

Hear Me Out Again: Grounding and Repair After Mid-Conversation Voice Conversion in Full-Duplex AI

Felix Ölander

KTH Royal Institute of Technology, Stockholm University
Stockholm, Sweden
felixola@kth.se

Éva Székely

KTH Royal Institute of Technology
Stockholm, Sweden
szekely@kth.se

Syed Mohammad Fahim Abrar

KTH Royal Institute of Technology, Stockholm University
Stockholm, Sweden
smfabrar@kth.se

Shree Harsha Bokkahalli Satish

KTH Royal Institute of Technology
Stockholm, Sweden
shbs@kth.se

Abstract

Full-duplex conversational models jointly process linguistic and vocal cues while managing turn-taking and context across an ongoing conversation. Yet, controlled evaluations of voice-related differences have largely used single-turn, stable-voice inputs and isolated responses. In ongoing interactions, users may change their vocal presentation while their task or communicative goal remains intact. For a full-duplex model, these cues may shape when it speaks or yields, and how incoming speech is processed. This raises a question stable-voice evaluations cannot address: *How does changing the voice presented to the model mid-conversation affect the grounding of task-relevant corrections and the repair work required of the user?*

We investigate this question by activating and deactivating zero-shot voice conversion (VC) to alter the acoustic presentation of the user’s voice during PersonaPlex conversations without resetting the session. In a counterbalanced within-participant user study, 24 participants complete four goal-oriented record-correction scenarios under four voice conditions: stable natural, stable converted, natural-to-converted, and converted-to-natural speech.

We traced two predefined task-critical information units per scenario through acknowledgement, claimed update, operational incorporation, delayed retention, and the final account. Of 154 units with checkable transmission, 148 reached the model completely, yet only 6 of 71 classifiable interactions met the strict end-to-end uptake criterion. False update confirmation occurred in 45 of 64 observable interactions, and only 6 of 59 final accounts were fully accurate. In the amended, hypothesis-guided pooled analysis, neither sequence contrast showed a detectable difference in information-use indicator satisfaction. The deactivation sequence had a higher expected post-checkpoint repair count than stable natural input (ratio of expected repair counts 1.64, 95% CI [1.11, 2.41]), but neither H1 sequence hypothesis met its joint decision rule after Holm adjustment. Participants often rated the assistant as accurate even when human review found downstream failures. Full-duplex evaluation should follow corrections beyond fluent acknowledgement to their downstream task consequences.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; *Natural language interfaces*; • **Computing methodologies** → *Speech recognition*.

Keywords

full-duplex, grounding, repair, voice conversion, conversational AI, human-AI interaction

1 Introduction

Full-duplex conversational systems listen while speaking, letting users interrupt or redirect a response. Commercial systems include ChatGPT Voice and Gemini Live [14, 26]. Research systems include Moshi [13] and PersonaPlex [28]. This fluidity makes turn boundaries less explicit and leaves users fewer natural points at which to verify a correction.

Acoustic presentation can affect speech systems. Recognition performance varies across demographic groups [16, 30], and some SpeechLLMs respond differently across accent and perceived-gender presentations of the same words [4, 9, 21, 34]. Voice conversion (VC) alters that presentation while aiming to preserve linguistic content [33]. Low-latency methods can convert live speech [24, 39]. Yet isolated-utterance evaluations cannot show whether information remains effective across later turns [15, 17].

This limitation matters when a user corrects a record. An assistant may acknowledge the correction and claim to have applied it, yet later act on the original information or repeat it in the final account. Conversational models mishandle corrections and lose track of information across turns [8, 18]. We therefore examine subsequent action and retention rather than treating immediate acknowledgement as grounding.

We switched X-VC after 45 seconds of cumulative participant audio without resetting the PersonaPlex session. We paired each switching condition with a stable condition sharing its post-checkpoint voice state. The contrasts therefore compare voice-state sequences rather than isolate the switch. Participants did not hear the converted signal live. Twenty-four participants contributed 88 included record-correction interactions. Of 154 units with checkable transmission, 148 reached the assistant completely, but later use was uncommon. In the amended, hypothesis-guided pooled analysis, we found no detectable information-use difference for either sequence contrast. The deactivation sequence had a higher expected post-checkpoint repair count than stable natural input, but H1b was not supported after adjustment. Acknowledgement and update

claims also often diverged from later action and final-account accuracy, and participant reports did not consistently identify those discrepancies.

We make three contributions:

- (1) We show that the assistant often acknowledged corrections without incorporating or retaining them, and that its update claims often preceded contradictory actions or accounts.
- (2) We run the first analysis on the effect of VC on full-duplex conversational models in terms of repair burden and information use.
- (3) We contribute a within-participant VC protocol and an evaluation framework that follows corrections from delivery through later action and retention.

2 Related Work

2.1 Voice-Conditioned Behaviour and Streaming Voice Conversion

Automatic speech recognition exhibits disparities across demographic and dialect groups [16, 30]. Controlled SpeechLLM studies report response differences across accent and perceived-gender presentations of fixed linguistic content [4, 9, 21, 34]. These studies treat acoustic presentation as an input condition without isolating one causal characteristic.

Voice conversion alters vocal characteristics while aiming to preserve content [33]. Low-latency systems such as X-VC can operate on a live channel [24, 39]. Converted and natural speech may still differ in intelligibility, latency, naturalness, target similarity, and artifacts. Converted speech can also shape trust and attribution [5], so we analyse participant perception separately from assistant behaviour. Fixed-voice and isolated-response studies, however, do not test a change within one continuing conversation.

2.2 Grounding, Repair, and Task-State Persistence

Repair addresses trouble in speaking, hearing, understanding, or action [29]. Voice assistants often leave users to repeat, rephrase, or clarify [2, 23]. Because repair concerns what the system heard, understood, and said, supporting it requires more than semantic correction [22]. Interface barriers discourage feedback [32]. Users prefer actionable repair [1], and self-correction improves evaluations [12]. Earlier interaction patterns can also signal abandonment [38].

Grounding establishes enough evidence of mutual understanding for the current purpose [6, 11]. Conversational models use fewer grounding moves than humans [31], struggle to repair revealed misunderstandings [8], and may persist with early assumptions [18]. Dialogue-state tracking addresses a related requirement: representing the user’s evolving goal across turns [35]. Errors in that representation can affect later responses [36].

We predefined each scenario’s task-critical information and distinguished registration from later use and retention (Section 5.1). This distinction lets us identify apparent acceptance followed by an incorrect action or account.

2.3 Evaluating Full-Duplex Interaction

Full-duplex benchmarks measure latency, overlap, and interruption handling [20, 25]. These measures capture conversational access, but they do not show whether the assistant later uses what a participant said. Newer benchmarks add correction, entity tracking, and multi-turn instruction following [19], but they still use automated examiners with a stable input voice.

Hear Me Out[3] provides an interactive platform for evaluating full-duplex conversational models. The present study uses this platform as experimental infrastructure while focusing on within-session changes in acoustic presentation and their relation to downstream information use and repair.

To our knowledge, prior work has not tested how a within-session change in acoustic presentation relates to later grounding. We address this gap by comparing voice-state sequences ending in the same state and following predefined information from delivery through acknowledgement to later task consequences.

3 Research Questions

We compare VC activation with stable converted input and VC deactivation with stable natural input. Each pair shares its post-checkpoint voice state but differs in its earlier sequence. The contrasts therefore compare voice-state sequences rather than isolate the switch.

RQ1. Does a mid-conversation VC switching sequence differ from its matched stable sequence in the assistant’s use of task-critical information or the participant’s post-checkpoint repair burden?

RQ2. What do participants report about the interaction and about their converted speech when heard retrospectively, and how do those reports relate to annotated information use, task outcome, false update confirmation, and repair?

For RQ1, we examine how the assistant used task-critical information and how often participants repaired the conversation (Section 5). After observing that the planned interaction-level uptake outcome was sparse, we amended the hypothesis-guided analysis before model fitting to pool unit-level indicators. We retain the original outcome as a descriptive measure (Section 5.4).

For RQ2, we use post-scenario ratings, post-session reports, and retrospective playback ratings. Because participants heard converted playback only after the interactions, those ratings do not measure awareness of a live change. We report experience by condition descriptively without condition contrasts.

H1a and H1b compare each switching sequence with its matched stable sequence. H2a and H2b test whether participants’ reports track the outcomes identified through human review:

H1a. VC activation will show greater post-checkpoint repair burden and lower information-use indicator satisfaction than stable converted input.

H1b. VC deactivation will show greater post-checkpoint repair burden and lower information-use indicator satisfaction than stable natural input.

H2a. Greater information-use indicator satisfaction will be associated with higher task-specific trust and greater confidence in the resulting task state.

H2b. Greater repair burden will be associated with greater perceived effort and frustration.

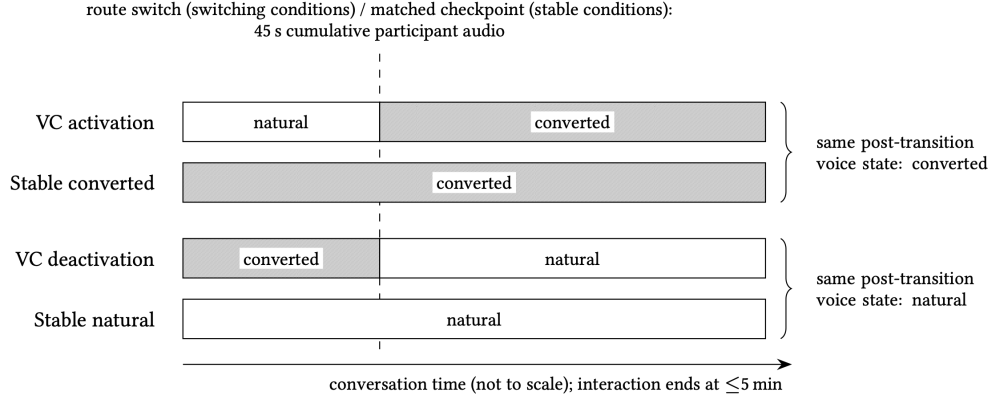


Figure 1: The four within-participant voice-state conditions.

Each hypothesis requires support from both of its component outcomes. We describe the models and multiplicity adjustment in Section 5.4.

4 Study Design and Four VC Conditions

We assigned each participant four goal-oriented record-correction conversations, one per voice-conversion condition (Figure 1 and Appendix A).

4.1 System and Voice-Conversion Pipeline

Participants interacted through the Hear Me Out platform [3], using `nvidia/personalex-7b-v1` [28] as the full-duplex conversational model. For the present study, the platform routed participant speech either unchanged or through X-VC before transmission to the model (Figure 3).

In the two switching conditions, we configured the proxy to change the VC state after 45 seconds of cumulative participant audio without resetting the session, context, or dialogue state. Based on pilot interactions, we chose this timing to place the checkpoint after the initial task presentation while leaving time to observe information use, repair, and a final account within five minutes. We stored separate unmodified, transmitted, and assistant recordings and a route-change log (Appendix A.1).

Target voices and deferred assignment. We used two fixed 10-second VCTK excerpts [37]: p229 (feminine-presenting, Southern English) and p279 (masculine-presenting, Leicester English). Before recruitment, we selected them using UTMOS, word error rate (WER), speaker similarity (SIM), and listening checks. After consent and the background questionnaire, we assigned the masculine-presenting target to women and the feminine-presenting target to men. For all other gender responses, the system selected a target at random.

Voice-conversion quality. For each converted region, we computed UTMOS, cosine similarity to the fixed target, and WER between recognitions of time-matched unmodified and converted speech (Appendix A.4). We used route events to isolate speech

transmitted while conversion was active. The 65 interactions containing converted speech had median WER 0.19 (20% exceeded 0.5), median similarity 0.58, and median UTMOS 2.61.

4.2 The Four Voice-Conversion Conditions

We used four scenarios (Appendix B):

- reopening a declined warranty request.
- correcting an inaccurate accident statement in a bicycle-injury insurance claim.
- requesting a review of an incorrectly graded assignment.
- requesting a refund for food delivered to the wrong address.

We used the counterbalancing scheme to assign each scenario to one of four conditions (Figure 1):

- **Stable natural:** natural voice throughout.
- **Stable converted:** converted voice throughout.
- **VC activation:** natural voice before and converted after the switch.
- **VC deactivation:** converted voice before and natural after the switch.

