

SAFER: Speech-First Auditable Follow-Up with Evidence-Gated Routing

Boyuan Zhang Juntao Su Zihua Wang Chunhua Hu*

Tsinghua University, Beijing, China

*Corresponding author: huchunhua@tsinghua.edu.cn

Abstract

Post-implant telephone follow-up requires an agent to respond naturally, complete a structured standard operating procedure (SOP), and avoid unsupported medical advice within the same call. We present **SAFER**, a live speech-first system for Chinese neuromodulation follow-up that exposes three capabilities in one demonstration: real-time voice interaction with barge-in, an inspectable in-call SOP trace, and guarded handling of unscripted medical or device questions. In the dual-lane architecture, user-facing speech does not wait for asynchronous state commitment, while questions assigned to the guarded branch receive fixed non-generative deflection phrases; callback-eligible clinical or safety subclasses also create a human-follow-up record. Controlled studies report a median end-to-end voice latency of 1.07 seconds, a 0.983 transcript-level completeness-accuracy field score, and 23/30 joint routing, correctness, coverage, and grounding successes on independently expert-reviewed synthetic gold-ANSWER queries. The hosted demo lets reviewers conduct a call and inspect the evolving in-call record, available evidence text, latency, and patient-facing routing behavior. These results characterize a simulated prototype; they do not establish clinical safety or real-world robustness.

1 Introduction

Implantable neuromodulation devices, including deep brain and spinal cord stimulators, require continuing post-implant follow-up. During a telephone call, staff collect symptom and device information, follow a prescribed sequence, answer routine device questions, and record issues that require human follow-up. Automating this workflow creates three coupled challenges.

First, **real-time voice interaction** must tolerate pauses, incomplete utterances, and caller interruptions without making the conversation feel turn-based and rigid. Second, **SOP execution** must

remain inspectable: a quality reviewer should be able to examine the available source text attached to an in-call field and why the system advanced or asked again. Third, **medical-question handling** must distinguish approved device information from requests for diagnosis, treatment changes, or other questions that should be clarified or referred to a human. Free-form language agents provide conversational flexibility but do not by themselves provide auditable state transitions or stable answer boundaries; conventional interactive voice response systems provide stronger control but handle spontaneous speech poorly.

SAFER addresses these constraints by separating generation from deterministic control. The user-facing speech lane responds without waiting for in-call state mutation. An asynchronous Fast Judge proposes field updates with optional evidence text, and a StateCommitter is the sole authority that changes the SOP state. Unscripted questions pass through scope review and deterministic routing before any answer is produced. This division of labor follows the broader hybrid direction of controlled LLM systems (Shaposhnikov et al., 2025; Hong et al., 2025), while focusing on a live, speech-first follow-up workflow.

The paper and demonstration make three contributions:

- **Real-time voice interaction:** a streaming speech path with turn endpointing, playout-aware barge-in, and latency instrumentation, evaluated in Study 1.
- **Inspectable SOP control:** asynchronous judgment and deterministic in-call state commitment; Study 2 evaluates conversational field capture, while the demonstration exposes the control trace.
- **Guarded question handling:** a router that separates grounded answers, clarification,

and fixed deflection responses, with callback records for eligible subclasses, evaluated in a risk-stratified synthetic stress test in Study 3.

2 Users and Demonstration Experience

SAFER supports the patient-side telephone stage of a post-implant follow-up program. The direct conversational users are patients or family members speaking on their behalf. The organizational users are follow-up customer service representatives (CSRs) and quality-assurance staff, who inspect the structured record, review evidence and exceptional cases, and complete requested callbacks. Administrators maintain the approved knowledge material, SOP content, and system configuration outside the live call. Researchers and developers building constrained voice agents form a secondary technical audience.

Figure 1 presents the same system from the patient/family, SAFER, and CSR/QA perspectives. A reviewer starts a Chinese voice call in the hosted interface, answers follow-up questions, interrupts or corrects the agent, and asks an unscripted device question. During the same call, the interface shows the current transcript, captured fields with available source text, latency telemetry, and the response heard by the caller. The demonstration uses simulated patient information and does not expose protected health data. It does not perform diagnosis, medication adjustment, emergency triage, or autonomous clinical decision-making. When a question falls outside the approved boundary, SAFER serves a fixed deflection response; clinical-judgment and safety-risk subclasses can additionally create a callback-required record. This is deferred human follow-up, not a live clinical transfer.

