because all ten validation items were classified correctly by both the classifier and the human annotator, making observed agreement equal to expected chance agreement in the κ formula, a small-n artifact rather than a classification failure (raw accuracy is 100% for apology).

The family-level breakdown is not uniform. Apology and offer items are handled relatively well by all three models, while preference, thanks, and some correction or frustration items remain more fragile. These errors are not random; they cluster around specific pragmatic obligations. This is one reason we frame the benchmark as a structured stress test rather than a generic style evaluation.

Table 2 shows per-family uptake accuracy before and after mitigation for all three models. Frustration is the weakest family before repair for GPT-4o (48.0%) and a notable gap also exists for Claude Sonnet 4.6 in corrections (50.0%) and preferences (52.0%). GPT-5.2 is more balanced before mitigation but gains less on preferences (74.0% $\rightarrow$ 78.0%). After mitigation, all families reach at least 70%, and correction, offer, and frustration achieve the largest absolute gains.

Near-miss items are the hardest split before mitigation: 63.4% for GPT-4o, 59.4% for Claude Sonnet 4.6, and 70.3% for GPT-5.2. Adapted-from-real-conversation items are also consistently harder than, or at best roughly comparable to, synthetic items.

7.2 Automatic Mitigation Gains

Table 4 summarizes the full results. SpeakFix produces substantial automatic gains on all three models. For GPT-5.2, uptake accuracy rises from 77.3% to 89.3% (+12.0 points) while the void-commitment rate falls from 8.0% to 0%. For GPT-4o, uptake rises from 73.0% to 93.3% (+20.3

	GPT-4o		Claude Sonnet 4.6		GPT-5.2	
Family	Before	After	Before	After	Before	After
Apology	0.940	0.940	0.920	0.940	0.760	0.760
Thanks	0.740	0.900	0.800	0.840	0.580	0.860
Offer	0.720	1.000	0.780	0.980	0.880	1.000
Correction	0.900	0.920	0.500	0.700	0.860	0.960
Frustration	0.480	0.980	0.740	0.880	0.820	1.000
Preference	0.600	0.860	0.520	0.880	0.740	0.780
Overall	**0.730**	**0.933**	**0.710**	**0.870**	**0.773**	**0.893**

Table 2: Per-family uptake accuracy before and after SpeakFix for all three models (50 items per family). Apology is consistently strong, while frustration and preference are the hardest families before repair. After mitigation, every family and model combination reaches at least 70%.

Metric	Value
Precision	0.869
Recall	0.869
F1	0.869
Accuracy	0.822
Cohen's κ	0.593

Table 3: Held-out evaluator validation on 90 manually labeled responses.

points) and void commitments drop from 9.0% to 0%. For Claude Sonnet 4.6, uptake rises from 71.0% to 87.0% (+16.0 points) and void commitments likewise drop from 9.0% to 0%.

The mitigation works best when the missing move is short and local. The largest gains are concentrated in offer, frustration, correction, and preference slices, depending on the model. Preference items remain the hardest family overall, though all three models still gain.

The same pattern appears by difficulty. Near-miss items, the hardest setting before repair, improve substantially for every model: 63.4% to 88.1% for GPT-4o, 59.4% to 82.2% for Claude Sonnet 4.6, and 70.3% to 87.1% for GPT-5.2. Mixed-act items also improve to 95.7%, 89.4%, and 90.4%, respectively. Source-type differences narrow after repair as well: adapted-from-real-conversation items move to 86.3-93.5% uptake, and synthetic items to 87.8-93.2%.

These gains are achieved with relatively sparse intervention. The pipeline changes 86 responses for GPT-5.2, 96 for GPT-4o, and 91 for Claude Sonnet 4.6 out of 300 benchmark items.

Figure 2 visualizes the overall uptake gains. Table 5 summarizes the benchmark slices that are most diagnostic before and after repair.

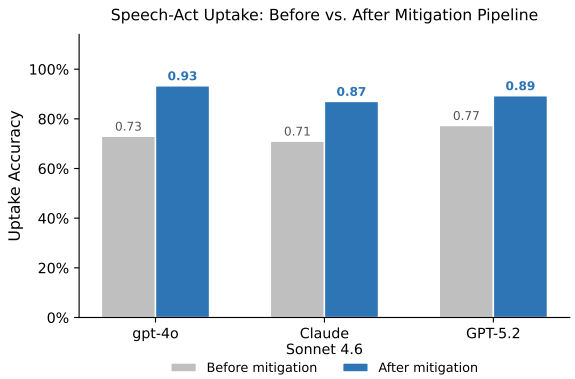

Figure 2: Uptake accuracy before and after the SpeakFix mitigation pipeline for all three models. Grey bars show baseline performance (71.0–77.3%); blue bars show post-repair accuracy (87.0–93.3%). GPT-4o gains the most (+20.3 points), while GPT-5.2 starts highest and gains the least (+12.0 points). All three models reach above 87% after repair, and void-commitment rates drop to 0% in every case.