Before data collection, we created four Graeco–Latin counterbalancing configurations. Across the four configurations, every scenario appeared in every position and condition, and every condition appeared in each position. Within each assignment stratum, the system selected the least-used configuration, which fixed scenario order, condition mapping, and target voice.

4.3 Participants and Recruitment

Participants were at least 18 and able to complete spoken tasks in English. We recruited through academic and social channels and collected language background, English proficiency, prior conversational AI use, age range, and optional demographics.

We targeted 40 complete datasets and issued 40 invitation codes. Twenty-eight invitees consented and completed the background questionnaire. The other 12 codes had no consent record. Of the 28 who consented, 26 began an analytical scenario and 24 contributed at least one included interaction. Two consented participants never began an analytical scenario, and two who began contributed no included interactions. Our included sample comprised 18 men and

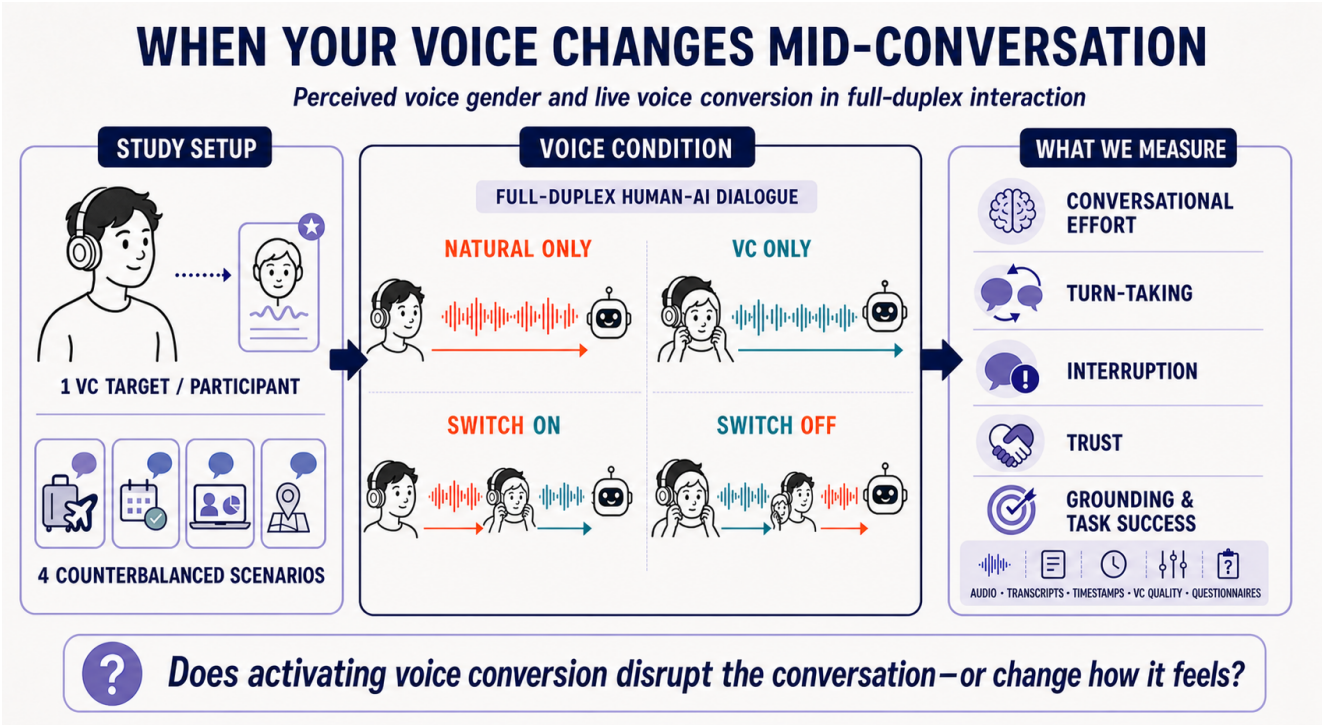


Figure 2: Overall setup in our study

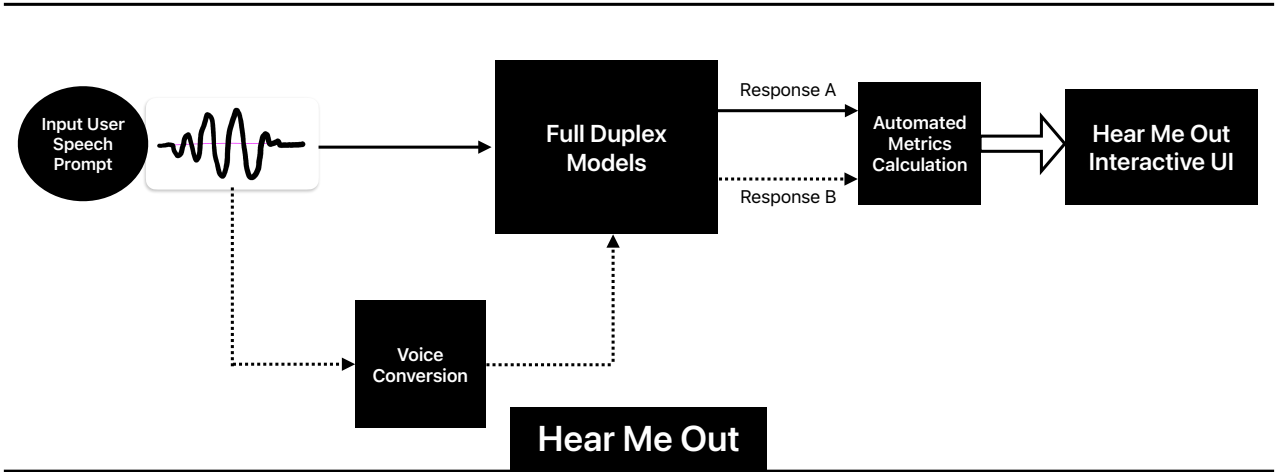


Figure 3: Hear Me Out platform architecture[3]

6 women. Twenty were aged 25–34, 21 did not list English among their first languages, and 22 had never used conversational AI or used it less than monthly.

4.4 Which Interactions We Analysed

We included an interaction only when three conditions held. First, the assigned condition, including any switch, ran as planned. Second, all required recordings were available. Third, no system failure

prevented meaningful interaction. When a technical failure occurred, we excluded only the affected interaction. We treated timeouts and participant-initiated endings as outcomes, not technical failures.

Of 125 attempts, 88 met the inclusion criteria. We excluded 13 unsubmitted attempts, 9 technical failures, and 15 whose validation was incomplete when data collection closed. Most exclusions occurred in the switching conditions, leaving fewer included interactions there than in the stable conditions (Appendix A.1).

4.5 Scenarios and Procedure

We designed each scenario around an inaccurate or contested record, two task-critical units known to the participant, a bounded institutional action, and a required final account. Participant cards specified the role, record, corrective facts, and goal. Hidden system prompts specified the assistant’s role, record, permitted actions, and required conclusion but omitted the corrective facts (Appendix B). Participants chose how to pursue the standardised goal.

Participants attended individual headphone-based sessions. After consent and the background questionnaire, the system assigned a target voice and counterbalancing configuration. We scheduled participants for an audio check, an excluded practice interaction, and four analytical scenarios in the assigned order. Before ending each interaction, they requested a final summary of the recorded information, decision, and next step. They completed a questionnaire after each interaction. After the fourth, they completed a post-session questionnaire, listened to and rated converted-speech playback, and received a debriefing.

Each interaction ended after five minutes. Nine included interactions reached this limit. Because participants had to request a final summary, timeouts or early endings could leave no eligible account. When no eligible account existed, we treated final-account accuracy as missing and calculated account-level outcomes only among interactions with an account.

4.6 Participant Questionnaire

After each interaction, participants completed eight seven-point items on effort, frustration, communication and intervention, information use and reporting, trust, and outcome confidence. They also answered categorical items on misunderstanding, false update confirmation, perceived outcome, and desire to end early. After all four interactions but before playback, they reported processing-change awareness, speaking adaptation, and their overall experience. They then heard two non-overlapping excerpts from their stable-converted interaction. They could play each excerpt one to three times before rating the converted voice on seven items. Because participants never heard conversion live, these ratings were retrospective (Appendix C).

4.7 Ethics, Consent, and Data Governance

Participants gave informed consent before data collection and could stop without giving a reason. The participant information described rights of access, correction, deletion, restriction, and consent withdrawal (Appendix C).

Invitation codes contained no direct identifiers. We used pseudonyms in analysis, excluded contact details from the dataset, and supplied fictional reference details on scenario cards.

5 Outcomes and Analysis

5.1 Using Task-Critical Information

Before data collection, we divided each scenario into two task-critical information units. For each unit, we recorded whether the participant attempted delivery, produced it completely in the unmodified recording, and transmitted it completely to the assistant. We also recorded whether delivery occurred before, after, or across the checkpoint, or remained incomplete.

For every unit that met the downstream-annotation rule below, we annotated four indicators:

- **Acknowledgement:** the assistant indicates that it has registered the unit.
- **Update claim:** the assistant states that the operative record, decision, or action has changed.
- **Operational incorporation (use in the next relevant action):** the assistant’s next relevant action reflects the unit.
- **Delayed retention (later evidence in the interaction):** the assistant reflects the unit correctly at a later probe or in the final account.

We distinguished two cases in which apparent acceptance did not change task behaviour. *Acknowledgement without uptake* occurs when the assistant acknowledges a unit but does not incorporate it operationally. *False update confirmation* occurs when the assistant explicitly claims to update the record but a later action or final account contradicts that claim.

For each indicator, we distinguished an observed failure from a missing value. If the transmitted recording showed that a unit was incomplete, we assigned no downstream indicator labels. If transmission completeness could not be checked, we included the unit only when the participant’s unmodified recording showed complete production (Appendix A.3). We marked an indicator as missing when its opportunity was unobservable or inapplicable. Each indicator therefore has its own denominator of observed opportunities. We assessed final-account accuracy from the first assistant response to the participant’s request for a final summary, or from an earlier turn only when it clearly provided a spontaneous final account. When no eligible account existed, we marked accuracy as missing. We defined the interaction-level outcome demonstrated task-information uptake as one only when both units showed correct operational incorporation and final-account accuracy. We report indicator results for observed opportunities among units admitted by the transmission or production-proxy rule, and we report transmission completeness separately.

We classified task outcome at the interaction level according to the bounded institutional action reached. The ordinal levels are scenario-specific, but their highest and lowest levels carry a common interpretation across scenarios. We measured behavioural repair burden from human-reviewed repair annotations. The post-checkpoint repair count is the co-primary repair outcome.

5.2 Participant Measures

Post-scenario items measured participants’ effort, frustration, trust, and confidence in the resulting task state (Section 4.6). We measured behavioural repair separately through human-reviewed repair annotations. We analysed participants’ reports of misunderstanding, false update confirmation, task outcome, and desire to end early alongside the corresponding human-reviewed labels.

5.3 LLM-Assisted Transcript Annotation and Human Review

We classified semantic outcomes and repair behaviour through LLM annotation followed by human review. The predefined labels covered information acknowledgement, information use and retention, repair, and task completion. Using pilot data, we finalised the annotation guide, annotator and checker prompts, output format, model and provider, decoding settings, and review rules before analysing condition differences.

LLM annotation and checking. We made separate annotator and checker calls to `openai/gpt-oss-120b` through Berget AI. Both used temperature 0.0 and a 16,384-token output limit.

We submitted only pseudonymised, speaker-labelled text: aligned unmodified and transmitted transcripts, assistant turns, the system prompt, information units, permitted actions, and annotation guide. We withheld audio, condition and target-voice labels, questionnaire responses, contact details, and direct identifiers. The provider retained no text, and personal data remained within the EU/EEA.

The annotator returned fixed-format labels, supporting turns, repair events and categories, task outcome, confidence, and review flags. The checker compared these labels with cited evidence and flagged omissions, contradictions, and rule violations.

We derived interaction-level information use from unit labels using fixed rules rather than asking either LLM to assign it (Appendix A.3).

Human review and final labels. We referred an interaction for human review if the LLM output was invalid, lacked supporting evidence, contained contradictory labels, violated a predefined rule, received a confidence score below 0.8, or disagreed with the checker. We withheld condition labels from the human reviewer.

