SpeakPatient: Improving End-to-End Spoken Virtual Patients with Patient-Level Conditioning

Dongliang Li[1], *Syaheerah Lebai Lutfi*[2,1,*], *Keng Hoon Gan*[1], *Muhammad Firdaus Akbar*[1]

[1]Universiti Sains Malaysia, Malaysia
[2]Sultan Qaboos University, Oman

pillerli123@gmail.com, s.lutfi@squ.edu.om, khgan@usm.my, firdaus.akbar@usm.my

Abstract

Medical history-taking is intrinsically speech-centric, yet most virtual-patient systems remain text-based or rely on cascaded ASR–LLM–TTS pipelines that add latency and error accumulation. End-to-end speech-to-speech (S2S) models enable more natural spoken interaction, but general backbones lack patient-role behavior, and physician-side adaptations do not yield a patient agent. Even after patient-role adaptation and preference alignment, replies can still drift on identity and disclosed case content over long consultations. This paper proposes SpeakPatient, an end-to-end spoken virtual patient on LLaMA-Omni2 for interactive history-taking. SpeakPatient combines (1) patient-role adaptation on about 100k patient-text turns, (2) preference alignment with about 5.1k chosen/rejected pairs, and (3) patient-level conditioning on about 29k OSCE-style turns—a persistent representation of who the patient is and what has been disclosed that guides spoken decoding—followed by spoken re-alignment. Evaluation uses matched interactive speech-in/speech-out rollouts with text-side simulation metrics and automatic speech quality measures. Experimental results show that SpeakPatient improves automatic simulation quality over a preference-aligned spoken-patient baseline, with additional gains on late identity and presenting-problem probes and competitive speech quality. Against a cascaded ASR–LLM–TTS reference, 0.5B SpeakPatient further delivers much lower first-audio latency (362.8 ms vs 5283.7 ms), underscoring the latency benefit of end-to-end S2S. Code is available at https://github.com/pillerli123/SpeakPatient. All claims here concern automatic simulation metrics, not clinical or educational validity.

Index Terms: spoken virtual patient, speech-to-speech, patient-level conditioning, preference optimization, SpeakPatient

1 Introduction

Medical education needs scalable ways for trainees to practice multi-turn history-taking: clinical teacher time is limited, standardized patient encounters are costly to schedule, and bedside practice is uneven across institutions [1, 2, 3]. Virtual patients help close this gap, including stations in the Objective Structured Clinical Examination (OSCE)—timed clinical scenarios in which candidates interact with a patient under structured scoring. Existing virtual-patient systems are often text-based [2, 1] or rely on cascaded ASR–LLM–TTS pipelines rather than end-to-end spoken interaction [4, 5]. End-to-end speech-to-speech (S2S) models make low-latency spoken interaction feasible [6, 7], but a usable spoken virtual patient requires more than real-time speech.

Modern S2S models can listen and reply [6], but they typically lack medical patient-role knowledge, and existing S2S adaptations such as SpeechMedAssist mainly fine-tune the backbone as a physician-side assistant rather than as a patient [8]. Even when a spoken patient is obtained, consistency can break down over long consultations. Early turns may disclose a name and a presenting problem; after many further questions, a late re-ask may change the name, rewrite onset or timeline, or deny facts already stated [9, 10]. Figure 1 illustrates this failure: a patient who initially reports a chronic cough with yellow-brown sputum for ten years may later deny the sputum color or shorten the onset.

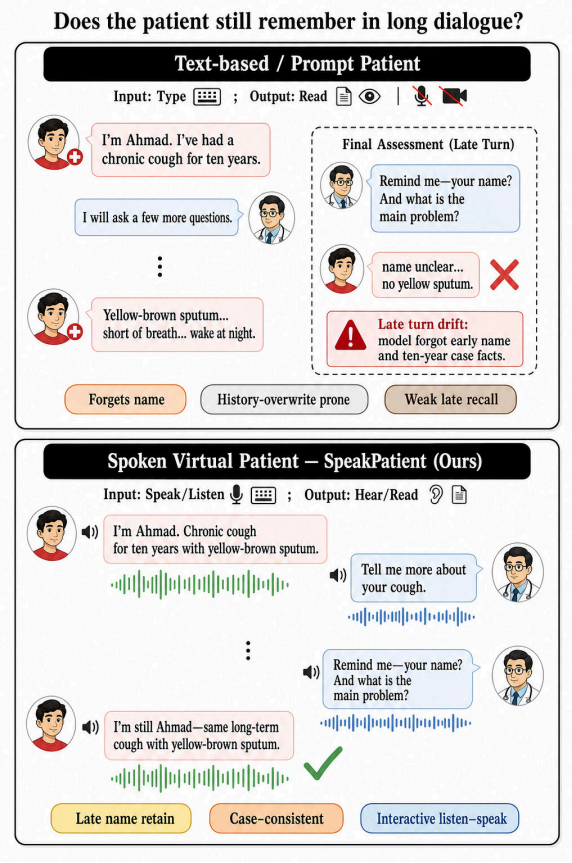

Figure 1: *Motivation for SpeakPatient on long spoken consultations. Early turns disclose name and case facts; after many turns the doctor may re-ask both. Top: a preference-aligned patient can still drift on identity and case content. Bottom: SpeakPatient adds patient-level conditioning under such stress.*