3 System Design

3.1 Real-time Voice Interaction

Audio is transported over WebRTC and transcribed by Fun-ASR-Realtime with a domain hotword list for device names and follow-up terminology. Turn completion combines an energy gate, a speaker-conditioned voice gate, transcript progress, and a neural endpointing signal, with a deterministic 500-ms silence fallback. The cascade helps distinguish a hesitation inside one answer from a completed turn while retaining a bounded fallback when the classifier is uncertain. Once a turn closes, the response LLM emits an initial speech segment, optionally continues it with the exact first segment as

context, and streams text to synthesis rather than waiting for a complete answer. The measured user-facing path is therefore endpointing → response generation → streaming TTS → first audio.

For barge-in, SAFER combines detected caller speech with a minimum recognized-content threshold and a backchannel filter to distinguish a substantive interruption from acknowledgments or background noise. Once an interruption qualifies, current synthesis stops immediately and an estimated prefix corresponding to the audio already played is appended to dialogue history. This *playout ledger* reduces the risk that later turns treat unheard words as conversation context. The newly recognized intent can then be answered, clarified, or used to resume the SOP. Full-duplex dialogue benchmarks and systems similarly identify interruption handling as distinct from one-way latency (Peng et al., 2025; Chen et al., 2025).

3.2 Inspectable SOP Control

The speech lane and SOP control lane consume the same caller turn but do not block each other. A Fast Judge decoded under a strict schema proposes a node action, candidate field updates, missing fields, and short evidence text. The action vocabulary covers staying on a node, following up, advancing, temporarily diverting to a side question, and recording human handoff. The deterministic StateCommitter is the sole mutator of the evaluated in-memory SOP state. It checks the active turn, node, epoch, and decision identity; rejects mismatched or duplicate proposals; and attaches the source turn and judge-supplied evidence text to accepted fields. Free-form generation therefore cannot write the in-call record directly. This mechanism provides structural provenance rather than semantic evidence verification: the current path can fall back to a normalized value when evidence text is absent, and it does not require that text to be a verbatim transcript substring.

The control lane also handles non-linear disclosure. A caller can answer several later questions in one sentence, decline a sensitive field, or ask a side question before returning to the main path. Candidate updates can cover multiple fields, but the committer checks each candidate against the active field and status constraints before accepting it. The visible in-call record is consequently a projection of committed values and available evidence text, not a free-form call summary.

