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

Sahana Varadaraju **Bharathwaj Vijayakumar** **Lathish Balaji Baskaran**
Rowan University
{varadaraju, vijayakumar, baskaran}@rowan.edu

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. This benchmark is intended as a curated core set: small enough for careful agreement analysis, but structured so that additional balanced batches can be added under the same ontology. On the current core set, we find 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%; the paired uptake gains are statistically significant for all three models ($p < 10^{-10}$). In a two-reviewer manual evaluation of 54 repaired failure cases (108 total judgments), 70.4% of repairs are rated better than the original, 28.7% equivalent, and 0.9% worse; helpfulness and naturalness are preserved in 100% and 99.1% of judgments, respectively. A separate blinded study with four independent, non-author raters is more critical: it confirms the repairs are at least as helpful as the original, but finds that the minimal prepend-style edits are often judged less natural, so reliable uptake gains do not yet translate into uniformly preferred rewrites. 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 frustration expression 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)*

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 3880–3893
November 6-10, 2026 ©2026 Association for Computational Linguistics

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 proprietary chat models, baseline uptake accuracy under our final evaluator is only 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 largely preserves the original response content while restoring the missing conversational move.

3. **Empirical evaluation.** Across three proprietary chat models, 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%, with statistically significant paired gains for all three models and larger improvements than system-prompt or LLM-rewrite baselines. Human evaluation is mixed on surface quality: two author reviewers rate 70.4% of repairs better than the original, whereas a blinded panel of four independent raters finds the repairs at least as helpful but often less natural, so the reliable uptake gains do not yet yield responses that independent readers prefer overall.

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 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, Sravanthi et al. (2024) evaluate LLM competence across implicature, presupposition, and reference tasks in the PUB benchmark, 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 fol-

lowing, 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. Complementary to alignment training, applied deployments increasingly use prompt engineering and post-hoc guardrails to steer model outputs in context (Vijayakumar et al., 2026); SpeakFix is a lightweight, pragmatics-specific instance of such post-generation control. The notion of a void commitment connects to work on long-term conversational memory in LLM agents (Maharana et al., 2024), 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., 2024; Liu et al., 2023) cautiously and only on failure cases, given known judge biases (Shi et al., 2025). 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* $\hat{r} \in \Phi(f)$.

Each family has one *primary* expected response act; individual benchmark items may additionally specify acceptable alternatives, yielding the item-level acceptable set $\Phi(f)$ used for evaluation.

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 choice is deliberate. Our goal is not to impose a maximally fine- grained theory of dialogue, but to isolate a subset of user moves that (i) occur often in assistant interaction, (ii) create practically important expectations about the assistant's next turn, and (iii) can be labeled with high reliability. We therefore prefer a compact ontology with explicit alternative-act allowances over a larger label space that would make annotation less stable and evaluation harder to interpret.

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 current benchmark contains 300 user turns with gold speech-act metadata. We treat this release as a curated core benchmark rather than a maximal collection: it is large enough to support comparative evaluation across families and conditions, yet still small enough for careful item design, agreement analysis, and targeted error inspection. 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.

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. Mixed-act is a difficulty designation, not a multi-label family annotation: each item retains a single primary speech-act family while permitting acceptable alternative response acts where appropriate.

The benchmark design follows three principles. First, balance across families prevents a single high-frequency 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 small enough for careful curation and agreement analysis, but large enough to expose stable cross-model differences at the family and difficulty level.

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.

This design also makes later scaling straightforward. Because family, difficulty, and source-type are explicit fields rather than implicit dataset properties, new batches can be added while preserving the same balancing logic and label ontology. That matters for this setting: a larger benchmark is useful only if it increases source diversity and stress-test coverage without weakening annotation consistency.

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. This rewrite step also reduces the risk that benchmark success depends on memorizing artifacts of a particular public log or provider style.

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% raw agreement). 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: 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 proprietary 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. We use the same benchmark and mitigation policy for all three models and do not perform per-model prompt tuning.

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.

Two-reviewer human evaluation on failure cases. Automatic gains do not tell us whether repairs still sound good to a human reader. We therefore conduct a manual evaluation on failure cases only: responses that originally missed uptake or contained a void commitment and were materially changed by the pipeline. We sample 54 such cases, stratified across the three models and six speech-act families, and ask two reviewers to compare the original and repaired responses using the same rubric: *better/same/worse*, *helpfulness preserved*, and *naturalness preserved*. We report both pooled judgment rates and inter-reviewer agreement. To guard against author bias, we additionally run a fully blinded A/B study with four independent, non-author raters over 60 randomized pairs: 36 genuine repairs and 24 already-adequate responses that the pipeline nonetheless edited (over-trigger controls), with no cue as to which response is the system output.