Building a better Omni-based spoken patient therefore requires patient-role adaptation on an open S2S backbone, then further improving interaction quality beyond preference alignment

Corresponding author.

alone [8, 11]. We organize the study around three questions:

- RQ1. How can a general end-to-end S2S foundation model be adapted into a spoken virtual patient?
- RQ2. Does patient-level conditioning improve the quality of an adapted spoken virtual patient beyond preference alignment?
- RQ3. Can patient-level conditioning reduce identity and clinical case drift during long-horizon interaction under late re-ask probes?

To answer these questions, we develop SpeakPatient with (i) patient-role adaptation, (ii) preference alignment, and (iii) patient-level conditioning that keeps a persistent control of who the patient is and what has been disclosed during spoken decoding [6]. We evaluate on held-out interactive S2S dialogues using text-side simulation metrics (including textual naturalness) and automatic speech metrics.

Contributions.

1. We adapt LLaMA-Omni2 into an interactive end-to-end spoken virtual patient via patient-role supervised fine-tuning (SFT), Direct Preference Optimization (DPO), and spoken stages.
2. We introduce SpeakPatient, which adds patient-level conditioning (identity + case state as an LM soft prefix) and improves interactive simulation quality over the DPO spoken-patient baseline.
3. We provide secondary analyses of long-horizon identity/case probes (late re-ask) and patient-state fusion ablations that study how different conditioning choices affect these gains.

2 Related work

We review prior work from three perspectives and emphasize what each line leaves unresolved for long-horizon spoken virtual patients.

2.1 Virtual patients for medical education

Virtual patients have moved from scripted scenarios to LLM-mediated adaptive dialogue for clinical training [12, 13, 14, 15, 3, 16]. Scripted and template-driven systems can enforce case fidelity, but they transfer poorly to open-ended questioning and rarely support natural spoken interaction [17, 13]. Later systems inject structured medical knowledge, case curricula, knowledge graphs, or retrieval over prior turns and clinical facts [1, 18, 19, 20]. Although these designs improve local medical grounding, they typically treat consistency as an inference-time context problem: if the relevant fact is not retrieved or not kept in the prompt window, the patient can still contradict earlier disclosures. We therefore treat prompt/RAG/DST injection as important controls rather than as solved consistency, and contrast them with installing a persistent patient-state signal inside S2S decoding. LLM virtual-patient evaluations emphasize medical accuracy, hallucination, and short-horizon dialogue quality [10, 21, 2], while late re-ask of identity and presenting problem under long spoken consultations remains weakly stressed.

2.2 Spoken interaction and S2S patient adaptation

Spoken medical dialogue is commonly realized with cascaded ASR–LLM–TTS pipelines, which add latency and accumulate recognition and synthesis errors across modules [4, 5, 22]. In contrast, end-to-end S2S backbones such as LLaMA-Omni2 enable direct listen–speak interaction [6, 7, 23], and related work further reduces response delay through streaming and speculative decoding [24, 25, 26]. Speed alone does not yield a medically credible patient, because general S2S models are not trained for patient-role clinical dialogue. Spoken standardized patients are often obtained by prompting a general S2S service, without a published patient-role adaptation recipe on an open backbone [27]. SpeechMedAssist shows that backbone fine-tuning matters for medical speech dialogue, but it adapts LLaMA-Omni2 as a physician-side assistant rather than as a patient [8]. Doctor-side adaptation does not supply patient-side role grounding, and neither prompting nor physician-side fine-tuning addresses long-horizon consistency under late re-ask.

2.3 Long-dialogue consistency and persistent patient state

Long-horizon consistency remains a failure mode of conversational agents: local fluency can coexist with forgotten commitments or rewritten facts across turns [9, 10]. Preference optimization improves response preference alignment, but it supervises preferred replies rather than explicit cross-turn identity commitments [11]. Prompt- or knowledge-augmented virtual patients improve style and local grounding, yet continuity failures persist in multi-turn history-taking [19, 10]. Inference-time remedies re-supply context through dialogue-state tracking [28, 29, 30], growing spoken/text history [30], or retrieval over case or knowledge memory [1, 31]. These mechanisms help when the right evidence is available, but they do not install a persistent patient-state control signal inside S2S decoding. Persona and character-consistency methods improve textual role stability [32, 33, 34], but they target general character traits rather than medically constrained case trajectories and are not designed to condition end-to-end speech generation. Prior remedies either re-inject context at inference time or stabilize text personas; neither provides an explicit, persistent patient-state conditioning path for long-horizon spoken virtual patients.