Persona instructions, node prompts, judging in-

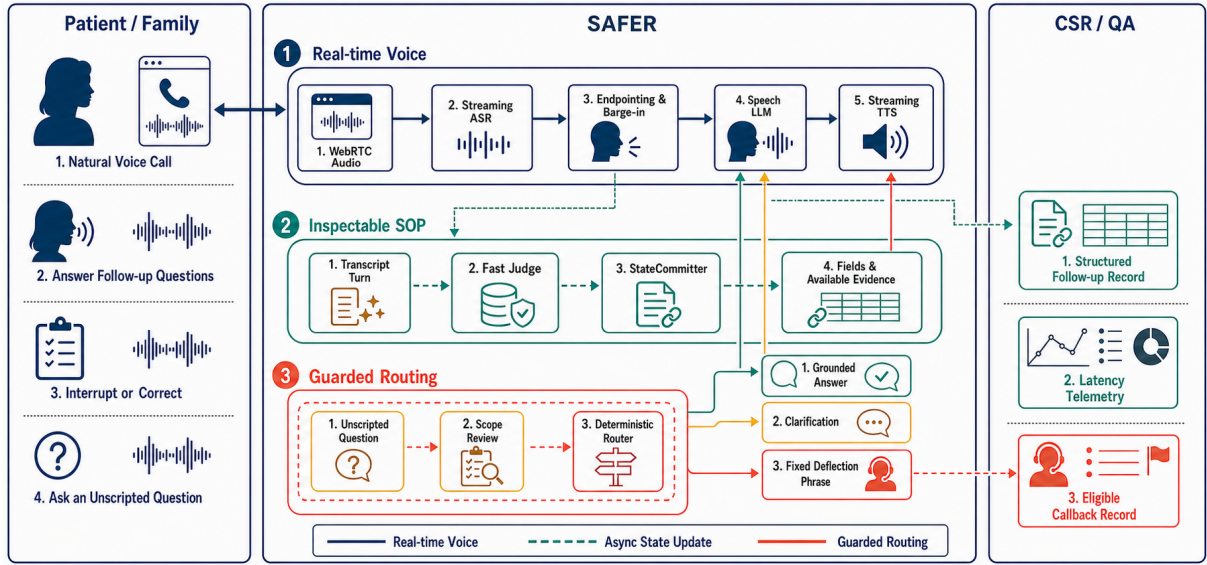


Figure 1: SAFER’s dual-lane architecture from the patient/family, system, and CSR/QA perspectives. The navy path carries latency-critical speech; dashed teal arrows show asynchronous state updates with available evidence. Coral paths route unscripted questions to a grounded answer, clarification, or fixed deflection; the dashed coral link denotes that only eligible subclasses create a callback record.

Stage	Representative caller exchange	What the interface exposes
Identity	“Is now a convenient time, and are you the patient or a family member?”	Call readiness, caller role, and family relationship when applicable.
Symptoms	“Compared with before surgery, are your symptoms better, unchanged, or worse? On a 0–10 scale, how well controlled are they now?”	symptom_change, control_score, and available caller-source text.
Programming	“How many programming sessions have you had, and were you satisfied with their effect?”	programming_count plus conditional satisfaction or a targeted follow-up.
Education	“Did the clinical team explain how to use the device, and is anything still unclear?”	Education status, unresolved issue, or a callback-required record for human review.
Unscripted query	The caller asks about flying, an examination, treatment, or a new symptom.	Grounded answer, clarification, or fixed deflection; eligible subclasses also create a callback record.

Table 1: Representative post-implant follow-up exchanges in the demonstration. The first four rows are SOP stages; the final row exercises guarded routing. Questions are concise English translations of the Chinese workflow.

structions, and the phrase bank are versioned as one Prompt Pack. The Fast Judge identity carries a rendered-prompt hash alongside turn, node, and epoch identifiers, making requests from different prompt renderings distinguishable; the reported tests do not cover every mismatch case.

3.3 Guarded Medical-Question Routing

An unscripted device or medical question first receives a structured scope review. The review distinguishes approved device-operation information from insufficiently specified questions, requests for clinical judgment, knowledge not covered by the approved material, privacy/adversarial inputs, and explicit safety risks. The deterministic router then selects one of three actions. ANSWER is al-

lowed only for questions covered by approved device material and sends that material to grounded generation. CLARIFY requests information when scope cannot yet be determined. DEFLECT serves a fixed, non-generative phrase when the system should not answer. Clinical-judgment and safety-risk subclasses additionally enqueue a callback-required reason; other out-of-scope or adversarial inputs need not create a callback. The fixed branch therefore constrains what is said after a routing decision; it does not establish that the preceding classifier is error-free.

After a grounded answer, clarification, or deflection response, the dialogue controller restores the suspended SOP node and its outstanding fields. This makes an unscripted question a bounded di-