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 $\kappa = 0.593$ (Figure 1). For the benchmark-level before/after comparisons, we additionally report Wilson 95% confidence intervals for each proportion and exact paired sign tests on discordant items.

The automatic evaluator is therefore best understood as an operational measurement layer aligned to the benchmark ontology, not as a general dialogue understanding system. We use it to support consistent large-scale comparison across models and conditions, and we pair it with targeted human

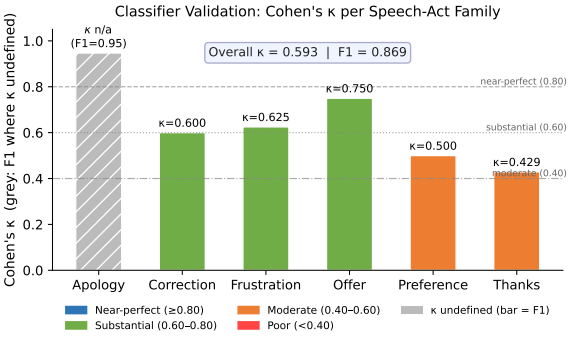


Figure 1: Classifier validation Cohen's κ by speech-act family. For the apology family, Cohen's κ is undefined because the classifier's predictions have no variance (it assigned the same label to all ten items). Classification quality nevertheless remains high ($\mathrm{F1} = 0.95$, one false positive), so we plot F1 (grey, hatched) rather than κ for this family.

review on materially changed failure cases.

7 Results and Analysis

We report four results: baseline contract-violation rates, automatic gains from SpeakFix, a comparison against prompt-based mitigation baselines (including an over-triggering check), and human evaluation.

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.

The family-level breakdown is not uniform. Performance varies substantially by family and model: apology is strong for GPT-4o and Claude Sonnet 4.6, offer is strongest for GPT-5.2, while frustration, preference, thanks, and correction expose different model-specific weaknesses. 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, every family-model combination reaches at least 70%, although the magnitude of improvement varies substantially across families and models.

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

7.2 Automatic Mitigation Gains

Table 3 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 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%.

These changes are large relative to benchmark uncertainty. The uptake rates move from 73.0% [67.7, 77.7] to 93.3% [89.9, 95.6] for GPT-4o, from 71.0% [65.6, 75.8] to 87.0% [82.7, 90.3] for Claude Sonnet 4.6, and from 77.3% [72.3, 81.7] to 89.3% [85.3, 92.3] for GPT-5.2 (Wilson 95% intervals, percentages). On paired discordant items, two-sided exact sign tests also favor the repaired responses for all three models ($p < 10^{-10}$ for uptake and $p \leq 1.2 \times 10^{-7}$ for void-commitment removal).

The mitigation works best when the missing move is short and local. The largest gains are family- and model-specific: frustration improves most for GPT-4o, preference for Claude Sonnet 4.6, and thanks for GPT-5.2; preference nonetheless remains among the harder families after repair, 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: 70.3% to 87.1% for GPT-5.2, 63.4% to 88.1% for GPT-4o, and 59.4% to 82.2% for Claude Sonnet 4.6. Mixed-act items also improve to 90.4%, 95.7%, and 89.4%, respectively. Source-type differences narrow af-

	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). Performance varies substantially by family before repair: frustration is weakest for GPT-4o, correction and preference for Claude Sonnet 4.6, and thanks for GPT-5.2. After mitigation, every family and model combination reaches at least 70%.

Model	**Automatic Benchmark**		**After SpeakFix**		**Changed**	**Void Fixed**
	Uptake	Void rate	Uptake	Void rate		
GPT-4o	0.730	0.090	0.933	0.000	96	27
Claude Sonnet 4.6	0.710	0.090	0.870	0.000	91	27
GPT-5.2	0.773	0.080	0.893	0.000	86	24

Table 3: Main benchmark results before and after SpeakFix. "Changed" counts responses modified by the mitigation layer.