3 Method

SpeakPatient improves end-to-end spoken virtual patients by adding patient-level conditioning on LLaMA-Omni2. A preference-aligned patient can still answer fluently while drifting on who the patient is or what was disclosed. SpeakPatient encodes identity and disclosed case content into a representation that is routed into the LM soft prefix so patient-side control shapes spoken decoding.

3.1 Overview: Patient-Conditioned LM Control

Implementation builds on LLaMA-Omni2 [6]. We separate two notions that late re-asks jointly stress:

- **Identity**: who the patient is (name, demographics, persona);
- **Case**: what happened to the patient (chief complaint, symptoms, history already disclosed).

A dual-state memory $\mathbf{M}_t = [\mathbf{M}_t^{\mathrm{id}} \| \mathbf{M}_t^{\mathrm{case}}]$ accumulates both streams from dialogue history. They map to a patient-consistency representation

$$z_{\mathrm{patient}} = [z_{\mathrm{identity}};\, z_{\mathrm{case}}] = f(\mathbf{M}_t), \tag{1}$$

and turn-level response context h_t is formed from the current doctor speech (pooled acoustic encoding; no ASR transcript)

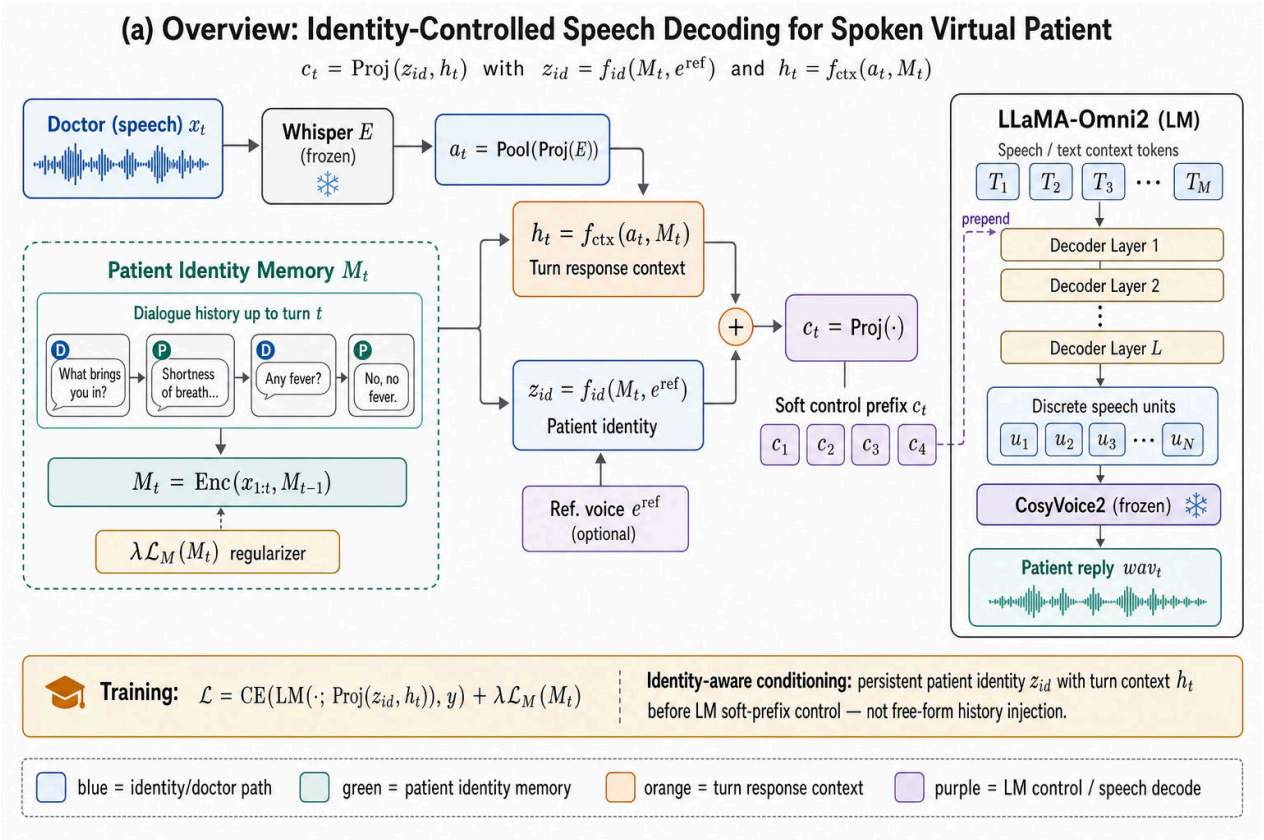

Figure 2: *SpeakPatient overview. Dual-GRU memory yields* z_{identity} *and* z_{case}*; soft prefix* c_t *(*L_p*=4) conditions the LM before CosyVoice2.*