7.3 Five-Reviewer Human Evaluation on Failure Cases

To validate quality directly, we manually reviewed 54 repaired failure cases with five reviewers using the same rubric (Table 6). Across the pooled 270 judgments, 80.0% of repaired responses were rated *better* than the original, 15.9% were rated *equivalent*, and 4.1% were rated *worse*. Helpfulness was preserved in 98.1% of judgments and naturalness in 96.3%.

Agreement was strongest on the binary quality dimensions. Mean pairwise agreement is 0.963 for helpfulness-preservation judgments and 0.930 for naturalness judgments. On the three-way *better/same/worse* label, mean pairwise agreement is 0.678, with unanimous judgments on 23 of 54 cases. Most disagreements were *better* versus *same* or *better* versus *worse*, which suggests that the reviewers often agreed on the repaired response's basic acceptability while differing about the size of

Model	Before		After		Changed	p
	Uptake (95% CI)	Void rate	Uptake (95% CI)	Void rate		
GPT-4o	0.730 [0.677, 0.777]	0.090	0.933 [0.899, 0.956]	0.000	96	1.39×10^{-17}
Claude Sonnet 4.6	0.710 [0.656, 0.758]	0.090	0.870 [0.827, 0.903]	0.000	91	7.11×10^{-15}
GPT-5.2	0.773 [0.723, 0.817]	0.080	0.893 [0.853, 0.923]	0.000	86	2.91×10^{-11}

Table 4: Main benchmark results before and after SpeakFix. Confidence intervals are Wilson 95% intervals. The final column reports paired exact sign tests on before/after uptake changes after discarding unchanged items.

Slice	GPT-4o	Claude Sonnet 4.6	GPT-5.2
Hardest family before repair	frustration 0.48	correction 0.50	thanks 0.58
Largest family gain	frustration +0.50	preference +0.36	thanks +0.28
Hardest difficulty before repair	near-miss 0.634	near-miss 0.594	near-miss 0.703
Near-miss after repair	0.881	0.822	0.871
Adapted source before / after	0.719 / 0.935	0.686 / 0.863	0.725 / 0.889
Synthetic source before / after	0.741 / 0.932	0.735 / 0.878	0.823 / 0.898

Table 5: Stress-test slices derived from the 300-item core benchmark. Near-miss prompts are the hardest condition for all three models before repair, and adapted-from-real-conversation items are generally harder than synthetic ones before mitigation.

the improvement.

Disagreements cluster around a small set of cases where the repair adds a redundant acknowledgment or a template-like opener. This is useful: it shows that the manual audit is not merely echoing the automatic metric, and it localizes where prepend-not-replace most often trades fluency for safer pragmatic uptake. Because the sample covers only repaired failure cases, we treat this study as targeted quality evidence rather than as full human validation of the benchmark.

8 Discussion

Three takeaways stand out. First, speech-act contract violations are measurable and non-trivial even in frontier chat models. The raw failure rates are not catastrophic, but they are large enough to matter in repeated interaction. Second, a simple post-generation layer goes surprisingly far: it improves both uptake and capability-truthfulness without retraining the base model. Third, not all failures are equally easy to repair. Preference items remain the hardest part of the benchmark because they often require the model to acknowledge the user's request and change tone, style, or format at the same time, even though human reviewers still judged those repairs favorably.

Caveats for interpreting the results. The automatic gains reported here rely on the same rule-based classifier used to produce the benchmark labels, so they overstate real-world impact to the extent that the classifier is imperfect (precision, recall, and F1 all at 0.869 on the held-out sample). The human evaluation provides an independent check, but it covers only 54 repaired failure cases rather than the full benchmark. So it should be read as targeted evidence about repair quality, not as definitive end-to-end human validation of the benchmark.

Implications and next steps. Even with these limitations, the results suggest that dialogue evaluation benefits from looking beyond factuality and generic helpfulness to the local conversational act the model performs. A natural next step is to learn more fluent repair policies, especially for preference and multi-constraint cases, and to extend the benchmark to richer corpora and more languages.