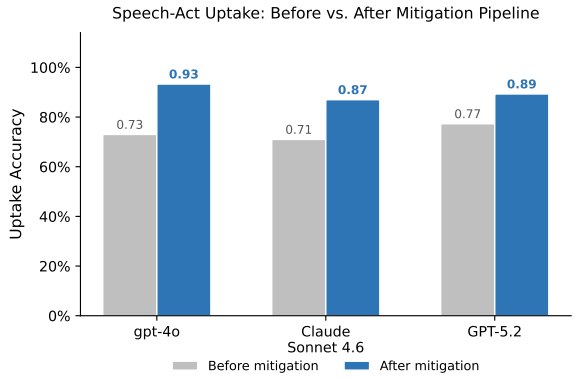


Figure 2: Uptake accuracy before and after the SpeakFix mitigation pipeline for all three models. Grey bars show baseline performance (71.0–77.3%); blue bars show post-repair accuracy (87.0–93.3%). GPT-4o gains the most (+20.3 points), while GPT-5.2 starts highest and gains the least (+12.0 points).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

ter 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.

7.3 Comparison to Baselines

A natural question is whether SpeakFix's structured repair is needed at all, or whether the same gains follow from simply asking the model to behave well. We compare it against two alternatives under identical metrics: a *system-prompt* baseline that instructs the model to acknowledge the user's move and avoid unbounded memory claims (regenerating each response), and an *LLM-rewrite* baseline that asks the same model to fix each draft's pragmatics using no families, lexicons, or rules (Table 4).

The pattern is consistent across all three models. SpeakFix improves uptake by a wide, highly significant margin in every case (exact paired sign tests: $+62/-1$ for GPT-4o, $+48/-0$ for Claude Sonnet 4.6, $+36/-0$ for GPT-5.2; all $p < 10^{-10}$) and drives void commitments to 0%. The prompt-based baselines do not: the system-prompt baseline lowers GPT-4o uptake to 60.0% ($p = 5.6 \times 10^{-4}$), with no statistically significant change for Claude Sonnet 4.6 or GPT-5.2, while the LLM-rewrite baseline lowers Claude Sonnet 4.6 uptake to 59.0% ($p = 7.7 \times 10^{-4}$) and does not significantly change the others; the LLM-rewrite pass also removes some void commitments but reintroduces others. Under our evaluation setup, neither generic system prompting nor unconstrained LLM rewriting matches the uptake improvements produced by the targeted, obligation-specific SpeakFix repair.

	GPT-4o		Claude Sonnet 4.6		GPT-5.2	
Condition	Uptake	Void	Uptake	Void	Uptake	Void
Base	73.0	9.0	71.0	9.0	77.3	8.0
System-prompt	60.0	0.0	68.3	2.3	72.0	0.0
LLM-rewrite	70.3	8.7	59.0	4.0	73.3	4.0
SpeakFix	**93.3**	**0.0**	**87.0**	**0.0**	**89.3**	**0.0**

Table 4: Uptake accuracy and void-commitment rate (percent) for the base model and three mitigation strategies, on the full 300-item benchmark. Among the strategies evaluated here, only SpeakFix reliably improves uptake; the two prompt-based baselines do not, and sometimes reduce it.

Does SpeakFix over-trigger? Because the automatic evaluator recognizes the same uptake moves that SpeakFix's prefixes supply, we must check that the pipeline does not simply rewrite responses that were already correct. We isolate the drafts that were already fine under the evaluator (correct uptake and no void commitment) and pass them through SpeakFix. Of these, SpeakFix modifies only a small fraction and never harms them: the no-op correctness rate is 93.6% for GPT-4o, 93.8% for Claude Sonnet 4.6, and 89.7% for GPT-5.2, and in all three models zero of the touched already-fine responses lost uptake or gained a void commitment (over-trigger rates 6.4%, 6.2%, and 10.3% respectively, with 0 harmful edits). The over-triggered cases are redundant rather than incorrect, typically a short uptake prefix added to a response that already handled the move, which is a naturalness cost rather than a correctness cost, and is precisely the kind of effect the human evaluation is designed to detect.

7.4 Human Evaluation

To validate quality directly, we manually reviewed 54 repaired failure cases, with both reviewers evaluating the same sample (full breakdown in Appendix C). For transparency, the two reviewers in this study are authors of this work; the automatic over-triggering check in Section 7.3 provides a complementary benchmark-internal measure of whether the pipeline degrades responses that were already acceptable under the evaluator. Across the pooled 108 judgments, 70.4% of repaired responses were rated *better* than the original, 28.7% were rated *equivalent*, and only 0.9% were rated *worse*. Helpfulness was preserved in 100% of judgments and naturalness in 99.1%.