together with shared memory cues. SpeakPatient prepends a short soft control prefix to the LM:

$$c_t = \text{Proj}(z_{\text{identity}},\ z_{\text{case}},\ h_t), \tag{2}$$

implemented as additive fusion then a linear decoder to L_p=4 prefix tokens (Sec. 3.2). Spoken replies follow the standard Omni2 stack under this prefix:

$$\text{units}_t = \text{LM}(\cdot\,; c_t), \qquad \text{wav}_t = \text{CosyVoice2}(\text{units}_t). \tag{3}$$

CosyVoice2 does not take a continuous consistency latent, so control is applied at the LM interface (Fig. 2). Unlike a preference-aligned patient without this path, the LM sees c_t plus the usual speech/text context, not raw retrieved transcript chunks as the sole consistency mechanism.

3.2 Dual-State Memory and Turn Context

z_{patient} must not be inferred from the current turn alone, or the model can rewrite who the patient is every turn. SpeakPatient therefore maintains a dual-state memory

$$\mathbf{M}_t = [\mathbf{M}_t^{\text{id}} \| \mathbf{M}_t^{\text{case}}] \in \mathbb{R}^{128+128}, \tag{4}$$

updated recurrently from streaming doctor speech (no ASR transcript in the memory path). Turn audio x_t is mapped by the Omni2 Whisper encoder and speech projector, then mean-pooled to $h_t \in \mathbb{R}^{896}$; a 14-d process vector u_t is concatenated, and each slice is a dedicated GRUCell:

$$\mathbf{M}_t^{(\cdot)} = \text{GRUCell}^{(\cdot)}([h_t; u_t],\ \mathbf{M}_{t-1}^{(\cdot)}), \quad (\cdot) \in \{\text{id}, \text{case}\}. \tag{5}$$

Soft control reads identity and case latents from $\mathbf{M}_t$:

$$\begin{aligned} z_{\text{identity}} &= f_{\text{id}}([\mathbf{M}_t; e^{\text{ref}}]), \\ z_{\text{case}} &= f_{\text{mot}}([h_t; \mathbf{M}_t]), \\ c_t &= \text{reshape}\big(W_{\text{dec}}(z_{\text{identity}} + z_{\text{case}})\big) \\ &\in \mathbb{R}^{4\times 896}, \end{aligned} \tag{6}$$

where f_{id} and f_{mot} are Linear–GELU–Linear maps to $\mathbb{R}^{896}$, e^{ref} is a learnable reference embedding, and c_t is scaled then prepended to LM embeddings. Turn context h_t may change every turn with the question; z_{identity} should stay person-stable, while z_{case} tracks disclosed-case cues. History does not bypass this path as free-form context injection: case and identity enter generation by shaping z_{patient}, then c_t.

3.3 Training Strategy

The LLM is the decision core, while the speech encoder and CosyVoice2 generator provide listening and speaking. Patient-role medical behavior is largely linguistic: who the patient is, what has been disclosed, and how to answer under clinical questioning. We therefore adapt Omni2 with a staged strategy that first installs patient capability in text, then re-aligns speech modules with limited spoken data (Table 1; Figure 3).

Stage 0: inject patient role via text. We freeze speech-related modules and train the LLM core on patient-role medical dialogues, updating text-conditioned reply generation while leaving the encoder/adaptor/generator unchanged. This stage turns a

Table 1: *Training stages for matched comparison. Both systems share Stage 0→DPO and 1b→2; SpeakPatient adds Stage 1a.*

Stage	Purpose	Trainable
Stage 0	Patient-role text adaptation	LLM
DPO	Preference alignment	LLM
Stage 1a	Conditioning (c_t)	Id/case + LM
Stage 1b	Text→units	Speech gen.
Stage 2	Spoken S2S alignment	S2S path

general Omni2 backbone into a medically constrained patient interlocutor in the text modality.

DPO: preference alignment. Stage 0 alone can still produce locally fluent but clinically undesirable replies (unsupported claims, identity slips, off-role answers). We therefore continue LLM training with Direct Preference Optimization [11] on patient chosen/rejected pairs, yielding the preference-aligned DPO Base patient used in our matched comparison.