In addition to every flagged interaction, the human reviewer independently annotated at least 25% of otherwise valid interactions without seeing the LLM annotator’s labels. We balanced this sample across conditions and scenarios and measured agreement before resolving differences. Under the predefined rule, if a variable fell below its agreement threshold, the human reviewer examined that variable in every interaction. In fact, most variables fell below threshold, so the reviewer examined all 92 annotated interactions, including the four excluded from the main analysis. We used the human-reviewed labels in place of the LLM output and retained a record of every change.

Compared with the LLM annotator’s initial output, the human reviewer changed 27.2% of values across the 15 label types evaluated. Eighty-two of the 92 interactions had at least one changed value. Changes occurred most often in repair counts, final-account accuracy, and labels indicating whether the assistant claimed to update the relevant information.

Table 1: Research questions, outcomes, and hypotheses.

RQ	Outcomes (phenomenon level)	Hypotheses
RQ1	Use of task-critical information and post-checkpoint repair burden (co-primary, with an amended hypothesis-guided analysis for information use)	H1a, H1b
RQ2	Participant experience, task outcome, final-account accuracy, false update confirmation, and retrospective converted-voice ratings (secondary, descriptive), together with relations between behaviour and effort, frustration, trust, and outcome confidence	H2a, H2b

Agreement between the LLM annotator and human reviewer. We measured agreement using 28 interactions that the human reviewer annotated without seeing the LLM annotator’s output. Twenty-four of these interactions were included in the main analysis, with six from each condition.

For the remaining 64 interactions, the human reviewer examined and corrected label sets initially populated from the LLM annotator’s output. We excluded these interactions from independent agreement estimates because the reviewer had seen the LLM labels.

Only the task-outcome label met its predefined agreement threshold, with a linear-weighted κ of 0.60 across 24 interactions. Agreement for all other measured variables fell below threshold. Across the six labels assigned to individual critical-information units, κ ranged from -0.03 to 0.40 , based on 23 to 48 paired units. Final-account accuracy had a κ of 0.33, the presence of any repair had a κ of 0.50, and repair count had an ICC(2, 1) of 0.43. We therefore used the LLM for preliminary annotation only. All subsequent analyses use human-reviewed labels.

5.4 Analysis

We begin with descriptive results for information use, task outcomes, repair, and participant reports. We then test the matched RQ1 sequence contrasts and RQ2 associations. Table 2 lists each hypothesis-guided outcome, analytical unit, model, and effect measure. We treat the stable-condition contrast, individual indicators, task outcomes, and self-report summaries as secondary analyses.

We originally planned one information-use measure per interaction: whether the assistant used both units correctly and reported them correctly at the end. Only 6 of 71 classifiable interactions met this criterion, too few to support a four-condition comparison. After observing this sparsity but before fitting any model, we amended the analysis to use one row per unit per observed indicator. This produced 584 rows: 152 acknowledgement, 151 update-claim, 152 incorporation, and 129 retention observations. The rows came from 176 predefined unit slots across 88 interactions and 24 participants and were therefore not independent. Because we made this decision after observing the data and pooled indicators with different meanings, we describe the amended analysis as hypothesis-guided rather than fully confirmatory. We did not change the hypotheses, comparisons, or predicted directions. We report the original

Table 2: Hypothesis-guided outcome, analytical unit, model, and reported effect.

Outcome	Unit	Model	Effect
Information-use indicators (H1a, H1b, H2a)	Critical-information unit $\times$ indicator	Logistic mixed	Odds ratio
Post-checkpoint repair (H1a, H1b)	Interaction	Poisson mixed	Ratio of expected repair counts
Effort and frustration (H2b)	Interaction	Cumulative-link mixed	Odds ratio

interaction-level outcome descriptively. All models include condition, scenario, and ordinal position as fixed effects and a participant random intercept. Unit-level models also include indicator type, scenario-specific unit identity, and an interaction random intercept. A sensitivity model adds a unit intercept. We adjust for scenario because the condition-by-scenario cells are small and uneven and the four scenarios elicited different repair counts. We present unadjusted condition summaries only as descriptive results. Appendix A.2 gives the model checks, software, and fitting details.

We use all valid observed values without imputation. Each indicator uses its observed opportunities. Delayed retention requires a later probe or final account. Final-account accuracy requires a final account. We treat “not applicable” responses to the interrupt-and-correct item as missing.

To control family-wise error across four hypotheses, we treat the hypothesis-guided tests as one family and apply Holm’s step-down procedure. Both outcomes must support a hypothesis, so its larger component p -value enters the adjustment. Tests are two-sided, and we report two-sided Wald intervals. We do not interpret the H2 associations causally.

6 Generative AI Usage Disclosure

We used AI to assist with parts of the code for the study interface and for the analysis, figures, and tables, and with drafting prompts for the scenarios. We reviewed and modified all of it before use. Every reported statistic, table entry, and plotted value was computed from the recorded data by that code rather than generated by a model.

7 Results

7.1 Data Included in the Analysis

We analysed 88 interactions from 24 participants and 176 predefined critical-information units: 23 stable natural, 23 stable converted, 21 VC activation, and 21 VC deactivation. Because the realised condition–scenario cells were uneven, we adjusted comparisons for scenario. We report unadjusted condition summaries.

A human reviewer finalised every label used in the analysis (Section 5.3). Twenty-nine interactions had no final account. Four “not applicable” responses left 84 interrupt-and-correct ratings. We obtained 88 responses for each of the other post-scenario items.

7.2 Use of Task-Critical Information and Task Outcomes

Participants attempted 172 of 176 task-critical information units. They produced 159 of 175 observable units completely (90.9%) and

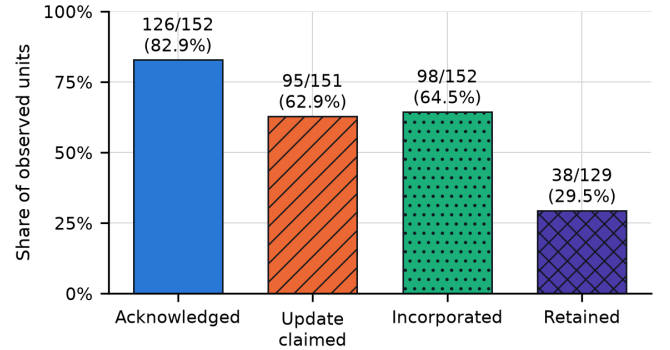


Figure 4: The assistant acknowledged information more often than it retained it. Each bar reports its numerator, denominator, and percentage. The indicators are not stages of a sequence.

transmitted 148 of 154 checkable units completely (96.1%), but downstream use was inconsistent. We exclude uncheckable units from these denominators.

The assistant acknowledged information much more often than it retained it (Figure 4). It acknowledged 126 of 152 observable units, claimed an update for 95 of 151, incorporated 98 of 152, and retained 38 of 129. We assessed each indicator independently when the interaction made it observable, so these counts do not form a step-by-step funnel.

Six of 71 classifiable interactions (8.5%) satisfied the stricter interaction-level outcome. We classified 18 interactions as failures before a final account because incorporation had already failed. Among the 17 unclassified interactions, 11 lacked both a final account and an earlier observed failure, and 6 had an unobservable incorporation or retention value.

Participants reported fully achieving the goal in 44 interactions, whereas human review classified 36 interactions at the strongest permitted action and 9 at the lowest outcome level (Figure 5). Of 59 final accounts, the human reviewer classified 6 as fully accurate, 37 as partly accurate, and 16 as inaccurate. The reviewer identified false update confirmation in 45 of 64 observable interactions (70.3%). In 30 of these 45 interactions (67%), participants denied the event or reported that the assistant had made no such claim. Three reports expressed uncertainty. Reaching an action therefore did not ensure an accurate account, and participants’ reports often missed the discrepancy.

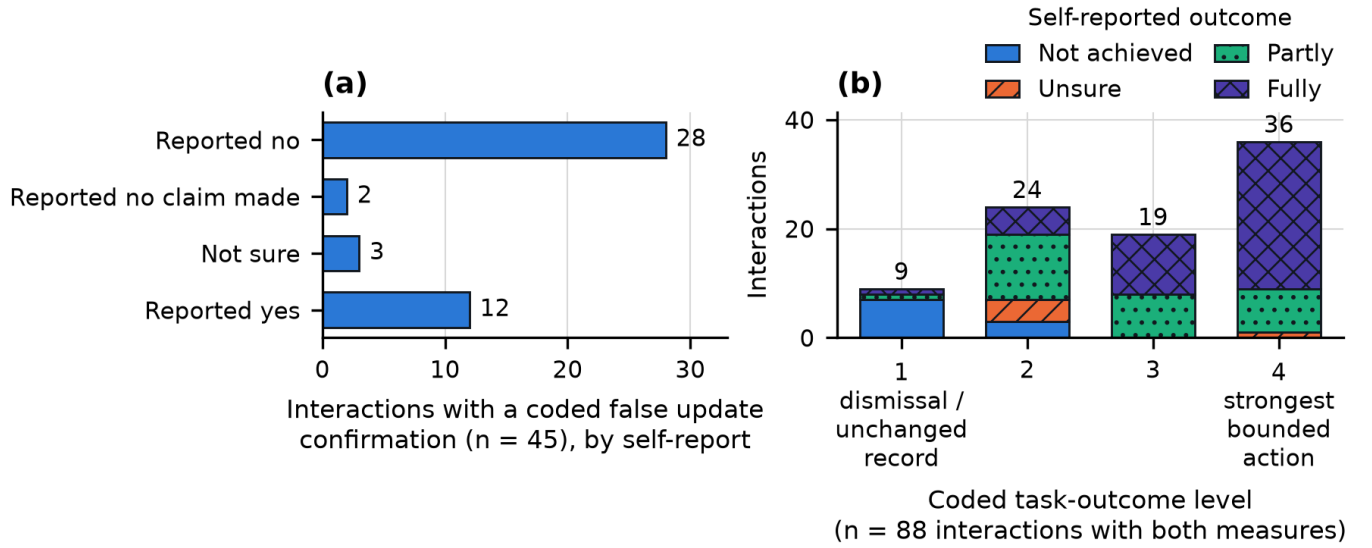


Figure 5: Participant reports often diverged from human-reviewed outcomes. Panel (a) compares false update confirmations identified through human review with subsequent self-reports. Panel (b) compares human-reviewed task-outcome levels with self-reported outcomes.

7.3 Voice-Condition Comparisons and Repair

Repair was common, but we found no support for either H1 sequence hypothesis. Participants produced 283 annotated repair moves across 88 interactions (mean 3.22, median 3, maximum 12), with none in 7 interactions. Of these moves, 246 (86.9%) occurred at or after the switch or matched checkpoint. We structured the tasks to elicit correction later in the interaction, so we do not interpret this distribution as a condition contrast (Figure 6).

H1a compared VC activation with stable converted input. The ratio of expected post-checkpoint repair counts was 1.09 (95% CI [0.76, 1.56], $p = 0.641$), and the odds ratio for a satisfied information-use indicator was 1.10 (95% CI [0.52, 2.32], $p = 0.797$). These estimates did not support the predicted activation penalty.

H1b compared VC deactivation with stable natural input. The ratio of expected post-checkpoint repair counts was 1.64 (95% CI [1.11, 2.41], $p = 0.013$). The estimated odds that an information-use indicator was satisfied were lower, but the 95% CI included 1 (odds ratio 0.56, 95% CI [0.27, 1.17], $p = 0.121$). Because H1b required support from both outcomes, we found no support for the hypothesis after Holm adjustment ($p = 0.383$).

The deactivation repair-count estimate pointed in the same direction in every sensitivity analysis, but no sensitivity analysis supported a hypothesis. The smallest adjusted hypothesis-level p was 0.089 (Appendix Table 6). The estimates shifted across analyses, and the intervals were too wide to support an equivalence claim.

In the secondary stable converted–stable natural comparison, the ratio of expected post-checkpoint repair counts was 1.50 (95% CI [1.02, 2.22], $p = 0.040$). The odds ratio for a satisfied information-use indicator was 0.63 (95% CI [0.30, 1.30], $p = 0.209$). We did not include this secondary comparison in the multiplicity-controlled hypothesis family, so we treat it as exploratory.