Agreement was strongest on the binary quality dimensions: the two reviewers agreed on 100% of helpfulness-preservation judgments and 98.1% of naturalness judgments. Agreement on the three-way *better/same/worse* label was 68.5%. Most disagreements were *better* versus *same*, not *better* versus *worse*, which suggests that the reviewers largely agreed on acceptability and disagreed mainly about how much improvement the repair provided.

The family-level pattern is also intuitive. Correction and preference repairs were rated better in 100% of pooled judgments, and apology repairs in 91.7%. Frustration repairs were stronger than not, with 75.0% *better*, 20.0% *same*, and 5.0% *worse*. Offer and thanks repairs were more often judged *same* than *better*, suggesting that in these families the final system often makes a response safer or more explicit without dramatically changing the perceived quality.

Blinded evaluation with an independent panel. Because authors evaluating their own system may be biased, we ran a second, fully blinded study with four raters who are neither authors nor collaborators on this work. Each rater saw the same 60 randomly ordered response pairs: 36 genuine repairs (a base failure vs. its SpeakFix repair) and 24 *over-trigger controls* (an already-adequate base response vs. the response SpeakFix nonetheless edited), with no indication of which side was the system output. Raters chose which response was better overall, more helpful, and more natural. Table 5 reports the pooled results; all four raters selected the same overall preference on 55 of 60 pairs (91.7%).

The independent panel is markedly more conservative than the author reviewers. On genuine repairs, SpeakFix is judged at least as helpful as the original in 86.9% of judgments (43.1% better, 43.8% tie), but on overall preference and naturalness the raters favor the original 59.7% of the time, and on the over-trigger controls they pick the untouched base response 96.9% of the time. The automatic evaluation indicates that SpeakFix reli-

Stratum	Dim.	SpeakFix	Tie	Base
Repairs (n=144)	Overall	40.3	0.0	**59.7**
	Helpful	**43.1**	43.8	13.2
	Natural	40.3	0.0	**59.7**
Controls (n=96)	Overall	3.1	0.0	**96.9**
	Helpful	3.1	**71.9**	25.0
	Natural	3.1	0.0	**96.9**

Table 5: Blinded A/B evaluation by four independent (non-author) raters, pooled judgments (percent). SpeakFix is at least as helpful as the base response on genuine repairs, but its prepend-style edits are judged less natural, so raters prefer the original overall, especially on already-adequate responses (over-trigger controls) that were edited unnecessarily.

ably adds the operationalized missing move, while the independent panel finds the resulting responses generally at least as helpful but often less natural; editing an already-adequate response is usually a step backwards. The automatic evaluation shows substantial gains in operationalized uptake, but converting them into responses independent readers prefer would need a more fluent, less intrusive repair policy than the minimal prepend.

8 Discussion

Three takeaways stand out. First, speech-act contract violations are measurable and non-trivial even in proprietary 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 produces substantial gains on the operationalized uptake metric: it improves uptake and eliminates detected void commitments without retraining the base model, and it does so where prompt-based alternatives do not. As Section 7.3 shows, neither instructing the model to be pragmatically careful nor asking it to rewrite its own draft reliably improves uptake, and both sometimes make it worse; the obligation-specific structure of SpeakFix is important to the observed gains. A blinded panel of independent raters, however, shows the limit of the current minimal-repair style: while they find the repairs at least as helpful as the original, they often prefer the original wording overall on naturalness grounds, especially when SpeakFix edits a response that was already adequate. The uptake gains thus show the contract is satisfied, not that the repaired surface form is preferred; closing that gap needs a more fluent repair policy. 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.

Internal validity and scope. The automatic results should be interpreted as benchmark-internal evidence: the same ontology informs benchmark labels, rule-based recognition, and repair triggers, which suits controlled measurement but does not by itself establish broad real-world effect size. The held-out validation sample partly addresses this, showing that the classifier tracks human uptake judgments with 0.869 precision/recall/F1, but it does not remove the dependence on the benchmark's operational label space. Both human studies, moreover, cover materially changed or control cases rather than the entire 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 along three axes at once: more items, more source diversity, and broader human validation. In practice, that means treating the present 300-item benchmark as a curated core set and building larger balanced extensions, stronger external annotation studies, and open-weight model comparisons on top of the same evaluation protocol.