Stage 1a: learn patient-level conditioning. Preference alignment improves reply quality, but it does not install an explicit, persistent control signal for identity and disclosed case state. SpeakPatient therefore inserts Stage 1a after the shared DPO checkpoint: dual-GRU / projection modules and the LM are trained jointly so that soft prefix c_t conditions generation,

$$\begin{aligned} \mathcal{L} &= \mathrm{CE}\big(\mathrm{LM}(\cdot;\, c_t),\, y\big) + \lambda\, \mathcal{L}_{\mathbf{M}}(\mathbf{M}_t), \\ \mathcal{L}_{\mathbf{M}}(\mathbf{M}_t) &= \lambda_{\mathrm{traj}}\, \mathrm{MSE}\big(G_\theta(\mathbf{M}_t),\, u_t\big) \\ &\quad + \lambda_{\mathrm{persona}}\, \mathrm{CE}\big(C_\theta(\mathbf{M}_t^{\mathrm{id}}),\, y^{\mathrm{per}}\big) \\ &\quad + \lambda_{\mathrm{clinical}}\, \mathrm{MSE}\big(\mathrm{normalize}(\mathbf{M}_t^{\mathrm{case}}),\, v_t\big), \end{aligned} \tag{7}$$

where y is the gold patient reply, y^{per} is a discrete persona label, G_θ / C_θ are linear readouts, and $v_t \in \mathbb{R}^{128}$ is an L2-normalized rolling hash-bag over disclosed clinical free-facts (defaults $\lambda_{\mathrm{traj}}{=}\lambda_{\mathrm{persona}}{=}0.05$, $\lambda_{\mathrm{clinical}}{=}1.0$). Thus $\mathcal{L}_{\mathbf{M}}$ is *not* contrastive and *not* memory-token CE: it is MSE reconstruction of process features u_t, persona-label CE on the identity slice, and MSE alignment of the case slice to v_t. The first CE term trains replies under c_t and back-propagates into z_{identity}, z_{case}, and Proj; $\mathcal{L}_{\mathbf{M}}$ directly supervises the dual-GRU memory $\mathbf{M}_t$ that feeds those latents. DPO Base skips this stage; both arms then share the same spoken stages below. Because Stage 1a also exposes SpeakPatient to ∼29k extra supervised turns, the matched comparison evaluates the full SpeakPatient recipe (conditioning path + Stage 1a data) rather than a pure architectural ablation.

Stages 1b–2: re-align speech with limited spoken data. Stage 1b freezes the backbone and trains the speech generator on text→unit pairs; Stage 2 performs spoken speech→speech alignment on the S2S path. Because patient knowledge is already in the LLM, only a moderate amount of paired speech is needed to restore listen–speak behavior.

3.4 Data Construction

Existing medical corpora are dominated by physician-side QA or text-only consultations, which mismatch our target: a spoken *patient* under multi-turn history-taking. We therefore build a patient-centric pipeline with a text corpus for role/preference/conditioning learning and a speech corpus for modality re-alignment (Table 2).

Table 2: *SpeakPatient training / evaluation data overview.*

Split	Scale	Role
Stage 0 text	∼100k	Patient-role SFT
MediTOD train	216 dlg. / ∼10k	OSCE core
OSCE short mix	+ shorter cases	Coverage
Stage 1a flats	∼29k turns	Conditioning
DPO pairs	∼5.1k	Preference
Stage 1b speech	∼15k–47k	Text→units
Stage 2 speech	∼35k	Spoken S2S
Eval (held-out)	30 dlg. / 1,312	Interactive test

TextPatient data. For patient-role adaptation and conditioning, we start from MediTOD OSCE-style history-taking dialogues [35] (216 training cases; long multi-turn encounters with case sheets and personas) and expand with shorter OSCE-style patient dialogues to increase case coverage. Each training turn is paired with patient-side supervision: a case sheet / persona prompt, the doctor utterance, and the gold patient reply. For SpeakPatient Stage 1a, we further attach identity and disclosed-case targets (persona labels and free-fact clinical deltas accumulated along the dialogue) so $\mathcal{L}_{\mathbf{M}}$ can supervise dual-state memory $\mathbf{M}_t$. Preference data for DPO are constructed as trajectory-level chosen/rejected patient replies (about 5.1k pairs), covering identity drift, unsupported clinical claims, and off-role answers. Stage 0 uses a larger English patient-text pool (about 100k turns sampled for SFT) to inject broad patient-role behavior before MedTOD-centric conditioning.

SpeechPatient data. Speech targets are synthesized with CosyVoice2 under a shared speaker prompt so unit learning and interactive decoding share the same voice interface as Omni2. Stage 1b trains text→unit generation on MedTOD/OSCE patient utterances mixed with longer Omni2 instruct speech samples [6], because MedTOD turns alone are often short. Stage 2 uses paired doctor speech / patient speech turns for end-to-end spoken alignment. Held-out evaluation never reuses these training cases: interactive tests use 30 unseen MediTOD scenarios (1,312 generated patient turns).