Answer to RQ1. In the amended, hypothesis-guided pooled analysis, we found no support for either H1 sequence hypothesis. The sequence contrasts do not isolate an instantaneous switch effect or establish equivalence.

7.4 Participant Experience by Condition

Participant ratings were widely dispersed (pooled SD s 1.82–2.12). Participants gave the highest mean ratings to the assistant’s use of decisive information (5.56) and the accuracy of its final response (5.40). Means for effort, frustration, communicating and correcting, trust, and outcome confidence ranged from 4.33 to 4.91 (Table 5).

Participants rated the assistant relatively favourably on information use and final-response accuracy, even though it incorporated information in only 98 of 152 observable opportunities (64.5%) and produced only 6 fully accurate final accounts among 59 observed accounts. The questionnaire and transcript measures use different criteria and denominators, but the ratings still failed to reveal many problems identified through human review.

We report the unadjusted ratings by condition in Table 5. We did not estimate condition contrasts for these measures.

7.5 Post-Session Awareness, Adaptation, and Converted-Playback Ratings

Twenty-two participants answered the post-session questionnaire before playback. Six noticed or suspected a processing change, 12 did not, and 4 were unsure. Eleven reported intentionally changing how they spoke, 10 did not, and 1 was unsure. The most common adaptations were greater directness or assertiveness (9) and repetition or rephrasing (8). Because participants reported once after all four interactions, we cannot link these recollections to a particular interaction.

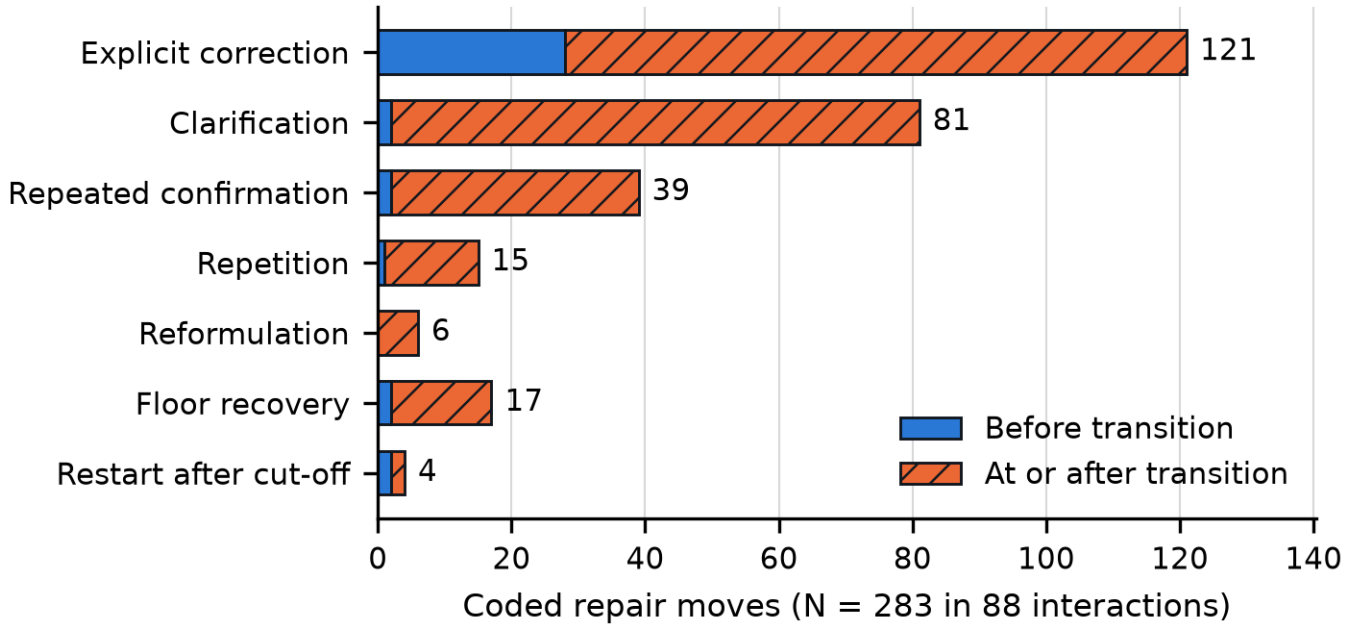


Figure 6: Coded repair moves by category. Explicit corrections were most common (42.8%), followed by clarifications (28.6%) and repeated confirmations (13.8%). Most repairs occurred at or after the switch or matched checkpoint. Categories follow the frozen annotation-guide order. The stable-condition split uses the matched checkpoint.

Table 3: Unadjusted descriptions of demonstrated task-information uptake and repair by condition. The uptake row reports the strict interaction-level outcome. Adjusted contrasts use the unit-level indicators and appear in Table 4.

	Stable natural	Stable converted	VC activation	VC deactivation
Interactions	23	23	21	21
Demonstrated uptake	1 / 16	2 / 20	2 / 17	1 / 18
Repair moves, mean (SD)	2.52 (1.90)	3.57 (2.86)	3.52 (2.04)	3.29 (2.51)
at or after the checkpoint	2.04 (1.97)	3.09 (2.56)	3.05 (1.75)	3.05 (2.40)

Participants’ mean overall ratings were 4.32 for trust, 3.55 for pleasantness, and 3.18 for willingness to use a spoken assistant for a similar task.

Afterwards, the 22 respondents rated two converted-speech excerpts as understandable ($M = 5.77$, $SD = 1.23$) and different from their natural voice ($M = 6.09$, $SD = 1.02$). Their mean ratings for naturalness, self-representation, comfort, and willingness ranged from 3.77 to 4.59. Their mean audible-artifact rating was 4.00 ($SD = 2.37$). Because participants heard these excerpts only after the session, the ratings cannot tell us how they experienced the live interactions.

7.6 Behavioural–Perceptual Associations

We found no support for either behavioural–perceptual hypothesis after adjustment. For H2a, outcome confidence was positively associated with the odds that an information-use indicator was satisfied (odds ratio 1.17, 95% CI [1.03, 1.33], $p = 0.013$). The trust estimate’s 95% CI included 1 (odds ratio 1.12, 95% CI [0.98, 1.29], $p = 0.096$). For H2b, frustration was positively associated with

repair burden (odds ratio 1.32, 95% CI [1.05, 1.65], $p = 0.018$). The effort estimate’s 95% CI included 1 (odds ratio 1.16, 95% CI [0.95, 1.42], $p = 0.137$). Each hypothesis required both component associations. Neither met this rule, and both had a Holm-adjusted hypothesis-level $p = 0.383$ (Table 4). We do not interpret these observational associations causally.

Answer to RQ2. Participant reports diverged from human-reviewed task outcomes and often missed false update confirmations. Among the 45 cases identified through human review, 30 reports denied the event or said that no claim had been made, and 3 were uncertain. We did not estimate condition contrasts for self-report, and neither paired association hypothesis survived Holm adjustment. Because awareness was session-wide and playback was retrospective, neither measure captures experience of a specific live transition.

Table 4: Estimates for H1a–H2b. Both outcomes must support a hypothesis. Component p -values are unadjusted. The final column reports the Holm-adjusted hypothesis-level p -value.

Endpoint	Estimate	95% CI	p	Holm p
H1a VC activation vs stable converted				
Post-checkpoint repair moves (ratio of expected counts)	1.09	[0.76, 1.56]	0.641	
Information-use indicator satisfied (odds ratio)	1.10	[0.52, 2.32]	0.797	0.797
H1b VC deactivation vs stable natural				
Post-checkpoint repair moves (ratio of expected counts)	1.64	[1.11, 2.41]	0.013	
Information-use indicator satisfied (odds ratio)	0.56	[0.27, 1.17]	0.121	0.383
H2a Information use and perceived state				
Trust, per scale point (odds ratio)	1.12	[0.98, 1.29]	0.096	
Outcome confidence, per scale point (odds ratio)	1.17	[1.03, 1.33]	0.013	0.383
H2b Repair burden and perceived cost				
Effort, per repair move (odds ratio)	1.16	[0.95, 1.42]	0.137	
Frustration, per repair move (odds ratio)	1.32	[1.05, 1.65]	0.018	0.383

Table 5: Post-scenario ratings by condition: means (SDs) on seven-point scales. Higher scores mean more effort and frustration on the first two rows and a better experience on the others. We excluded not-applicable responses from interrupt-and-correct and did not estimate condition contrasts.

Measure	Stable natural	Stable converted	Activation	Deactivation	Overall
Effort	4.43 (2.02)	4.74 (2.03)	5.24 (1.55)	4.52 (2.18)	4.73 (1.95)
Frustration	3.65 (1.85)	4.61 (2.06)	4.86 (1.88)	4.24 (2.14)	4.33 (2.00)
Finished communicating	4.57 (2.06)	5.09 (1.98)	4.52 (1.97)	4.86 (2.10)	4.76 (2.01)
Could interrupt or correct	4.55 (2.24)	4.59 (1.94)	4.25 (2.20)	5.25 (1.59)	4.65 (2.01)
Decisive information understood	5.52 (1.81)	5.22 (2.02)	6.14 (1.42)	5.38 (1.94)	5.56 (1.82)
Final response accurate	5.04 (2.12)	5.70 (1.96)	5.33 (1.62)	5.52 (2.06)	5.40 (1.94)
Trust	4.61 (2.02)	4.57 (2.02)	4.67 (1.93)	5.00 (1.92)	4.70 (1.95)
Outcome confidence	4.78 (2.35)	4.96 (1.87)	4.52 (2.16)	5.38 (2.13)	4.91 (2.12)
Interactions	23	23	21	21	88

8 Discussion

8.1 Main Findings

We followed predefined task-critical information from the participant’s speech through transmission, acknowledgement, explicit update claims, operational use, and later retention. Participants usually delivered the information. Among the 154 units with checkable transmission, 148 reached the assistant completely. The more consequential failures followed delivery. The assistant acknowledged information substantially more often than it retained it, and explicit update claims often diverged from later incorporation. Only 6 of 59 final accounts were fully accurate, and 45 of 64 interactions with an observable opportunity contained a false update confirmation. Our central finding is that conversational evidence of acceptance often diverged from later behaviour, not simply that the conversations contained errors.

In the amended, hypothesis-guided pooled analysis, we found no support for either H1 sequence hypothesis after Holm adjustment. The deactivation sequence had a higher expected post-checkpoint repair count than stable natural speech, but the corresponding

information-use estimate remained uncertain. We cannot distinguish disruption caused by switching from the accumulated effects of earlier converted speech, conversion artifacts, or differences in where the checkpoint fell in the evolving dialogue. The secondary comparison between the stable conditions also suggested more repair with converted speech, but that comparison was unadjusted and exploratory. We therefore do not interpret the results as evidence of an isolated transition effect or as evidence that the conditions were equivalent.

Participants’ impressions provided a different view of the same conversations. They rated the assistant relatively favourably on understanding decisive information and producing an accurate final response, even though delayed retention and fully accurate final accounts were uncommon. In most interactions with an annotated false update confirmation, participants did not report detecting it. Outcome confidence and frustration each showed one endpoint-level association in the predicted direction. After adjustment, however, we found no support for either paired hypothesis. Self-report

remains necessary for understanding experienced effort, frustration, and trust, but it cannot replace checking what the assistant subsequently did with the information.

Participants never heard their converted voice during the live conversations. Their playback ratings therefore describe retrospective reactions to a recording rather than the experience of hearing their voice change while interacting. The post-session awareness and adaptation questions likewise combine recollections across four interactions and cannot show whether participants perceived a particular transition. In this model and task configuration, fluent acknowledgement and confirmation were unreliable indicators of later task behaviour. Evidence about differences between voice-state sequences remained limited.

8.2 What This Changes for Full-Duplex Evaluation

Full-duplex benchmarks commonly measure whether users can gain the conversational floor, interrupt, and receive a timely response [20, 25]. Floor access and response timing are necessary for usable interaction, but they do not establish whether the system subsequently acts on what the user said. In our data, acknowledgement, update claims, incorporation, and retention diverged often enough that no one measure could stand in for the others.

Evaluators should therefore trace task-critical information across multiple observable points: production in the participant signal, presence in the transmitted signal, acknowledgement, any explicit update claim, incorporation into the next relevant action, and retention at a later probe or final account. They should retain a separate opportunity denominator for each measure rather than treat an unobservable event as a failure or success. A final readback is particularly important because users may not challenge an immediate confirmation when an interaction ends before they request an account.