9 Conclusion

We introduced speech-act contract violations as a measurable failure mode in LLM dialogue, focusing on uptake failures and void commitments. Using a 300-item benchmark across six speech-act families, we showed that these failures occur across three proprietary chat models and that SpeakFix, a lightweight post-generation repair layer, substantially improves operationalized uptake while eliminating detected void commitments. Independent human evaluation also reveals an important trade-off: repairs are generally at least as helpful as the originals but can be less natural, highlighting the need for more fluent repair policies. More broadly, the results suggest that dialogue alignment should evaluate not only what a model says, but whether it makes the appropriate conversational move.

Limitations

The mitigation pipeline and evaluator rely on lexical rule-based classifiers designed for 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; as our blinded evaluation shows, independent raters often find the prepended form less natural than the original, and 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 chat models but not an open-weight baseline, so the generality of the observed rates across model families remains to be established. More broadly, the notion of a void commitment is deployment-relative: a statement that is misleading in a stateless chat setting may be accurate in a system with durable memory or explicit preference storage. Our labels should therefore be interpreted with respect to the capabilities of the evaluated deployment rather than as universal judgments about the surface form alone.

Ethics Statement

This work studies conversational patterns in LLM-based assistants. The benchmark contains no private user conversations or personally identifiable information. Synthetic items are hand-authored; the adapted items are manually rewritten from publicly observable chat logs and conversational patterns rather than copied verbatim, and are intended for evaluating conversational behavior rather than inferring anything about individual users. The source patterns were drawn from publicly accessible online discussions and publicly shared chatbot interactions; no private or authenticated conversations were collected. The mitigation pipeline is designed to improve conversational alignment rather than introduce new persuasive or deceptive behavior. The two reviewers in the failure-case evaluation are co-authors of this work. The blinded A/B evaluation additionally involved four independent raters who are not authors or collaborators (two graduate students, one PhD student, and one professional colleague); they participated voluntarily and without compensation, gave informed consent, and judged only short assistant responses containing no personally identifiable information.

Code and data availability. The benchmark, SpeakFix implementation, evaluation and baseline scripts, prompts, and human-evaluation materials are available at `https://github.com/sahanakv01123/speech-act-contracts`.

Acknowledgments

The authors thank Jacqueline M. Ring, Vice Chancellor and Chief Institutional Research Officer at Rowan University, for her support and guidance throughout this work.

Generative AI tools were used in a limited supporting role for coding assistance and language refinement during manuscript preparation. The research design, methodology, experiments, analysis, interpretation of results, and substantive manuscript development were conducted by the authors. All AI-assisted outputs were independently reviewed, revised, and verified by the authors, who take full responsibility for the final content.

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.

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, Jack Hessel, Liwei Jiang, Peter Clark, 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)*, page 317.

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.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314.

Settaluri Lakshmi Sravanthi, 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 12075–12097.

Bharathwaj Vijayakumar, Sahana K. Varadaraju, Lathish Balaji Baskaran, Nidhal C. Bouaynaya, and S. Alapati. 2026. AI-powered course discovery: Leveraging prompt engineering, feedback loops, and guardrails for institutional context. In *2026 IEEE Global Engineering Education Conference (EDUCON)*, pages 1–5. IEEE.

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. 2024. 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*.

A Complete SpeakFix Specification

SpeakFix is a deterministic, rule-based system, so the rules are the method. For reproducibility we document the complete lexical resources used by each stage: the user-act classifier (Appendix A.1), the response-act matcher (Appendix A.2), the uptake-prefix and over-trigger guard lexicons (Appendix A.3), and the void-commitment lexicon and bounding map (Appendix A.4). Appendix A.5 gives a before/after repair example for each of the six speech-act families. All text is lowercased and quote/dash-normalized before matching.

Rule precedence. When multiple rules could fire, precedence is resolved deterministically: the user-act classifier returns the first matching family in the listed order (Table 7); the response-act matcher applies its checks in the order listed in Appendix A.2, with the more specific families (frustration, correction, preference, offer) checked before the generic acceptance/acknowledgment families; and repair prepends at most one uptake phrase, skipped entirely when the draft already opens with a recognized uptake cue for its family (Table 9). Because user-act precedence is by first match, a turn that blends acts (e.g., an apology that also corrects) is assigned to its earliest matching family; this is a deliberate deterministic policy rather than a mixed-act disambiguation model.