3.5 Evaluation Metrics

At inference, SpeakPatient updates $\mathbf{M}_t$, forms z_{patient} and h_t, applies c_t, then runs LM+CosyVoice2. Evaluation uses interactive rollout (each reply conditions on the model's own earlier replies, not gold transcripts). We define complementary metric groups:

- **Overall reply quality.** Medical Accuracy (Med Acc↑): turn-level case alignment; Textual naturalness / Dialogue Coherence (Coher↑): whether the reply sounds like a natural patient turn (GPT judge, 0–2, reported as $100 \times \mathrm{mean}/2$)—this is *text* naturalness, not speech-prosody MOS; Hallucination rate (Halluc.↓): unsupported clinical claims; Turn Contradiction Rate (TCR↓): dialogue-level self-contradiction.
- **Early identity.** Opening identity error (Open. ID err↓): identity/CC/onset errors (align=0) in the first 20% of turns (not the late re-ask probe).
- **Depth scores.** Early / Mid / Late Med Acc↑ on dialogue-depth thirds.
- **Secondary fact retention.** Name retain↑ / Disease retain↑ at ∼75% depth (name: rule match to gold free-fact

names; disease: chief-complaint category consistency, GPT judge).

- **Automatic speech metrics.** ASR-WER↓ (%): Whisper-tiny vs. intended reply; SpkSim↑: Cam++ cosine to the CosyVoice prompt; UTMOS↑: SpeechMOS automatic naturalness on a ∼1–5 scale; Lat_{FU}↓ (ms): time to first speech unit (macro-averaged per-dialogue p50). Human speech MOS and learner-facing educational studies are left for future work.

4 Experiments

We answer RQ1–RQ3 under one matched interactive protocol (Section 4.1): RQ1 via the Stage 0→DPO→1b→2 adaptation recipe (DPO Base), RQ2 via SpeakPatient vs DPO Base quality gains, RQ3 via late re-ask probes and conditioning ablations; speech metrics are reported as a complement.

4.1 Experimental setup

Protocol. Interactive evaluation on 30 unseen MedTOD history-taking scenarios (1,312 generated patient turns) [35], with speech-in/speech-out rollout as in Section 3.5 (no ASR cascade).

Systems. We compare matched systems based on the same LLaMA-Omni2 0.5B backbone (Table 1; Figure 3). DPO Base applies patient-role adaptation and preference alignment followed by spoken stages (Stage 0→DPO→1b→2). SpeakPatient starts from the same DPO checkpoint and introduces patient-level conditioning before the shared spoken stages (Stage 0→DPO→1a→1b→2). Because Stage 1a also adds ∼29k supervised turns, gains should be read as conditioning path + extra Stage 1a supervision jointly (not a pure architecture-only delta). As inference-time context controls on the same DPO Base Stage 2 checkpoint, we further report Plain prompt, long-context history, retrieval-augmented (RAG), and dialogue-state (DST) injection under the same interactive S2S protocol (Table 3). As an external reference (not a matched Omni ablation), we also report a cascaded ASR–LLM–TTS pipeline on the same 30 scenarios (Table 3, bottom). Training and the matched Omni ablations use 4×V100-32GB GPUs; the bottom-row first-audio latencies (SpeakPatient 362.8 ms vs cascade TTFA 5283.7 ms) are measured on RTX 5090 with FlashAttention v4 and are not directly comparable to V100 Lat_{FU} in Table 4.

Metrics. Definitions follow Section 3.5. Table 3 reports interactive quality, late re-ask, fusion, and the cascaded ASR–LLM–TTS reference; Table 4 reports automatic speech metrics. GPT judges use three independent scores with median aggregation (pairwise agreement on Med Acc / Coher / Halluc. is 97.4% / 96.0% / 97.1% on SpeakPatient turns). All reported conclusions concern automatic simulation-quality metrics; they do not establish educational effectiveness or clinician-validated realism.

4.2 RQ1: Adapting a general S2S model into a spoken virtual patient

RQ1 asks how a general end-to-end S2S model can become a speech-in/speech-out patient without an ASR–LLM–TTS cascade. We study a staged adaptation strategy that first injects patient-role behavior, then preference-aligns replies, and finally re-aligns speech generation (Figure 3). The resulting DPO Base sustains interactive spoken history-taking with 47.29 Med Acc and 52.74 Coher (Table 3), and serves as the matched baseline for RQ2–RQ3.