Recent benchmarks extend timing-based evaluation with correction and entity tracking [19]. We argue that evaluators should follow corrections through their downstream consequences rather than stop at the response immediately following them. Doing so exposes false update confirmations and persistent inconsistencies in later task behaviour and final accounts that could otherwise appear to be successful grounding.

8.3 What This Means for Design

Our results suggest three design priorities across the tested conditions. First, systems should distinguish conversational acknowledgement from a completed task-state update. An assistant should not say that information has been recorded, corrected, or accepted unless the corresponding state is available for verification. When an update is provisional or cannot be completed, the response should state that limitation explicitly.

Second, users need a way to inspect and correct the operative state directly. A compact structured readback, editable summary, or visible record of the decisive facts could reduce reliance on fluent verbal confirmation. For consequential tasks, the interface should surface unresolved conflicts and request confirmation before presenting a decision as final.

Third, systems should support repair without requiring users to infer whether a correction persisted. Systems can do this by preserving interruption access, asking targeted clarification questions when the signal is uncertain, and providing a final account before timeout or termination. They should treat a change in acoustic input quality as a reason to handle uncertainty, not as evidence about speaker identity or intent.

These recommendations address failures we observed across all conditions.

9 Limitations and Ethical Considerations

9.1 Limitations

We evaluated one version of a full-duplex model, one X-VC configuration, two fixed target voices, and four scaffolded record-correction scenarios. We therefore limit our claims to this system and task configuration. Other models, conversion systems, target voices, languages, or interaction domains may behave differently.

The manipulation changed more than one vocal attribute. Each target combined timbre, pitch, accent, and perceived gender presentation, while conversion also introduced variable distortion, intelligibility loss, and latency. Any condition difference could therefore reflect signal quality or cumulative exposure to converted speech rather than the switching sequence itself. In addition, the reported WER uses an ASR transcript of the unmodified speech as its reference rather than a human transcript, so it measures disagreement between ASR outputs rather than human-verified word accuracy.

We linked target assignment to reported gender: participants who reported being men and women received different fixed targets, while we assigned target voices at random for all other responses. The design therefore confounds target voice with participant gender, and the groups were unequal in size. We neither estimate target-specific effects nor interpret either target as an independent manipulation of gender or another demographic category.

The final sample comprised 24 participants and fell below our prerecruitment target. Most participants were men, and most were aged 25–34. Most did not list English as a first language, and most had little prior experience with spoken conversational AI. The sample does not support broad generalisation across gender, age, language background, accent, or prior experience. Attempt-level exclusions and incomplete within-participant datasets may also introduce selection that counterbalancing and model adjustment cannot remove.

The four fictional service scenarios improved experimental control but did not reproduce the material or emotional stakes of real disputes. They also differed in evidence load, permitted actions, and prompt strictness. Although we adjusted the models for scenario and ordinal position, the realised condition-by-scenario cells were small and uneven. We triggered the transition by cumulative participant audio rather than at a common semantic event, so it occurred at different points in the task. The sequence contrasts consequently combine the transition, earlier voice exposure, and task history.

Twenty-nine interactions produced no eligible final account. Final-account accuracy and some retention measures are therefore conditional on reaching an observable opportunity, and those opportunities are unlikely to be missing at random. The corresponding proportions describe observed or probed interactions rather than all

interactions. Six units with uncheckable transmission entered downstream annotation under the prespecified unmodified-production proxy rule. The pooled analysis is therefore conditional on admission under either stated rule rather than on confirmed transmission alone.

One human reviewer determined the final semantic and repair labels under the predefined annotation guide. Most initial LLM labels failed their agreement thresholds. The reviewer changed 27.2% of the compared values. Only task outcome met the predefined agreement criterion. For most interactions, the reviewer corrected labels initially proposed by the LLM rather than producing an independent second annotation. We therefore treat the final labels as one trained reviewer’s final assessment, not as consensus annotation or validated automated measurement.

Our planned interaction-level uptake outcome was too sparse for the intended comparison. After observing this sparsity, but before fitting the condition models, we changed to a pooled unit-by-indicator analysis and retained the original outcome as descriptive. This amendment was not independent of the data and changed both the analytical unit and the measured construct. We therefore describe the resulting estimates as amended and hypothesis-guided, not fully confirmatory. With only 24 participant clusters, intervals remain wide and nonsignificant estimates do not demonstrate equivalence. We did not include the secondary stable-condition comparison in the multiplicity adjustment. The participant associations rely on single questionnaire items without reliability estimates.

Finally, participants did not hear the converted signal during the live interactions. Playback ratings measure retrospective reactions to recorded converted speech, while the awareness and adaptation questions summarise recollection across the complete session. The study cannot determine how hearing one’s own converted voice during an interaction would affect experience or behaviour.

9.2 Ethical Considerations

Participants gave informed consent before we collected background information or audio. We disclosed that their transmitted speech could remain natural, be converted, or change conversion state during an interaction. We withheld the assigned condition, direction, and timing until debriefing. Revealing that schedule in advance could have changed how participants spoke and how they attributed the assistant’s behaviour. Participants could stop without giving a reason. The participant information described their data rights and the institutional retention policy (Appendix C). We used fictional reference details in every scenario and asked participants not to provide personal information.

The conversion and annotation pipelines limited the data exposed outside the study interface. We kept converted audio within the study infrastructure, did not return it to participants during live interactions, and excluded it from published material. The LLM annotator and checker received pseudonymised, speaker-labelled text rather than audio, condition labels, questionnaire responses, contact details, or target identifiers (Section 5.3).

Voice conversion has benefits and potential for misuse: it can give new identity to a consenting speaker, but it can also enable impersonation or misrepresentation. A mid-conversation change could create an additional opportunity for such misuse. Moreover,

any apparent voice-linked difference in this study would remain a property of two individual target voices under one system, not evidence about a demographic group. Because target assignment depended partly on reported gender, we treat target voice and participant gender as confounded and make no causal claims about gender, accent, or other social groups.

Formal ethical review was not required for this study under national regulations and the host institution’s guidance for research with human participants. The host institution is the data controller for the recordings, transcripts and questionnaire responses. Participants received written participant information during recruitment before consent.

10 Conclusion

We examined whether reaching a post - checkpoint voice state through a switching sequence differs from remaining in that state throughout a full-duplex task. In the amended, hypothesis-guided pooled analysis of 88 interactions from 24 participants, we did not detect sequence-related differences in information-use indicator satisfaction. The deactivation sequence had a higher expected post - checkpoint repair count than stable natural speech, but neither H1 sequence hypothesis met its joint decision rule after Holm adjustment. The larger and more consistent finding concerned grounding itself. Although 148 of 154 units with checkable transmission reached the assistant completely, only 6 of 71 classifiable interactions met the end-to-end criterion and only 6 of 59 final accounts were fully accurate. We identified a false update confirmation in 45 of 64 observable interactions. Participants’ reports often failed to reveal these discrepancies. Evaluators and designers should therefore look beyond turn-taking and fluent acknowledgement. Evaluators should test whether corrections shape later actions and final accounts. Designers should give users a way to verify the resulting task state.

References

- [1] Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–12. doi:10.1145/3290605.3300484
- [2] Erin Beneteau, Olivia K. Richards, Mingrui Zhang, Julie A. Kientz, Jason Yip, and Alexis Hiniker. 2019. Communication Breakdowns Between Families and Alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–13. doi:10.1145/3290605.3300473
- [3] Shree Harsha Bokkahalli Satish, Gustav Eje Henter, and Éva Székely. 2025. Hear Me Out: Interactive evaluation and bias discovery platform for speech-to-speech conversational AI. In *26th Interspeech Conference 2025, Rotterdam, Netherlands, Kingdom of the, August 17-21, 2025*. International Speech Communication Association, 2151–2152.
- [4] Shree Harsha Bokkahalli Satish, Christoph Minixhofer, Maria Teleki, James Caverlee, Ondřej Klejch, Peter Bell, Gustav Eje Henter, and Éva Székely. 2026. The Voice Behind the Words: Quantifying Intersectional Bias in SpeechLLMs. *arXiv preprint arXiv:2603.16941* (2026). arXiv:2603.16941 [cs.CL] <https://arxiv.org/abs/2603.16941>
- [5] Shree Harsha Bokkahalli Satish, Maria Teleki, Christoph Minixhofer, Ondřej Klejch, Peter Bell, and Éva Székely. 2026. “Walk a Mile in My Voice”: Voice Conversion Shapes Trust, Attribution, and Empathy in Human–AI Speech Interactions. In *Companion Proceedings of the 31st International Conference on Intelligent User Interfaces*. Association for Computing Machinery, 126–130. doi:10.1145/3742414.3794737
- [6] Susan E. Brennan and Eric A. Hulteen. 1995. Interaction and Feedback in a Spoken Language System: A Theoretical Framework. *Knowledge-Based Systems* 8, 2–3 (1995), 143–151. doi:10.1016/0950-7051(95)98376-H