A.1 User-Act Classifier

The user turn is mapped to a speech-act family by the first matching rule in the order below (first match wins); if none fire the family is *other*. Each family maps deterministically to a single expected response act (Table 6).

A.2 Response-Act Matcher

The response-act matcher is deterministic and applies the pattern groups below in the order listed. The first matching rule is returned; if no rule matches, the response is labeled *other*. All tests are

User act	Expected response act
apology	acceptance_reassurance
thanks	acknowledgment
offer	accept_or_decline_offer
correction	acknowledge_update
frustration	repair_acknowledgment
preference	bounded_acknowledgment
other	other

Table 6: Expected response act for each user-act family.

Family	Trigger cues (substring match, in order)
apology	*sorry*
thanks	*thank*
offer	*i can*, *i could*
frustration	*frustrat*, *ignoring*, *annoying*, *still not following*
preference	*please keep*, *be concise*, *keep your answer brief*, *use plain language*, *give me the answer first*, *avoid tables*, *please answer*, *plain prose*, *one paragraph*, *shorter sentences*, *under 120/100 words*, *don't switch back*, *no extra caveats*, *answer directly*, *don't use bold*, *don't use bullet(s)*, *no bullet*, *short and to the point*
correction	*actually*, *", not "*, *" no, "*
other	(default; no cue matched)

Table 7: User-act classifier cues. Rules are evaluated top to bottom and the first match determines the family.

case-insensitive substring matches on the normalized response (Appendix A). These are the complete pattern sets used in all experiments; no model-specific patterns or thresholds are used.

0. acceptance_reassurance (early exit): the normalized response starts with *no problem*.

1. repair_acknowledgment: *you're right to be frustrated*, *i understand the frustration*, *understand your frustration*, *sorry about that*, *i'm sorry this has been frustrating*, *i'm sorry you're dealing with that*, *i'm sorry you're feeling*, *i didn't follow your constraints*, *that's on me*, *i'll keep the format consistent*, *i won't use bullet points*, *i'll keep my next reply within it*, *i'll rewrite my response to match it exactly*, *you're right to flag that*, *i'll stay within the constraints you gave*, *i'll stick to it exactly in my next reply*, *i'm sorry-i*, *i missed that*, *i'll avoid that*, *i'll avoid doing*, *i apologize for that frustration*, *i apologize for the frustration*, *i'm sorry for the confusion*, *you're right, i'm sorry*, *you're right, and i'm sorry*, *you're right, and i apologize*, *you're right, i apologize*, *you're right to call that out*.

2. acknowledge_update: *thanks for the correction*, *thank you for the correction*, *thanks for catching that*, *thank you for catching*, *you're right*, *you're absolutely right*, *thanks, that's correct*, *thanks for clarifying*, *thank you for clarifying*, *got it-thanks for clarifying*, *got it- thanks for clarifying*, *i'll treat the appendix as optional*, *i'll switch to the camera-ready citation*, *i'll attribute that quote to reviewer*, *got it-i'll*, *got it, i'll*, *ah, got it*, *good catch*, *i'll move that*.

3. bounded_acknowledgment: *in this conversation*, *in this chat*, *i'll keep my answers concise*, *i will use plain language and short sentences*, *i'll use plain language*, *i'll put the direct answer first*, *i'll avoid using tables unless*, *i'll keep my replies to a maximum of two sentences*, *i won't use bold text unless*, *i'll answer in plain prose*, *i'll keep my next reply under*, *i'll keep responses direct*, *i'll answer directly*, *i'll respond in plain prose*, *i'll keep the same tone in my next reply*, *i'll keep the same structure as before for the rest of this conversation*, *i'll keep my replies short and to the point*, *i'll use shorter sentences*, *i won't add extra caveats unless you ask*, *i'll avoid using bold unless you ask*, *i'll keep using paragraph form*, *i'll keep future*, *i'll stick with*, *keep it concise*, *i'll keep it short*, a response starting with *understood*, *i'll keep my reply*, *i'll keep my responses*, *i'll keep things*, *i'll make sure to keep*.