4.3 RQ2: Does patient-level conditioning improve quality beyond preference alignment?

RQ2 asks whether persistent patient-level information further improves an already preference-aligned spoken patient. SpeakPatient raises Med Acc $+8.35$ pp and textual naturalness (Coher) $+7.85$ pp, and lowers Halluc. -4.95 pp, TCR -0.21, and Open. ID err -14.45 pp over DPO Base (Table 3). Early/Mid/Late Med Acc all rise, so the gain is not confined to one dialogue phase. Inference-time controls on the same DPO Base checkpoint show that plain / long-context prompting alone stays near DPO Base on Med Acc / Coher, whereas RAG and DST prompt injection reach aggregate Med Acc / Coher in a similar band to SpeakPatient (Table 3, middle). On late Name/Disease probes, however, prompt stuffing does not match SpeakPatient's Name retain (30.0%): Plain stays at DPO Base (20.0%), while RAG/DST drop to 10.0%/6.7% despite RAG's high Disease score (53.3%). Thus SpeakPatient is competitive with strong prompt-side grounding on automatic simulation quality, while installing a persistent soft-prefix control inside Omni decoding rather than stuffing retrieved text each turn. We interpret the DPO Base→SpeakPatient lift as evidence for the full conditioning recipe (Stage 1a path + data), not as a claim that soft prefixes uniquely dominate every prompt baseline on every aggregate metric.

4.4 RQ3: Can conditioning maintain identity and case consistency over long horizons?

RQ3 asks whether conditioning reduces long-horizon identity/case drift under late re-ask—not whether consistency is fully solved. At ∼75% dialogue depth, SpeakPatient improves Name $+10.0$ pp and Disease $+13.3$ pp over DPO Base (Table 3), but absolute rates remain modest (Name 30.0%, Disease 43.3% on n=30 scenarios; Wilson 95% CIs in the table caption overlap those of DPO Base). With only 30 held-out scenarios and automatic judges, we treat these probes as supportive evidence of reduced drift rather than definitive proof of maintained patient consistency. Open. ID err also drops -14.45 pp, aligning early and late identity evidence under the same automatic protocol.

To test how the patient-state path is used, we keep SpeakPatient Stage 2 weights fixed and change only how information enters the LM prefix $\hat{z}$: Prefix only, or identity latent only ($\hat{z}{=}z_{\text{identity}}$). Most of the Med Acc / Coher lift over DPO Base appears as soon as a trained SpeakPatient prefix path is present; Prefix / Identity-latent even slightly outperform full SpeakPatient on several aggregates (within ∼1.5 pp). Name / Disease and Open. ID err are more sensitive to injection choice: Name rises Prefix→Identity→SpeakPatient (23.3→26.7→30.0), and Disease jumps mainly under full SpeakPatient (43.3 vs 30.0 for identity-only). A complementary switch $\hat{z}{=}z_{\text{case}}$ (identity latent off) yields Name 30.0 / Disease 46.7 / Open. ID err 20.48. Overall, prefix / identity-only variants can edge Med Acc / Coher, while late Name/Disease probes favor keeping both identity and case streams in the full SpeakPatient fusion. We therefore claim that the conditioning path is useful and that joint identity+case fusion helps selected long-horizon retention measures, not that dual-state always dominates every metric.

4.5 Speech-level evaluation

As a speech-side complement (not an RQ), Table 4 reports ASR-WER, SpkSim, UTMOS, and Lat_{FU} on all interactive patient turns (1,312 clips/system). Both systems are in the high UTMOS band (≈4.3–4.4) with strong SpkSim (≈0.80–0.84); ASR-WER is usable (≈14–16%), slightly better for SpeakPatient; Lat_{FU} is