- [7] Mollie E. Brooks, Kasper Kristensen, Koen J. van Benthem, Arni Magnusson, Casper W. Berg, Anders Nielsen, Hans J. Skaug, Martin Maechler, and Benjamin M. Bolker. 2017. glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal* 9, 2 (2017), 378–400. doi:10.32614/RJ-2017-066
- [8] Javier Chiyah-Garcia, Alessandro Suglia, and Arash Eshghi. 2024. Repairs in a Block World: A New Benchmark for Handling User Corrections with Multi-Modal Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11523–11542. doi:10.18653/v1/2024.emnlp-main.643
- [9] Junhyuk Choi, Ro-hoon Oh, Jihwan Seol, and Bugeun Kim. 2025. VoiceBBQ: Investigating Effect of Content and Acoustics in Social Bias of Spoken Language Model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Suzhou, China, 28725–28736. doi:10.18653/v1/2025.emnlp-main.1461
- [10] Rune Haubo Bojesen Christensen. 2023. *ordinal—Regression Models for Ordinal Data*. R package version 2023.12-4.1. <https://CRAN.R-project.org/package=ordinal>
- [11] Herbert H. Clark and Susan E. Brennan. 1991. Grounding in Communication. In *Perspectives on Socially Shared Cognition*, Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley (Eds.). American Psychological Association, Washington, DC, 127–149. doi:10.1037/10096-006
- [12] Andrea Cuadra, Shuran Li, Hansol Lee, Jason Cho, and Wendy Ju. 2021. My Bad! Repairing Intelligent Voice Assistant Errors Improves Interaction. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–24. doi:10.1145/3449101
- [13] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: A Speech-Text Foundation Model for Real-Time Dialogue. *arXiv preprint arXiv:2410.00037* (2024). arXiv:2410.00037 [cs.CL] <https://arxiv.org/abs/2410.00037>
- [14] Google. 2025. *Gemini Live audio updates help conversations feel more natural*. Google. <https://blog.google/products-and-platforms/products/gemini/gemini-live-audio-updates/>
- [15] Advait Gosai, Tyler Vuong, Utkarsh Tyagi, Steven Li, Wenjia You, Miheer Bavare, Arda Uçar, Zhongwang Fang, Brian Jang, Bing Liu, and Yunzhong He. 2026. Audio MultiChallenge: A Multi-Turn Evaluation of Spoken Dialogue Systems on Natural Human Interaction. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, San Diego, California, United States, 35740–35770. doi:10.18653/v1/2026.acl-long.1654
- [16] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial Disparities in Automated Speech Recognition. *Proceedings of the National Academy of Sciences* 117, 14 (2020), 7684–7689. doi:10.1073/pnas.1915768117
- [17] Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. MT-Eval: A Multi-Turn Capabilities Evaluation Benchmark for Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Miami, Florida, USA, 20153–20177. doi:10.18653/v1/2024.emnlp-main.1124
- [18] Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2026. LLMs Get Lost in Multi-Turn Conversation. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Best Paper Award. arXiv:2505.06120. arXiv:2505.06120
- [19] Guan-Ting Lin, Shih-Yun Shan Kuan, Jiatong Shi, Kai-Wei Chang, Siddhant Arora, Shinji Watanabe, and Hung-yi Lee. 2025. Full-Duplex-Bench-v2: A Multi-Turn Evaluation Framework for Duplex Dialogue Systems with an Automated Examiner. *arXiv preprint arXiv:2510.07838* (2025). arXiv:2510.07838 [cs.CL] <https://arxiv.org/abs/2510.07838>
- [20] Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H. Liu, and Hung-yi Lee. 2025. Full-Duplex-Bench: A Benchmark to Evaluate Full-Duplex Spoken Dialogue Models on Turn-Taking Capabilities. *arXiv preprint arXiv:2503.04721* (2025). arXiv:2503.04721 [cs.CL] <https://arxiv.org/abs/2503.04721>
- [21] Yi-Cheng Lin, Wei-Chih Chen, and Hung-yi Lee. 2024. Spoken stereoset: on evaluating social bias toward speaker in speech large language models. In *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 871–878.
- [22] Robert J. Moore, Sungeun An, and Olivia H. Marrese. 2024. Understanding Is a Two-Way Street: User-Initiated Repair on Agent Responses and Hearing in Conversational Interfaces. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–26. doi:10.1145/3641026
- [23] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for How Users Overcome Obstacles in Voice User Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–7. doi:10.1145/3173574.3173580
- [24] Ziqian Ning, Shuai Wang, Pengcheng Zhu, Zhichao Wang, Jixun Yao, Lei Xie, and Mengxiao Bi. 2024. DualVC 3: Leveraging Language Model Generated Pseudo Context for End-to-End Low-Latency Streaming Voice Conversion. In *Interspeech* 2024. 197–201. doi:10.21437/Interspeech.2024-1857
- [25] Atsumoto Ohashi, Neil Zeghidour, Alexandre Défossez, and Eugene Kharitonov. 2026. Multi-Faceted Interactivity Alignment in Full-Duplex Speech Models. *arXiv preprint arXiv:2606.11167* (2026). arXiv:2606.11167 [cs.CL] <https://arxiv.org/abs/2606.11167>
- [26] OpenAI. 2026. *Introducing GPT-Live*. OpenAI. <https://openai.com/index/introducing-gpt-live/>
- [27] R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- [28] Rajarshi Roy, Jonathan Raiman, Sang-gil Lee, Teodor-Dumitru Ene, Robert Kirby, Sungwon Kim, Jaehyeon Kim, and Bryan Catanzaro. 2026. PersonaPlex: Voice and Role Control for Full Duplex Conversational Speech Models. *arXiv preprint arXiv:2602.06053* (2026). arXiv:2602.06053 [cs.CL] <https://arxiv.org/abs/2602.06053>
- [29] Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The Preference for Self-Correction in the Organization of Repair in Conversation. *Language* 53, 2 (1977), 361–382. doi:10.2307/413107
- [30] Chloe Sekkat, Fanny Leroy, Salima Mdahaffar, Blake Perry Smith, Yannick Estève, Joseph Dureau, and Alice Coucke. 2024. Sonos Voice Control Bias Assessment Dataset: A Methodology for Demographic Bias Assessment in Voice Assistants. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. ELRA and ICCL, Torino, Italia, 15056–15075. <https://aclanthology.org/2024.lrec-main.1309/>
- [31] Omar Shaikh, Kristina Gligorić, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2024. Grounding Gaps in Language Model Generations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, 6279–6296. doi:10.18653/v1/2024.naacl-long.348
- [32] Nikhil Sharma, Zheng Zhang, Daniel Lee, Namita Krishnan, Guang-Jie Ren, Ziang Xiao, and Yunyao Li. 2026. Feedback by Design: Understanding and Overcoming User Feedback Barriers in Conversational Agents. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. doi:10.1145/3772318.3791600
- [33] Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li. 2021. An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 132–157. doi:10.1109/TASLP.2020.3038524
- [34] Sheng-Lun Wei, Yu-Ling Liao, Yen-Hua Chang, Hen-Hsen Huang, and Hsin-Hsi Chen. 2026. Bias in the Ear of the Listener: Assessing Sensitivity in Audio Language Models Across Linguistic, Demographic, and Positional Variations. *arXiv preprint arXiv:2602.01030* (2026).
- [35] Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The Dialog State Tracking Challenge. In *Proceedings of the SIGDIAL 2013 Conference*. Association for Computational Linguistics, Metz, France, 404–413. <https://aclanthology.org/W13-4065/>
- [36] Hongyan Xie, Haoxiang Su, Shuangyong Song, Hao Huang, Bo Zou, Kun Deng, Jianghua Lin, Zhihui Zhang, and Xiaodong He. 2022. Correctable-DST: Mitigating Historical Context Mismatch between Training and Inference for Improved Dialogue State Tracking. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 876–889. doi:10.18653/v1/2022.emnlp-main.56
- [37] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). *The Rainbow Passage which the speakers read out can be found in the International Dialects of English Archive*: (<http://web.ku.edu/~idea/readings/rainbow.htm>). (2019).
- [38] Yu-Wei Yang, Chieh Hsu, Hsin-Chien Tung, Hong-Han Shuai, and Yung-Ju Chang. 2021. Tell Me When Users Leave: Predicting Users’ Abandonment of a Task-Oriented Chatbot Service Using Explainable Deep Learning. In *Proceedings of the 3rd Conference on Conversational User Interfaces*. Association for Computing Machinery, Article 35. doi:10.1145/3469595.3469630
- [39] Qixi Zheng, Yuxiang Zhao, Tianrui Wang, Wenxi Chen, Kele Xu, Yikang Li, Qinyuan Chen, Xipeng Qiu, Kai Yu, and Xie Chen. 2026. X-VC: Zero-shot Streaming Voice Conversion in Codec Space. arXiv:2604.12456 [eess.AS] doi:10.48550/arXiv.2604.12456

A Reproduction-Level Procedures

This appendix documents the procedures supporting Sections 4 and 5.

A.1 How We Decided That an Interaction Was Valid

Before semantic annotation, we automatically checked each attempt against the recorded session data and applied the inclusion rule in Section 4.4.

The export identifies every failed check. The nine attempts excluded for technical failures failed one or more checks in the following three groups.

- *The assigned voice route ran as scheduled* — `xvc_route_executed`, `input_route_sequence`, `transmitted_route_sequence`, `route_activation_lag`.
- *The required recordings were available* — `stream_start`, `stream_stop`, `event_stream_present`, and `capture_time_base`, the last of which checks that the recording clock is consistent with the frozen budget.
- *No system failure prevented a meaningful interaction* — `browser_artifacts_present`, `proxy_artifacts_present`, `end_reason`.

An attempt could fail more than one check. The 15 attempts excluded for incomplete validation had no recorded failed check because their validation had not finished when data collection closed.

For every session, we retained route regions, crosswalks between the capture and proxy clocks, and SHA-256 hashes of the stored audio. These records link each annotated interaction and quality score to its source audio.

We revised one validation rule before semantic annotation. When the end-of-interaction artifact check failed but an independent check showed continuous audio delivery, we recorded a warning rather than excluding the interaction. We applied this revision wherever that combination occurred; at the time, it added one interaction to the included set.

One participant’s four included interactions exceeded the frozen capture-gap budgets. Their recordings lost 1.2–1.5 s of audio per interaction in frame-length gaps distributed throughout the recording. The reconstruction rule for end-of-interaction loss did not apply because the loss was not confined to the end. Analyses that enforce those budgets exclude these interactions, but none of those analyses appears in this paper; we retained their human-reviewed and self-report measures here.

A.2 Model Estimation and Checks

We adjusted for scenario because condition-by-scenario cells contained 4–8 interactions and mean repair counts ranged from 2.67 to 4.62 (Section 5.4). The prespecified dispersion test did not indicate overdispersion ($p = 0.140$), so we used a Poisson repair model; negative-binomial models were available for overdispersed counts. We did not use duration as an offset because repair may affect duration. Because participant variance approached zero in the unit-level models, we also checked profile-likelihood intervals; they agreed with the Wald intervals. We fitted the models by maximum likelihood with a Laplace approximation in R [27], using `glmmTMB` [7] and `ordinal` [10].

The records use two timing systems: recorded duration uses scenario-step wall time, while the five-minute cap uses interaction time.

A.3 Annotation Pipeline

Transcripts came from two sources with different properties. Assistant text came from the model’s token stream and therefore preserved its disfluencies; participant text came from automatic speech recognition of the unmodified recording.

The LLM annotator received one interaction at a time and returned schema-constrained output containing unit-level labels, repair events and categories, task outcome, evidence-turn identifiers, confidence values, and review flags. It identified the participant’s first request for a final summary and selected the first assistant response to that request. It could select an earlier assistant turn only when the turn clearly provided a spontaneous final account of the recorded information, decision, and next step. If neither event occurred, it marked final-account accuracy as missing.

We assigned each repair move to exactly one prespecified category. When more than one category applied, we used this order: restart after cut-off, floor recovery, explicit correction, repeated confirmation after trouble, clarification, reformulation, and repetition. We did not treat ordinary elaboration, the first delivery of a critical-information unit, or the first protocol-required final-readback request as repairs. We classified escalation and withdrawal as separate outcomes rather than repair categories.

The schema enforces the distinction between zero and missing rather than leaving it to the human reviewer. A value of 0 means that the criterion failed in an observable opportunity; missing means that the stage was unobservable or inapplicable. When the transmitted recording showed that a unit was incomplete, we assigned no downstream indicator labels. When transmission completeness could not be checked, we admitted the unit only if the unmodified recording showed complete production.

The transmission rule held for all six units known not to have been transmitted: none received an indicator label. Among the 152 units with an observed acknowledgement label, 146 were transmitted completely and five had uncheckable transmission but entered the analysis through the unmodified-production proxy.

A.4 Voice-Conversion Quality Scoring

We scored conversion quality after fixing route regions from the recorded route events, so each region corresponded to a period when conversion was active. Within each region, we concatenated RMS-detected participant speech, applied a stability guard after the route change, and added symmetric boundary padding to exclude transition frames. WER compared recognitions of the converted region and its time-matched unmodified source; each row records the provenance of that reference. Speaker similarity compared the converted speech with the participant’s fixed target excerpt, and UTMOS measured the converted speech.

Every scored row records the region’s offsets on the proxy input clock, the number and duration of speech intervals scored, and SHA-256 hashes of the derived audio files, linking each score to its source audio. We retained failed or partial scoring as an explicit

Table 6: Primary estimates and sensitivity analyses. “Technical” excludes participants with any technically invalid scenario; “Quality” excludes converted-speech interactions with WER > 0.5; and “Unit intercept” adds the prespecified unit-level random intercept. Repair effects are ratios of expected counts; all other effects are odds ratios. Brackets show 95% confidence intervals.

Endpoint		Primary	Technical	Quality	Unit intercept
Interactions		88	60	75	88
Unit-by-indicator observations		584	400	490	584
H1a	Post-checkpoint repair	1.09 [0.76, 1.56]	1.50 [0.90, 2.51]	1.33 [0.87, 2.05]	1.09 [0.76, 1.56]
	Indicator attainment	1.10 [0.52, 2.32]	1.14 [0.52, 2.50]	1.93 [0.84, 4.43]	1.10 [0.51, 2.35]
H1b	Post-checkpoint repair	1.64 [1.11, 2.41]	1.75 [1.09, 2.81]	1.68 [1.12, 2.51]	1.64 [1.11, 2.41]
	Indicator attainment	0.56 [0.27, 1.17]	0.39 [0.18, 0.87]	0.61 [0.30, 1.24]	0.54 [0.25, 1.16]
H2a	Trust, per point	1.12 [0.98, 1.29]	1.00 [0.84, 1.19]	1.18 [1.02, 1.35]	1.12 [0.98, 1.29]
	Outcome confidence, per point	1.17 [1.03, 1.33]	0.99 [0.85, 1.16]	1.19 [1.05, 1.34]	1.17 [1.03, 1.33]
H2b	Effort, per repair move	1.16 [0.95, 1.42]	1.25 [0.97, 1.60]	1.22 [0.98, 1.52]	1.16 [0.95, 1.42]
	Frustration, per repair move	1.32 [1.05, 1.65]	1.39 [1.02, 1.88]	1.51 [1.16, 1.97]	1.32 [1.05, 1.65]
Secondary	Post-checkpoint repair	1.50 [1.02, 2.22]	0.97 [0.57, 1.64]	1.40 [0.92, 2.13]	1.50 [1.02, 2.22]
	Indicator attainment	0.63 [0.30, 1.30]	0.35 [0.16, 0.78]	0.58 [0.29, 1.17]	0.62 [0.29, 1.30]
Smallest adjusted hypothesis-level p		0.383	0.089	0.093	0.418

row with status and error fields and empty metric cells; we reported unavailable scores as missing rather than zero.