4. accept_or_decline_offer: fired if any *strong* cue is present, or any *weak* cue is present and no apology-accept cue (*no worries*, *that's okay*, *no need to apologize*, *my apologies*) is present. *Strong*: *yes, please*, *please share*, *please paste*, *please send*, *go ahead and paste*, *go ahead and upload*, *no thanks*, *yes, that would help*, *that would be helpful*, *that would be very helpful*, *if it's not too much trouble*. *Weak*: *feel free to send*, *feel free to share*, *feel free to upload*, *feel free to paste*, *please go ahead and*.

5. acceptance_reassurance: *no worries*, *that's okay*, *no need to apologize*, *my apologies*.

6. acknowledgment: *you're welcome*, *glad it helped*, *glad it's clearer*, *glad that helped*, *happy to help*, *glad it landed*, *glad to hear it*, *glad it's working better*, *thanks for the note*, *glad it hit*, *will do*, *got it*, *sure thing*.

7. advice: *try* (followed by a space) or *next time*.

8. other: returned if no rule above matches.

The pattern sets above are identical to those used to generate all automatic evaluation results in Sections 6–7; no patterns were added or modified after evaluation.

A.3 Uptake Prefixes and Over-Trigger Guard

When repair is required, SpeakFix prepends the short uptake phrase in Table 8 for the detected user act. To avoid over-triggering on responses that already perform the correct move, the prefix is *not* added when the response already opens with a recognized uptake cue for that family (Table 9); in that case the draft is returned unchanged.

User act	Prepended uptake phrase
apology	*No worries at all.*
thanks	*You're welcome. Happy to help.*
offer	*Yes, please share it.*
correction	*You're right, thanks for the correction.*
preference	*Understood. I'll keep that in mind in this conversation.*
frustration	*You're right to flag that.*

Table 8: Uptake prefixes prepended during minimal repair.

Family	Already-satisfied opening cues (repair skipped)
apology	*no worries*, *no problem*, *that's okay*, *no need to apologize*, *my apologies*
thanks	*you're welcome*, *happy to help*, *glad it helped/was helpful/to help*, *glad to hear it*, *glad it's clearer/working better*
offer	*yes, please*, *please share/paste/send*, *go ahead and paste/upload*, *that would be (very) helpful*, *feel free to send/share/upload/paste*
correction	*you're (absolutely) right*, *thanks for clarifying*, *thanks for catching that*, *got it*, *good catch*
preference	*understood*, *got it*, *i'll keep*, *i'll answer directly*, *i'll use plain language*, *in this conversation*, *in this chat*
frustration	*you're right to flag that*, *you're right to be frustrated*, *sorry about that*, *i'm sorry*, *i apologize*, *that's on me*, *i missed that*

Table 9: Over-trigger guard. If a draft already opens with any of these cues for its family, no prefix is prepended.

A.4 Void-Commitment Lexicon and Bounding Map

A response is flagged as a void commitment if it contains any phrase in Table 10 *and* is not explicitly bounded to the current interaction (the presence of *in this conversation*, *in this chat*, or *in this thread* suppresses the flag). During repair, each detected phrase is replaced in place by its bounded equivalent (Table 10, right column), preserving the surrounding content.

Void-commitment phrase	Bounded replacement
i'll remember / i will remember	i'll keep in mind for this conversation
i've made a note / i have made a note	i'll keep that in mind
i'll / i will save that preference	i'll apply that preference
save that preference permanently	apply that preference
for future sessions	in this conversation
going forward	in this conversation
from now on	in this conversation

Table 10: Void-commitment detection lexicon and in-place bounding map. Phrases explicitly scoped to the current interaction are never flagged.

A.5 Before/After Repair Examples

Table 11 shows one repair per family, drawn from the benchmark (GPT-4o drafts). Prepended uptake text and bounded rewrites are shown in *italics*.

B Reproducibility Details

All models are queried single-turn with temperature 0.2 and a 180-token completion cap. We draw a single completion per item (no self-consistency, sampling, or reranking), leave top-p and all other decoding parameters at each provider's default, and set no fixed random seed. The base generation system prompt is: *"You are a helpful research assistant. Respond naturally to the user's message in one short reply."* Table 12 lists the exact Azure deployment identifiers and API versions for each model. All model outputs used in the paper were collected in September 2026 using the model versions served by these deployments during that window; where a provider does not expose a finer-grained snapshot identifier, we report the deployment identifier and access window rather than an inferred date.