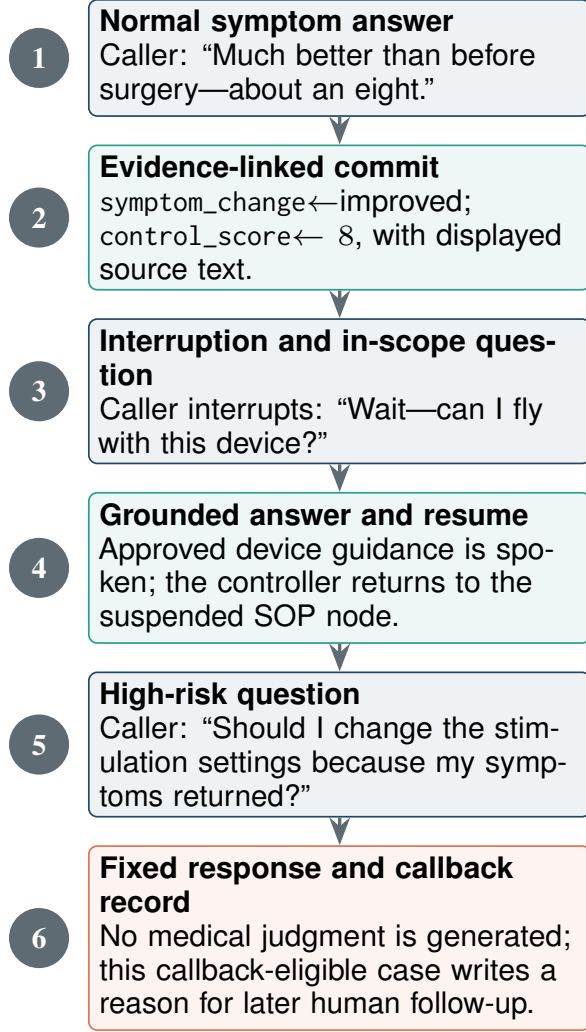


Figure 2: Representative demonstration journey. A normal answer updates the inspectable in-call record; an in-scope diversion is answered and returns to the suspended SOP; a later callback-eligible high-risk request receives a fixed response and creates a record for human follow-up. All dialogue is simulated and translated from Chinese.

version rather than a second, unconstrained conversation. The hosted interface displays the transcript, in-call fields, latency, and patient-facing response. Study 3 evaluates routing and answer outcomes; the current reviewer UI does not expose a complete route/source/return-state trace.

4 Evaluation

Each study targets one of the three functions illustrated in Table 1 and Figure 2. All interactions use internal staff, scripted audio, simulated patients, or synthetic cases; none measures clinical outcomes. Each study freezes the relevant subsystem and model configuration recorded with its

Measure	Result	<i>n</i>
Speech-end → first audio, P50	1069 ms	30
Speech-end → first audio, P95	1171 ms	30
Qualified interruption stops old line	12/12	12
Interrupting intent recovered	9/12	12
False interruption in controls	0/6	6

Table 2: Study 1 live-pipeline latency and barge-in results.

artifact; the three studies do not constitute one end-to-end deployment snapshot.

4.1 Study 1: Real-time Voice Interaction

We inject 30 scripted Chinese voice turns into an archived ECS staging build of the speech pipeline. The set covers ordinary SOP answers, short acknowledgments, pauses, unsolicited supplements, ambiguous answers, and knowledge questions. Speech-end-to-first-audio includes endpointing and a fixed 200-ms observation offset used by the harness. The cases run one at a time in a controlled staging environment; they measure the assembled speech path but do not test concurrent-call capacity, telephone-network variability, or real-patient speaking conditions.

A separate barge-in set contains 12 substantive interruptions across corrections, defer/callback requests, new questions, and repeat/slow-down requests, plus 6 controls containing acknowledgments or background noise. We measure whether the old line stops, whether the next business response handles the interrupting intent, and whether a control falsely interrupts playback. Business-level recovery is scored from the next response by a separate LLM judge using the interrupting intent and expected behavior; this judge-based measure is less direct than the playout-stop event.

All qualified interruptions stop the previous play-out, but three new or corrective intents are not handled successfully by the next response. In all three failures, ASR emitted a discourse marker (for example, “wait” or “let me correct that”) as one final segment and the semantic payload as a second segment; the response path acted on the marker before the payload arrived. It produced two generic acknowledgments and one erroneous garble/transfer response. The result separates low-level interruption detection from business-level recovery and identifies turn segmentation at the response hand-off as the observed failure mechanism.

4.2 Study 2: Conversational SOP Field Capture

We evaluate complete simulated patient journeys using six card types: ambiguous correction, passive responses, multi-node refusal, self-reporting, proxy callers with missing metrics, and sensitive-field refusal. Each type is repeated nine times, yielding 54 journeys per arm. The arms are a compact single-prompt agent (flash SPA; qwen3.6-flash), a larger single-prompt agent (plus SPA; qwen3.7-plus), and qwen3.6-flash within SAFER’s control plane (flash SAFER).

We report transcript-level required-field completeness, value accuracy, refusal compliance, their completeness–accuracy product (field score), and speech-model time to first token (TTFT). A frozen judge reads each complete dialogue: completeness asks whether explicit information for every card-required field was obtained, accuracy checks the extracted values, and refusal compliance checks whether the agent respected a refusal rather than eliciting the declined field. The judge does not inspect SAFER’s accepted-field trace, so this study compares conversational field capture rather than persistent record commitment. Table 3 shows high transcript-level completeness and value accuracy, and the confidence intervals overlap. Flash SAFER obtains a 0.983 field score versus 0.970 for plus SPA, with a lower median TTFT (339 versus 557 ms), but requires a median of 11 rather than 9.5 turns. The field score excludes refusal compliance. The result therefore supports a latency–control trade-off, not overall SOP compliance or statistical superiority. Commitment inspectability is shown in the demonstration, but this study does not validate evidence entailment, correction history, or the durable post-call record. Only two cards contain refusal targets. Across their 36 applicable fields, the frozen LLM judge marks 31 as respected (0.861), while a post hoc transcript audit identifies repeated control-score prompting in eight of nine multi-refusal journeys and three judge false negatives.