B System Prompts and Participant Scenario Cards

B.1 Shared Scenario Architecture

Each scenario contains a concise system prompt and a participant-facing scenario card. The system prompt gives the model its institutional role, the existing record, and its decision authority. Information needed to challenge that record appears only on the participant card and must therefore reach the model through the participant’s speech.

The full card is visible before the interaction and remains visible throughout the conversation. Cards use the same hierarchy: **Your situation** gives the essential task facts, **Your aim** states the intended outcome, **Start here** offers an optional opening, **Details you can use if asked** provides standardized supporting information, and **Before ending** requests a final account from the assistant. Supporting details help participants answer likely follow-up questions but are not additional critical units in the analysis.

Participants do not need to communicate every fact at once. They may use the suggested opening or their own words, introduce details when they become relevant, and clarify, correct, persuade, or interrupt the assistant when needed. If the assistant asks about something not on the card, participants say they do not know or do not remember rather than inventing new details. They end the call once they reach their aim or believe they cannot make further progress.

B.2 Practice Scenario: Fix a Restaurant Booking

B.2.1 System Prompt.

You are a cautious restaurant booking assistant. Ask for a reference number and accept 123. The current booking is for two people at 6 PM; keep it unchanged initially. Let the customer explain the problem. Ask for the correct number of guests and reservation time, clarify anything uncertain, and use confirmed details to update the booking. When asked, summarize the number of guests and time now recorded.

B.2.2 Scenario Card. Practice interaction: Fix a restaurant booking

Your situation. Your restaurant booking is recorded as:

- two people; and
- 6 PM.

You actually booked a table for four people at 7 PM.

Your aim. Have the booking corrected to four people at 7 PM.

Start here. You may begin with:

“I’m calling because some details in my restaurant reservation are incorrect.”

Details you can use if asked.

- Reference number: **123**.
- Booking name: **Alex Taylor**.
- Booking day: **this Saturday**.
- Use only fictional details.

Before ending. Ask the assistant:

“Please tell me the number of people and reservation time now recorded.”

Listen for the number of people and reservation time now recorded.

B.3 S1: Reopen a Declined Warranty Request

B.3.1 System Prompt.

You are a cautious warranty case handler. Ask for a reference number and accept 123. The request is declined because the paper receipt is missing, and no purchase evidence or product fault has been verified. Keep the decline initially. Before deciding, ask one question at a time until you understand what the purchase evidence shows, exactly when and how the product failed, what troubleshooting was attempted, and whether misuse occurred. Continue until these facts are clear, then make the strongest justified case decision yourself. You have authority to approve replacement headphones. Approve a replacement only if the customer has specifically described both the purchase evidence and how the headphones failed; otherwise keep the decline or open a review only. When asked, summarize the evidence, decision, and next step.

B.3.2 Scenario Card. **Scenario: Reopen a declined warranty request**

Your situation. Your warranty request for SoundsGood wireless headphones was declined because you do not have a paper receipt. However:

- your card statement shows a **\$120 purchase at TechStore two weeks ago**;
- the headphones charged normally the first two times, but on the third attempt the charging light stayed off;
- another cable and another power outlet did not solve the problem; and
- the headphones did not overheat and were not dropped or exposed to water.

The card statement names the store but does not identify the product.

Your aim. In order of importance:

- (1) replacement headphones approved;
- (2) a warranty review opened; and
- (3) the purchase evidence and product problem recorded.

Before asking for a replacement, make sure the assistant understands both the card-statement evidence and exactly how charging failed.

Start here. You may begin with:

“I’m calling about a declined warranty request, and I would like the case reviewed.”

Details you can use if asked.

- Reference number: **123**.
- You no longer have the receipt or packaging.

- You can submit the card statement through a secure process.
- Do not provide real card details.

Before ending. Ask the assistant:

“Please tell me what purchase evidence and product problem are now recorded, whether a review was opened, whether a replacement was approved, and what happens next.”

B.4 S2: Challenge a Wrong-Address Delivery

B.4.1 System Prompt.

You are a cautious food-delivery case handler. Ask for a reference number and accept 123. The current record says the customer entered building 18 as the delivery address. The courier delivered the order to building 18, and the delivery photo shows a red door matching that location. The order is therefore marked delivered. Explain this record to the customer and keep the decision initially. Let the customer challenge it. Ask one question at a time about what their order confirmation shows, their actual building and door, and whether they received the food. Require concrete details before reconsidering. Then make the strongest justified decision yourself. You have authority to fully resolve the dispute with a refund and replacement meal. When asked, summarize the record, decision, and next step.

B.4.2 Scenario Card. **Scenario: Challenge a wrong-address delivery**

Your situation. You paid **\$25** for a meal from Green Bowl but never received it.

The delivery platform says:

- you entered **building 18**;
- the courier delivered the order to building 18;
- the delivery photo shows a **red door**; and
- the delivered outcome and denied refund are therefore correct.

That record is wrong. Your order confirmation and confirmation email both show **building 16**, your building has a **blue door**, and nobody at your address received the food.

Your aim. In order of importance:

- (1) a refund issued and a replacement meal arranged;
- (2) the delivery record corrected to show the order went to the wrong address; and
- (3) the wrong-address dispute recorded.

Start here. You may begin with:

“I’m calling because an order I never received is marked as successfully delivered.”

Details you can use if asked.

- Reference number: **123**.
- The order was placed at approximately 7 PM.
- You can submit a screenshot showing building 16.

- You checked your entrance and nobody at your address accepted the order.
- Use only fictional details.

Before ending. Ask the assistant:

“Please tell me the delivery address and outcome now recorded, whether a refund and replacement meal were arranged, and what happens next.”

B.5 S3: Ask for a Grade Review

B.5.1 System Prompt.

You are a cautious university grade-review officer. Ask for a reference number and accept 123. The current record is Fail because the feedback says no source comparison or evidence limitation was identified. The grading criterion is satisfied when an assignment clearly compares two sources and explains how a limitation constrains its conclusions. Keep Fail initially. Ask one question at a time for the source names, the comparison finding, the limitation, and why it matters. Acknowledge concrete answers and ask a targeted follow-up for anything missing. After everything is clear, make the strongest justified decision yourself. You have authority to refer formal reassessment or change the grade to Pass. Decide on this call yourself; do not defer the decision to another person or process. Change the grade only if the student has specifically stated the comparison finding and how the limitation weakens the conclusions; otherwise refer reassessment or keep Fail. When asked, summarize the grounds, decision, and next step.

B.5.2 Scenario Card. **Scenario: Ask for a grade review**

Your situation. You received a Fail, but you believe your assignment was graded incorrectly. The grading criteria required you to compare two sources and explain a limitation in the evidence. Your assignment did both:

- it compared student interviews with university attendance records;
- most interview reports agreed with the attendance records, but two did not; and
- only eight volunteers from a class of 120 were interviewed, so this group might not represent the whole class.

Your aim. In order of importance:

- (1) the recorded grade changed from Fail to Pass;
- (2) the assignment referred for formal reassessment; and
- (3) both grading grounds recorded.

Start here. You may begin with:

“I’m calling because I disagree with the grade recorded for my assignment.”

Details you can use if asked.

- Reference number: 123.

- Eight students volunteered; six reported attending most classes.
- The attendance records supported four reports but not the other two.
- The comparison was in the findings section.
- The small volunteer group was discussed in the limitations section.
- The submitted assignment is already available in the university system.
- If asked why the limitation matters: a small volunteer group may not represent the whole class, so conclusions drawn from the comparison are less certain.

If asked to provide the assignment, say:

“The submitted assignment is already on file and can be checked during reassessment.”

Before ending. Ask the assistant:

“Please tell me which comparison and limitation are now recorded, whether the assignment was referred for reassessment, what grade is now recorded, and what happens next.”

B.6 S4: Correct a Wrong Accident Report

B.6.1 System Prompt.

You are a cautious accident-insurance handler. Ask for a reference number and accept 123. The claim and accident report say the claimant lost control of the bicycle and fell without another person involved; keep this record initially. Let the caller explain any dispute. Ask about cause, sequence, evidence, and treatment. Use confirmed details to read back one proposed statement. Decide separately whether to replace it and approve reimbursement; otherwise retain the record or decline. Decide both on this call; do not defer the decision to a later review. Replace the statement and approve reimbursement only for details the claimant has specifically confirmed. On request, summarize the active statement, reimbursement decision, and next step.

B.6.2 Scenario Card. **Scenario: Correct a wrong accident report**

Your situation. You hurt your ankle in a bicycle accident and asked the insurance company to pay your medical-treatment costs.

The insurance report says:

“You lost control of your bicycle and fell. Nobody else was involved.”

That is wrong:

- you were cycling straight in a bicycle lane;
- another cyclist passed you from behind and hit your **left handlebar**;
- the hit made you lose your balance and fall; and
- you hurt your ankle and received treatment that day.

Your aim. All of the following:

- the original statement removed and replaced with the correct account; and
- your medical-treatment costs reimbursed.

Make sure the assistant records that the other cyclist hit your left handlebar before you lost your balance and fell.

Start here. You may begin with:

“I’m calling because the accident description in my bicycle-injury claim is wrong.”

Details you can use if asked.

- Reference number: 123.
- A pedestrian saw the collision and can provide a statement.
- Your clinic record says that the injury followed a bicycle collision.
- You have the treatment invoice.
- The other cyclist did not stop, and you do not know their identity.
- There is no video or police report.
- Use only fictional details.

Before ending. Ask the assistant:

“Please read back the accident statement now on file, tell me whether my treatment costs will be paid, and explain what happens next.”

Listen for whether the original statement was removed, the accident description now recorded, the reimbursement decision, and the next step.

C Questionnaire Instrument and Participant-Facing Study Flow

This appendix presents the participant-facing wording and response formats implemented in the fielded study configuration. A separate participant-information document was supplied during recruitment before consent and is reflected in the consent and data-handling wording below.

C.1 Invitation Code, Welcome, and Eligibility

Before viewing study materials, participants enter a pre-generated pseudonymous invitation code. The invitation code is not generated from a name, institutional account, email address, or other direct identifier. At entry, it does not yet carry a target-voice or counterbalancing assignment.

Welcome

Thank you for your interest in this study. The study examines spoken interaction with an AI assistant. If you take part, you will first complete one short practice conversation, followed by four short task-based conversations and questions about your experience.

Your speech will be recorded. Depending on the interaction, your transmitted speech may remain natural, be digitally voice-converted, or change between natural and converted speech. You will not hear the converted version of your voice during the live interactions. After the four interactions, you will hear two short excerpts of your converted speech.

Participation is voluntary. The study is expected to take approximately 20–30 minutes.

Eligibility

Are you 18 years old or older?

- Yes
- No

C.2 Participant Information and Informed Consent

Project: Voice-Conditioned Grounding and Repair Study

Please read the participant information before deciding whether to take part. By consenting, you confirm that:

- You have received information about the study and have had the opportunity to ask questions. You may keep a copy of the written information.
- You consent to participate in this project.
- You consent to your personal data being handled within this project as described in the participant information.
- You understand that your speech will be recorded and may be digitally voice-converted or change between natural and converted speech during an interaction. You will not hear the converted version live, but you will hear two short excerpts of your converted speech afterward.

Your data rights

Study data that forms the basis of the research results, including audio recordings and questionnaire responses, will be stored securely by the host institution for at least 10 years after the project ends, in accordance with the institution’s archiving rules. Access will be restricted to authorised people.

Under EU data-protection rules and national supplementary legislation, you may have the right to:

- withdraw your consent without affecting processing that took place before withdrawal;
- request access to or correction or deletion of your personal data; and
- request that processing of your personal data be restricted.

These rights may be limited in some circumstances, including by confidentiality and archiving requirements. To exercise these rights, contact the researcher responsible for the project using the contact details in the participant information, or contact the host institution’s data protection officer at the address given there. You may also complain to the national data protection authority.