For the two prompt-based baselines (Section 7.3), the *system-prompt* baseline replaces the base system prompt with: *"You are a helpful research assistant. Respond naturally to the user's message in one short reply. Before answering, make sure your reply explicitly acknowledges the user's conversational move: accept and reassure after an apology, acknowledge thanks, accept or politely decline an offer, acknowledge and incorporate a correction, address frustration with a brief repair, and confirm a stated preference. Never claim to remember information across sessions, to save preferences permanently, or to change your behavior 'going forward' or 'from now on'. If you*

Family	Before	After
apology	Thanks for catching that, I'll update the label to 2c instead of 2b.	*No worries at all.* Thanks for catching that, I'll update the label to 2c instead of 2b.
thanks	Sure, go ahead and send the next reviewer comment.	*You're welcome. Happy to help.* Sure, go ahead and send the next reviewer comment.
offer	Sounds good, go ahead and send the draft, and I'll give concise, focused feedback.	*Yes, please share it.* Sounds good, go ahead and send the draft, and I'll give concise, focused feedback.
correction	Got it. I'll use the revised introduction and keep the explanation brief going forward.	Got it. I'll use the revised introduction and keep the explanation brief *in this conversation*.
frustration	Thanks for telling me. I'll keep my replies shorter and more to the point from here on out …	*You're right to flag that.* Thanks for telling me. I'll keep my replies shorter and more to the point …
preference	Sure, what would you like me to help you with?	*Understood. I'll keep that in mind in this conversation.* Sure, what would you like me to help you with?

Table 11: Before/after repairs for each speech-act family, drawn from the benchmark. Italic text marks SpeakFix's prepended uptake move or bounded rewrite; the correction row shows a void commitment (*going forward*) being bounded to the current conversation.

Model	Deployment	API version
GPT-4o	`gpt-4o`	`2024-02-01`
Claude Sonnet 4.6	`claude-sonnet-4-6`	`2024-02-01`
GPT-5.2	`gpt-5.2`	`2024-12-01-preview`

Table 12: Azure deployment identifiers and API versions. Full endpoint URLs and keys are provided through environment configuration and are not part of the release.

Human Review Summary	Value	Notes
Reviewed cases	54	Same sample, two reviewers
Pooled judgments	108	2 reviewers × 54
Better / Same / Worse	70.4 / 28.7 / 0.9	% of judgments
Helpfulness preserved	100.0	% of judgments
Naturalness preserved	99.1	% of judgments
Agreement (B/S/W)	68.5	Raw, 54 cases
Agreement (help./nat.)	100.0 / 98.1	Raw, 54 cases

Table 13: Two-reviewer (author) human evaluation of repaired failure cases.

commit to anything, bound it explicitly to the current conversation." The *LLM-rewrite* baseline uses the system prompt: *"You are an editor that improves the pragmatics of a chat assistant's reply. Given the user's message and a draft reply, rewrite the reply so it clearly responds to what the user is doing … while preserving all helpful content. Do not claim to remember anything across sessions or to store preferences permanently. Keep it to one short reply. Return only the rewritten reply, with no preamble or explanation."*

The benchmark, SpeakFix implementation, all generation and evaluation scripts, the baseline generators, the over-triggering check, and the blinded human-study exporter and scorer are released with the paper.

C Additional Human-Evaluation Details

Table 13 gives the full breakdown of the two-reviewer (author) evaluation of 54 repaired failure cases summarized in Section 7.

C.1 Blinded Independent A/B Study

Raters. The blinded study used four raters who are not authors or collaborators on this work: two graduate students, one PhD student, and one professional colleague. All participated voluntarily and without compensation and gave informed consent. They rated only short assistant responses, which contain no personally identifiable information.

Design. Every rater saw the same 60 randomly ordered response pairs: 36 genuine repairs (a base failure vs. its SpeakFix repair) and 24 over-trigger controls (an already-adequate base response vs. the response SpeakFix nonetheless edited). For each pair, the two responses were shown as *A* and *B* with the side assignment randomized independently per item, and no indication of which side was the system output. Raters chose *A*, *B*, or *Tie* independently for each of three dimensions (overall preference, helpfulness, and naturalness) and could optionally flag memory claims. Scoring maps each A/B choice back to base-vs-SpeakFix using a private key withheld from raters. We report pooled judgments across all four raters (Table 5) and raw inter-rater agreement on the overall judgment (all four raters agreed on 55 of 60 pairs, 91.7%). The exact rating sheet and instructions are released with the code.