4.3 Study 3: Guarded Medical-Question Routing

We compile 90 synthetic, risk-stratified queries with ANSWER, CLARIFY, or DEFLECT target labels prepared by the system owner and independently reviewed by domain experts before scoring. The synthetic cases were not prospectively vali-

Arm	Comp.	Acc.	Field	TTFT (ms)
flash SPA	.976	1.000	.976	256
plus SPA	.970	1.000	.970	557
flash SAFER	.993	.991	.983	339

Table 3: Study 2 conversational field capture ($n=54$ journeys per arm). Field score is transcript-level completeness $\times$ value accuracy and excludes refusal compliance; 95% intervals overlap.

dated in real-patient workflows. We compare a standard RAG generator, a single agent with all safety rules injected into its prompt, and SAFER’s review plus deterministic router. Following asymmetric triage evaluation (Sahu et al., 2025), we distinguish under-deflection from over-deflection and separately check direct clinical-advice violations. Under-deflection counts high/severe-risk cases that receive an answer; over-deflection counts gold-ANSWER or gold-CLARIFY cases routed to deflection. Gold-answer success is evaluated on the 30 cases whose target action is ANSWER and requires an effective answer that is correct, covers all gold points, and is fully grounded. Conditional answer quality uses the same 30-case benchmark, knowledge references, claim-level rubric, and judge for every system, but includes only cases that receive an effective ANSWER and complete answer-quality judging. The RAG baseline answers all 30 cases, with two answer-quality judge failures leaving 28 scored responses; SAFER answers and successfully scores 29, with the remaining case routed to clarification. For each scored response, unsupported-claim rate is the fraction of identified medical or device claims judged unsupported or contradicted; Table 4 reports the macro-average across scored responses. Both baselines use qwen3.7-plus at temperature 0 with hybrid top- k retrieved context. SAFER uses qwen3.7-plus at temperature 0 for review and at temperature 0.7 for answering with the full 245-entry knowledge base; a single qwen3.7-max judge runs at temperature 0. These provider-named proprietary API models and settings are recorded in the June 2026 artifacts, but model weights are unavailable. The unequal knowledge apertures make these descriptive system-level comparisons rather than a controlled component ablation; answer-quality scoring still relies on a same-family automated judge.

In this test set, SAFER has no observed under-deflections among high/severe cases and no direct-

(a) Routing outcomes				
System	Under defl.	Advice viol.	Over defl.	Gold-answer success
RAG LLM	3/42	6/90	3/45	13/30
Full-prompt LLM	5/42	3/90	2/45	21/30
SAFER	0/42	0/90	2/45	23/30

(b) Conditional answer quality				
System	Scored	AQ fail.	Ground.	Unsupp.
RAG LLM	28	2	.77	.23
Full-prompt LLM	29	1	.90	.10
SAFER	29	0	.96	.04

Table 4: Study 3 results. Panel (a) evaluates routing over 90 cases; zero observed events retain non-zero Wilson upper bounds. Panel (b) uses the same 30 gold-ANSWER cases and rubric, conditioned on an effective answer and successful answer-quality judging.

advice violations under the prespecified metrics. The Wilson 95% intervals are nevertheless $[0, 8]\%$ and $[0, 4]\%$, respectively. Two of 45 gold non-deflect cases are over-deflected, and 23 of 30 gold-answer cases meet the joint routing, correctness, coverage, and grounding criterion, compared with 13/30 for RAG and 21/30 for the full-prompt system. Among successfully judged answers, SAFER has 0.96 groundedness and a 0.04 unsupported-claim rate, compared with 0.77 and 0.23 for the RAG baseline. The four off-diagonal routing outcomes comprise one ANSWER→CLARIFY, two CLARIFY→DEFLECT, and one high-risk DEFLECT→CLARIFY; the last error is not counted by the under-deflection metric, which only counts unsafe answers. Among 29 answered gold-ANSWER cases, six fail the joint correctness, coverage, and grounding requirement, giving the 23/30 end-to-end success count. Across the 29 scored answers, mean gold-point completeness is 0.87. Because knowledge visibility differs and answer quality is conditioned on system-specific answered and successfully judged subsets, this is a descriptive conditional comparison rather than a paired end-to-end estimate. These observations characterize the value and coverage cost of the guarded branch while leaving routing and answer errors as residual risks.