Consent confirmation

I have read and understood the participant information and the consent statements above. I understand that participation is voluntary and that I may stop participating without giving a reason.

Do you consent to participate?

- I consent and wish to continue.
- I do not consent.

C.3 Background Questionnaire and Deferred Assignment

The following questions are used to describe the study sample. You may select “Prefer not to answer” where that option is provided.

Age range

What is your age range?

- 18–24

- 25–34
- 35–44
- 45–54
- 55–64
- 65 or older
- Prefer not to answer

Language background

What language or languages did you first learn to speak?

[Optional free-text response]

What language or languages do you currently speak most often?

[Optional free-text response]

English speaking and listening

How would you rate your ability to express yourself in English during a task-oriented spoken conversation?

1 = I would have great difficulty expressing the necessary information; 4 = I could express most of the necessary information with some difficulty; 7 = I could express the necessary information easily and precisely.

Response: 1–7; Prefer not to answer.

How would you rate your ability to understand English during a task-oriented spoken conversation?

1 = I would have great difficulty understanding the conversation; 4 = I could understand most of the conversation with some difficulty; 7 = I could understand the conversation easily and accurately.

Response: 1–7; Prefer not to answer.

Prior experience with spoken AI

How often have you used spoken conversational AI or a voice mode in a generative AI system? Examples include ChatGPT Voice, Gemini Live, Claude voice mode, and similar conversational AI assistants.

- Never
- Once or twice
- A few times
- Less than monthly
- Monthly
- Weekly
- Daily or almost daily

Before today, had you heard your own voice changed using voice conversion or similar software?

- Yes
- No
- Not sure
- Prefer not to answer

Gender

How do you describe your gender?

- Woman
- Man
- Non-binary
- Prefer not to answer
- Self-describe: [Free-text response]

C.4 Instructions, Equipment Check, and Practice

Required equipment

This study uses full-duplex spoken interaction, which means that you and the assistant may speak at the same time. To take part, you must use:

- a working microphone; and
- headphones, earphones, or in-ear monitors.

A headset with a built-in microphone is acceptable. You may also use a separate microphone together with headphones or earphones.

Do not use speakers. Sound from speakers may be picked up by the microphone and interfere with the assistant’s ability to detect speech, interruptions, and overlapping talk.

Please also make sure that you are in a quiet environment with minimal background noise.

Audio and microphone check

Select **Start audio check**, say “Hello, can you hear me?”, and wait for a short response. The microphone-level indicator should move while you speak. After the assistant responds, select **Yes, I heard the assistant clearly**.

Equipment confirmation

After the automated check passes or an override is recorded, confirm:

I am using a working microphone.

I am using headphones, earphones, or in-ear monitors.

Audio is not being played through device speakers.

My browser or device has permission to access the microphone.

I am in a quiet environment with minimal background noise.

Interaction instructions

During each interaction:

- The complete scenario card will remain visible.
- Try to achieve the goal on the card using your own words.
- You may repeat, rephrase, correct, or interrupt the assistant.
- You may speak while the assistant is speaking when you need to interrupt or correct it.
- You may end the interaction when you believe that the goal has been achieved or that no further progress is possible.
- Before the call can start, you must select “I have read the scenario and I am ready to begin.”
- Each analytical interaction will end automatically after five minutes.
- Do not provide real personal information. Use only the reference details shown on the scenario card.

Practice interaction

You will now complete a short practice conversation to become familiar with the study tasks and controls. The practice recording is retained for technical checks but is not included in the main analysis. Read the scenario carefully, then continue when you are ready. Use the practice to become familiar with speaking to the assistant, interrupting it while it is speaking, and ending the interaction.

After the practice interaction, confirm:

- (1) Was the assistant audio clear enough to continue?
 - Yes
 - No
- (2) Did the system detect your speech correctly?
 - Yes

- No
- (3) Were you able to interrupt the assistant while it was speaking?
 - Yes
 - I did not try
 - No
- (4) Were you able to end the interaction using the interface?
 - Yes
 - No
- (5) Did any other technical problem occur?
 - No
 - Yes

If “Yes,” briefly describe the technical problem.

[Required free-text response when shown]

Main study introduction

The practice interaction is complete. You will now begin the four main study conversations. Each conversation presents a different situation and aim. Read each scenario carefully and speak naturally using your own words. Continue when you are ready to begin.

C.5 Repeated Post-Scenario Questionnaire

Please answer the following questions based only on the interaction you have just completed.

For Items 1–8, use the following scale:

1 = Strongly disagree; 4 = Neither agree nor disagree; 7 = Strongly agree.

- (1) Reaching my goal in this interaction required a lot of effort.
- (2) I felt frustrated while trying to reach my goal.
- (3) I was able to finish communicating the information that was important.
- (4) When I needed to interrupt or correct the assistant, I was able to do so.
- (5) The assistant took into account that [insert the scenario-specific decisive information below].
- (6) The assistant’s final response accurately reflected the information I provided.
- (7) I trusted the assistant to use the information I provided correctly.
- (8) At the end of the interaction, I was confident about what had been recorded, decided, or would happen next.

Scenario-specific wording for Item 5

- **Warranty:** The assistant took into account that I had a card transaction showing the purchase and that the item stopped working after two uses.
- **Food delivery:** The assistant took into account that the delivery evidence showed the order was left at the wrong address.
- **Assessment record:** The assistant took into account that I compared student interviews with university attendance records and identified the small interview group as a limitation.
- **Incident statement:** The assistant took into account that another cyclist hit my left handlebar before I lost my balance and fell.

Perceived misunderstanding, repair, and outcome

- (9) Did the assistant misunderstand or incorrectly represent anything you said?
 - Yes
 - No
 - Not sure
- (10) Was the misunderstanding resolved?
 - Yes, completely
 - Yes, partly
 - No
 - Not sure
- (11) Did the assistant say that it had updated, recorded, accepted, or corrected something, but later act or speak as though it had not?
 - Yes
 - No
 - Not sure
 - The assistant did not make such a claim
- (12) Which option best describes the outcome of this interaction?
 - I fully achieved the goal on the scenario card.
 - I partly achieved the goal.
 - I did not achieve the goal.
 - I am not sure what the final outcome was.
- (13) At any point, did you want to end the interaction before reaching your goal?
 - No
 - Yes, briefly
 - Yes, strongly
 - I ended the interaction before reaching my goal

Reason for ending early

Why did you end the interaction before reaching the goal?

- I did not think I could make further progress.
- I was unable to communicate effectively.
- The assistant was not helpful enough.
- A technical problem occurred.
- Other: *[Free-text response]*

Optional comment

Briefly describe any misunderstanding, missed information, or update that did not persist.

[Optional free-text response]

Target completion time: approximately 2 minutes, with up to 3 minutes when branching questions are shown.

C.6 Post-Session Questionnaire Before Playback

You have now completed all four spoken interactions. Please answer the following questions before listening to any recording.

Awareness of speech-processing changes

Did you notice or suspect that the way your speech was processed changed during any interaction?

- Yes
- No
- Not sure

How confident are you in that answer?

1 = Not at all confident; 7 = Extremely confident.

Response: 1–7.

If you answered “Yes” or “Not sure,” in which scenario or scenarios did you suspect a change? Select all that apply.

- Reopen a declined warranty request
- Challenge a wrong-address delivery
- Ask for a grade review
- Correct a wrong accident report
- I do not remember

What cues influenced your answer?

[Optional free-text response]

Speaking adaptation

Did you intentionally change how you spoke during any of the interactions?

- Yes
- No
- Not sure

If yes, how did you change your speech? Select all that apply.

- I spoke more slowly or carefully.
- I spoke more loudly.
- I used shorter or simpler sentences.
- I repeated or rephrased information.
- I used more precise wording.
- I became more direct or assertive.
- I changed when I started speaking or how often I interrupted.
- Other: *[Free-text response]*

In which scenario or scenarios did you intentionally change how you spoke? Select all that apply.

- Reopen a declined warranty request
- Challenge a wrong-address delivery
- Ask for a grade review
- Correct a wrong accident report
- I do not remember

What caused you to change how you spoke?

[Optional free-text response]

Overall evaluation of the live interactions

For Items 1–3, use the seven-point agreement scale.

- (1) Overall, I trusted the assistant to handle my information correctly.
- (2) Overall, I would be willing to use a spoken AI assistant for a similar institutional task.
- (3) Overall, interacting with the assistant was pleasant.

Reflection on the live interactions

Briefly describe the hardest moment in getting the assistant to understand or act on what you said.

[Optional free-text response]

What evidence did you use to decide whether the assistant had actually updated, accepted, or retained the information you provided?

[Optional free-text response]

Did you feel that the assistant treated or responded to you differently across the interactions? Please explain.

[Optional free-text response]

C.7 Post-Session Playback Questionnaire

You will now hear two excerpts, one shorter and one longer, from your conversations in the study. Your voice is converted throughout these recordings. Each excerpt is a contiguous segment from the merged participant-and-assistant recording of the stable-converted interaction and may contain participant speech, assistant speech, silence, and overlap.

Please listen before answering. You must play each excerpt at least once and may play each up to three times.

Conversation recordings: [Play audio; each excerpt up to three plays]

For Items 1–6, use the following scale:

1 = Strongly disagree; 7 = Strongly agree.

- (1) The converted voice was easy to understand.
- (2) The converted voice sounded natural.
- (3) The converted voice felt like an acceptable representation of me in this interaction.
- (4) I felt comfortable hearing myself represented by the converted voice.
- (5) I would be willing to use this converted voice in a similar interaction.
- (6) I noticed audible distortion or other artifacts in the converted voice.

How different did the converted voice sound from your natural voice overall?

1 = Not at all different; 7 = Extremely different.

Response: 1–7.

Please describe any feature of the converted voice that particularly affected its intelligibility, naturalness, comfort, or acceptability.

[Optional free-text response]

After hearing the recording, in what situations, if any, would you choose to use voice conversion with an AI assistant?

[Optional free-text response]

C.8 Final Comment and Debriefing

Is there anything else you would like to tell us about the interactions, the converted voice, or the study experience?

[Optional free-text response]

Debriefing

Thank you for completing the study.

The study contained four speech-processing conditions: natural speech throughout, converted speech throughout, a change from natural to converted speech, and a change from converted to natural speech. The exact condition and change timing were not disclosed during the interactions because knowing them could have changed how participants spoke or evaluated the assistant.

The study examines whether speech processing and a change in processing affect conversational grounding, repair, turn-taking, task success, trust, and user experience. The study does not assume that voice conversion causes any observed difference; scenario, order, technical performance, and the timing of important information will also be considered in the analysis.

C.9 Measurement and Analysis Map

The questionnaire items are construct-specific and should not be combined into one overall score.

Repeated post-scenario measures

- Conversational effort: Item 1
- Frustration: Item 2
- Perceived ability to complete critical-information delivery: Item 3
- Perceived ability to interrupt and correct the assistant: Item 4
- Perceived use of the decisive scenario information: scenario-specific Item 5
- Perceived final retention or readback accuracy: Item 6
- Task-specific trust: Item 7
- Outcome confidence: Item 8
- Perceived misunderstanding and repair resolution: Items 9–10
- Perceived false update confirmation: Item 11
- Self-reported task outcome: Item 12
- Withdrawal pressure: Item 13 and the branching reason item

Post-session measures before playback

- Awareness or suspicion of processing changes
- Self-reported speaking adaptation
- Overall trust, willingness to use spoken AI, and pleasantness of the live interactions

- Qualitative critical incidents and evidence used to judge institutional updating

Playback-specific measures

- Converted-voice intelligibility and naturalness
- Acceptability as a representation of self and comfort
- Willingness to use the assigned converted voice
- Perceived artifacts
- Perceived overall difference from the participant's natural voice
- Optional qualitative reasons for acceptance or rejection

Protocol controls. Invitation codes contained no direct identifiers, and target-voice and counterbalancing assignments were made only after eligibility, consent, and the background questionnaire. Except for the predefined scenario-specific wording in Item 5, questionnaire wording, response anchors, and item order were held constant across scenarios, and each post-scenario questionnaire was completed before the next scenario was displayed. Post-session awareness and adaptation questions preceded playback and debriefing. Condition, checkpoint, route state, timeout, and technical-termination status were determined from system logs rather than participant recall. Internal condition labels were not shown before debriefing.