9 Conclusion

We introduced the problem of *speech-act contract* violations in LLM-based dialogue and argued that two failure types deserve explicit measurement: *uptake failures*, where the assistant performs the wrong next conversational move, and *void commitments*, where it implies memory or future behavioral guarantees that the evaluated deployment does not support.

To study these phenomena, we built a 300-item benchmark spanning six speech-act families, three difficulty tiers, and both synthetic and adapted-from-real-conversation prompts. Across three strong chat models, raw uptake accuracy under our final evaluation setup ranges from 71.0% to 77.3%, while void-commitment rates remain at 8-9%. We then presented SpeakFix, a lightweight

Human Review Summary	Value	Notes
Reviewed cases	54	Same sample reviewed by five people
Pooled judgments	270	5 reviewers $\times$ 54 cases
Better / Same / Worse	80.0 / 15.9 / 4.1	Percent of pooled judgments
Helpfulness preserved	98.1	Percent of pooled judgments
Naturalness preserved	96.3	Percent of pooled judgments
Mean pairwise agreement (better/same/worse)	67.8	Across 54 cases
Mean pairwise agreement (helpfulness / naturalness)	96.3 / 93.0	Across 54 cases

Table 6: Five-reviewer human evaluation of repaired failure cases.

post-generation repair layer that raises uptake accuracy to 87.0–93.3% and reduces void commitments to 0% across all three models, with statistically significant gains in every case. The largest improvements appear in the slices that are hardest before repair: near-miss items and families requiring pragmatic acknowledgment rather than generic helpfulness. In a five-reviewer manual evaluation of 54 repaired failure cases, 80.0% of judgments rate the repair as better than the original, with helpfulness and naturalness preserved in 98.1% and 96.3% of judgments.

The broader implication is simple: dialogue alignment is not only about what the model says, but also about what move it makes when it says it. We hope this benchmark and framing help make that part of LLM evaluation easier to measure and easier to improve.

Acknowledgements

The authors used generative AI tools only for writing support and minimal coding assistance during the preparation of this paper. The reported contributions, research ideas, experiments, analyses, and conclusions are the authors' original work. All experimental design, implementation decisions, citations, and final wording were reviewed and verified by the authors, who take full responsibility for the content of the submission.

Limitations

The mitigation pipeline and evaluator rely on keyword-based classifiers trained on English interaction patterns; performance may degrade on low-resource languages, code-switching, or highly idiomatic text. The benchmark covers only six speech-act families selected for their frequency and annotation reliability. It is not an exhaustive taxonomy of conversational obligations. The prepend-not-replace repair strategy prioritizes reliable uptake correction over maximal fluency; more capable rewrite policies may produce more natural outputs. Finally, our evaluation does not include a live user study, so downstream effects on real user trust and satisfaction remain an open question. The current experiments also cover three proprietary deployed models but not an open-weight baseline, so the generality of the observed rates across model families remains to be established.

Ethical Considerations

This work studies conversational patterns in LLM-based assistants. All benchmark items are synthetic or adapted from publicly available interaction logs, and no personally identifiable information is included. The mitigation pipeline is designed to improve conversational alignment rather than introduce new persuasive or deceptive behavior. Five human reviewers completed the manual repair evaluation; no crowd workers or vulnerable participants were recruited, and the annotation task posed no risk of harm, so institutional ethics review was not required. The benchmark and pipeline code will be made publicly available upon acceptance of this work.

References

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

John L. Austin. 1962. *How to Do Things with Words*. Oxford University Press, Oxford, UK.

Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. 2022. Keep me updated! memory management in long-term conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3769–3787.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, and Tom Henighan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics, Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. Technical Report 97-02, University of Colorado, Institute of Cognitive Science.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. ProsocialDialog: A prosocial backbone for conversational agents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4005–4029.

Stephen C. Levinson. 1983. *Pragmatics*. Cambridge University Press, Cambridge, UK.

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.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870.

Arunima Maitra, Dorothea French, and Katharina von der Wense. 2025. Dialogue acts as a lens on human–LLM interaction: Analyzing conversational norms in model-generated responses. In *Proceedings of the Fourth Workshop on Bridging Human–Computer Interaction and Natural Language Processing (HCI+NLP)*, pages 317–325.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, and Alex Ray. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Ayesha Qamar, Jonathan Tong, and Ruihong Huang. 2025. Do LLMs understand dialogues? A case study on dialogue acts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26219–26237.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, UK.

Sravanthi Settaluri, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12110–12125.

Lin Shi, Chieh-Yang Chen, Ting-Yao Liu, and Ting-Hao Huang. 2024. Judging the judges: A systematic study of position bias in LLM-as-a-judge.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, and Eric P Xing. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations*.