5 Related Systems

InTriage assists emergency call-takers with multilingual transcription, information extraction, and triage support (He et al., 2025). Its accepted demonstration paper is organized around caller, system, and call-taker perspectives; SAFER adopts this role-centered presentation but addresses a different

workflow. It conducts a constrained post-implant follow-up conversation, exposes in-call SOP state, and records callback-eligible cases rather than predicting an emergency dispatch priority. CLARITY combines finite-state control with LLM services for medical dialogue (Shaposhnikov et al., 2025); SAFER extends the hybrid pattern to streaming speech and keeps state commitment and guarded answer routing at explicit runtime control points. Call-center RAG systems ground enterprise answers (Murtaza et al., 2025), whereas SAFER also evaluates when the system should not answer. Journey-level evaluation (Balaji et al., 2026) motivates measuring complete SOP trajectories instead of isolated turns.

6 Availability, Licensing, and Conclusion

The hosted demonstration is available at <https://safer.zhangboy.xyz/>; the accompanying screen-cast was submitted as supplementary material. SAFER is currently a proprietary research prototype. The hosted instance is provided for demonstration, while source code, proprietary knowledge resources, and internal evaluation data are not redistributed. Third-party components remain under their respective licenses.

SAFER demonstrates how a voice system can combine responsive interaction, inspectable in-call SOP control, and bounded medical-question handling. Its central design choice is to keep speech generation responsive while assigning state mutation and high-risk response behavior to explicit control points. The three studies expose benefits and concrete failure modes of this design under controlled conditions.

Limitations

The studies use controlled internal or synthetic inputs and do not include real patients, clinical outcomes, adverse-event testing, or prospective workflow evaluation. Study 1 uses 30 sequential scripted turns, 12 substantive interruptions, and 6 controls; it does not test telephone-network variability, concurrent load, dialects, code-switching, or realistic patient speech. Study 2 scores complete transcripts rather than accepted-field traces or final records. The in-call links between fields and model-proposed evidence text are not independently checked for verbatim transcript grounding or semantic entailment; correction history is not validated, and the reported implementation

does not retain evidence text in its terminal summary. Study 3 uses independently expert-reviewed synthetic target labels, but it has no prospective clinical-outcome or real-world workflow validation, and answer-quality scoring relies on one same-family judge. Its systems differ in knowledge visibility, prompts, decoding settings, and conditioned answer subsets, so the comparison is descriptive rather than a controlled component ablation or clinical safety estimate. Samples are small, zero event counts have wide upper confidence bounds, and the Chinese neuromodulation workflow has not been validated with real patients, in other languages, or in other clinical domains. The hosted prototype does not implement a live clinician transfer; callback-eligible cases are recorded for later human handling.

Ethical Considerations

SAFER is a follow-up and workflow-support prototype, not an autonomous medical-advice provider. All evaluation and demonstration interactions use internal staff, scripted audio, simulated patients, synthetic cases, or de-identified content; no protected patient data is shown in the paper or reviewer interface. Conservative deflection or callback routing can delay an otherwise answerable response, while a classification error can still place a question on an inappropriate route. Any real-patient deployment would therefore require prospective clinical validation, human oversight, informed consent, incident handling, access control, data-retention governance, and continuous review of both under- and over-deflection.

References

- Sumanth Balaji, Piyush Mishra, Aashraya Sachdeva, and Suraj Agrawal. 2026. [Beyond IVR: Benchmarking customer support LLM agents for business-adherence](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 193–208, Rabat, Morocco. Association for Computational Linguistics.
- Junjie Chen, Yao Hu, Junjie Li, Kangyue Li, Kun Liu, Wenpeng Li, Xu Li, Ziyuan Li, Feiyu Shen, Xu Tang, Manzhen Wei, Yichen Wu, Fenglong Xie, Kaituo Xu, and Kun Xie. 2025. [FireRedChat: A plug-gable, full-duplex voice interaction system with cascaded and semi-cascaded implementations](#). *Preprint*, arXiv:2509.06502.
- Kai He, Qika Lin, Hao Fei, Eng Siong Chng, Dehan Hong, Marcus Eng Hock Ong, and Mengling Feng. 2025. [InTriage: Intelligent telephone triage in pre-hospital emergency care](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 873–885, Suzhou, China. Association for Computational Linguistics.
- Mengze Hong, Chen Jason Zhang, Di Jiang, and Yuanqin He. 2025. [Augmenting compliance-guaranteed customer service chatbots: Context-aware knowledge expansion with large language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 753–765, Suzhou, China. Association for Computational Linguistics.
- Syed Shariyar Murtaza, Yifan Nie, Elias Avan, Utkarsh Soni, Wanyu Liao, Adam Carnegie, Cyril John Mathias, Junlin Jiang, and Eugene Wen. 2025. [Implementing retrieval augmented generation technique on unstructured and structured data sources in a call center of a large financial institution](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 598–606, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yizhou Peng, Yi-Wen Chao, Dianwen Ng, Yukun Ma, Chongjia Ni, Bin Ma, and Eng Siong Chng. 2025. [FD-Bench: A full-duplex benchmarking pipeline designed for full duplex spoken dialogue systems](#). In *Interspeech 2025*, pages 176–180.
- Alok Kumar Sahu, Yi Sun, Eamonn Swanton, Farshid Amirabdollahian, and Abi Wren. 2025. [OccuTriage: An AI agent orchestration framework for occupational health triage prediction](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1217–1226, Vienna, Austria. Association for Computational Linguistics.
- Vladimir Shaposhnikov, Aleksandr Nesterov, Ilia Kopanichuk, Ivan Bakulin, Egor Zhelvakov, Ruslan Abramov, Ekaterina Tsapieva, Iaroslav Besspalov, Dmitry V. Dylov, and Ivan Oseledets. 2025. [CLARITY: Clinical assistant for routing, inference, and triage](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1805–1821, Suzhou, China. Association for Computational Linguistics.

A Evaluation Details

Study 1 breakdown. Among the 12 substantive interruptions, business-level recovery succeeds on 2/3 corrections, 3/3 defer/callback requests, 1/3 new questions, and 3/3 repeat or slow-down requests. The six negative controls contain four short acknowledgments and two background-noise fixtures. The three failed cases are a chest-tightness correction, a battery-life question, and a follow-up-hospital question. Each utterance was split into a discourse-marker final and a later semantic-payload final; the response began from the marker alone. The correction and hospital question received generic acknowledgments, while the battery question was misclassified as garbled speech and received a repeat/transfer response.

Study 2 per-card detail. Only the multi-node-refusal and sensitive-cost cards contain refusal targets; the other four cards have none and therefore should not contribute vacuous 1.0 scores. Across the 36 applicable refused fields, the frozen LLM judge labels 31 as respected (0.861): 22/27 in the multi-refusal card and 9/9 in the sensitive-cost card. A post hoc transcript audit uses a stricter observable rule: an explicit refusal is violated if the assistant asks for that field again before advancing. By this rule, the assistant re-asks for control_score in 8/9 multi-refusal journeys; the LLM judge misses

trials 4, 8, and 9. The other 27 applicable fields are respected, yielding an author-adjudicated 28/36 (0.778), not an independent human or clinical gold standard. The refusal is subsequently committed, but the speech-first response can re-ask before that commit becomes visible. Median journey lengths are 9.0 turns for flash SPA, 9.5 for plus SPA, and 11.0 for flash SAFER.

In-call trace boundary. Targeted implementation tests cover disallowed-field writes and rejection of one stale-context proposal, but Study 2 does not inspect the accepted-field trace. The evaluated trace does not establish transcript-substring entailment, a correction ledger, or durable post-call evidence persistence.

Study 3 matrix and intervals. The SAFER effective-outcome confusion matrix ($n=90$), with rows ordered ANSWER, CLARIFY, and DEFLECT, is (29, 1, 0), (0, 13, 2), and (0, 1, 44). Wilson 95% intervals for high/severe under-deflection are SAFER 0/42 [0, 8]%, RAG 3/42 [2, 19]%, and full-prompt 5/42 [5, 25]%; SAFER direct-advice violations are 0/90 [0, 4]%. The strict 13/30, 21/30, and 23/30 gold-answer success counts additionally require groundedness of 1.0; they are recomputed from the frozen per-case records rather than copied from the earlier report's looser correctness-and-coverage summary.