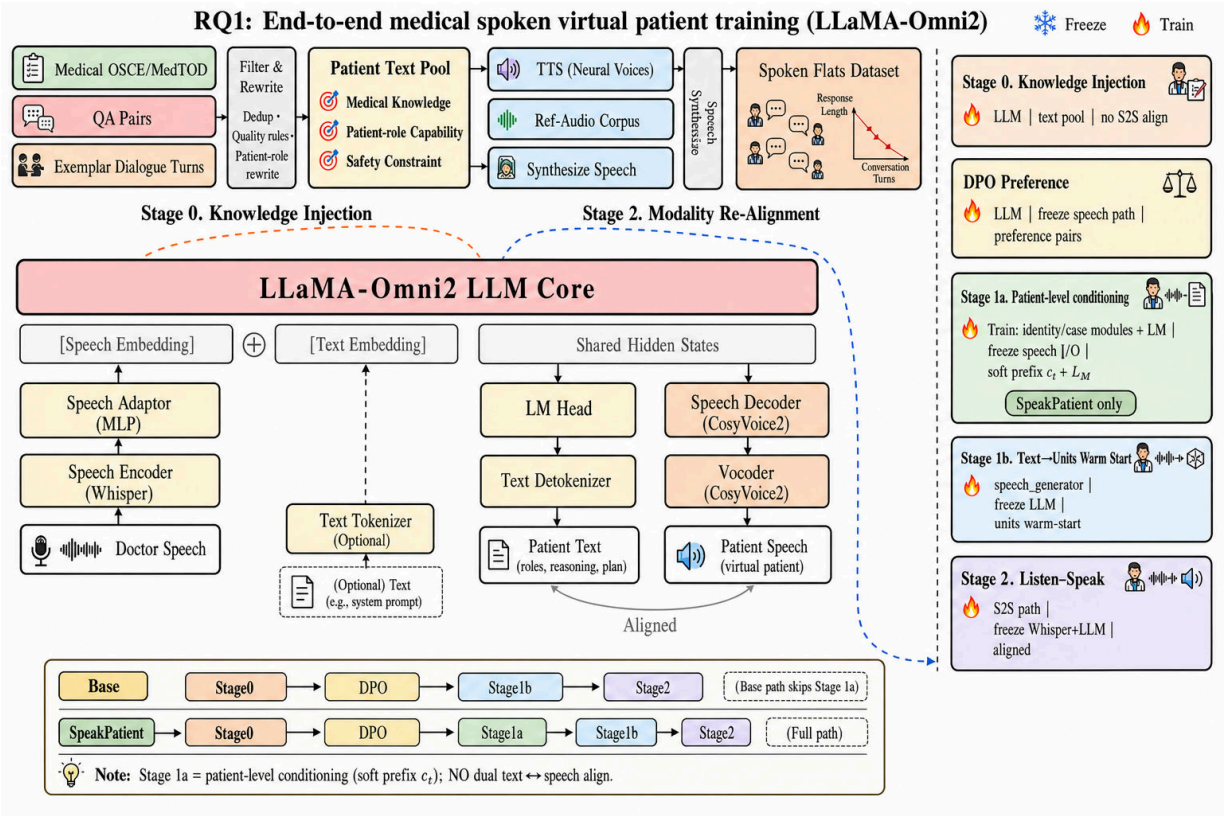

Figure 3: *SpeakPatient training on LLaMA-Omni2 (data → architecture → stages). Both arms share Stage 0→DPO; SpeakPatient inserts patient-conditioning Stage 1a before the shared spoken stages 1b→2.*

Table 3: *Interactive S2S evaluation on 30 unseen MedTOD scenarios (GPT×3). Top: matched DPO Base vs SpeakPatient (RQ1–RQ3; 1,312 turns). Then: inference-time prompt / RAG / DST on DPO Base; fusion ablations on SpeakPatient weights. Bottom: cascade reference—SpeakPatient 0.5B vs Whisper ASR→GPT-4o-mini→CosyVoice2 TTS (same 30 scenarios; parentheses = first-audio p50 / TTFA on RTX 5090+FlashAttention v4; other rows are V100-matched Omni runs). Name/Disease are dialogue-level rates (n=30); Wilson 95% CIs: DPO Name* 20.0% [9.5, 37.3], *SpeakPatient Name* 30.0% [16.7, 47.9]*; DPO Disease* 30.0% [16.7, 47.9], *SpeakPatient Disease* 43.3% [27.4, 60.8].

Model	Med Acc↑	Coher↑	Halluc.↓	TCR↓	Open. ID err↓	Early↑	Mid↑	Late↑	Name↑	Disease↑
DPO Base	47.29	52.74	23.93	0.81	33.73	45.77	53.65	42.45	20.0	30.0
SpeakPatient	55.64	60.59	**18.98**	0.60	**19.28**	54.23	63.81	48.86	**30.0**	43.3
+ Plain prompt	47.33	52.59	23.40	0.76	34.94	45.77	53.88	42.33	20.0	30.0
+ Long context	47.94	53.54	23.09	0.75	32.53	48.28	54.68	40.85	13.3	33.3
+ RAG prompt	55.91	61.66	18.98	0.63	37.35	54.23	64.50	48.97	10.0	**53.3**
+ DST prompt	56.40	61.66	**18.83**	**0.56**	44.58	51.37	67.58	50.23	6.7	30.0
+ Prefix only	56.29	61.59	20.05	0.59	26.51	**55.84**	63.24	49.77	23.3	33.3
+ Identity latent only	**56.75**	**62.04**	19.59	0.75	24.10	55.26	**64.16**	**50.80**	26.7	30.0
SpeakPatient 0.5B (362.8 ms)	54.48	59.66	19.02	0.69	21.18	54.36	61.89	47.28	30.0	43.3
ASR–GPT-4o-mini–TTS (5283.7 ms)	**84.28**	**85.88**	**1.70**	0.91	**1.18**	**86.87**	**84.42**	**81.57**	**100.0**	**100.0**

Table 4: *Automatic speech metrics on all interactive turns (1,312 clips/system). ASR-WER (%): intelligibility proxy; SpkSim: speaker similarity to CosyVoice prompt; UTMOS (∼1–5): automatic speech naturalness; Lat_{FU} (ms): time to first speech unit.*

Model	ASR-WER↓	SpkSim↑	UTMOS↑	Lat_{FU}↓
DPO Base	15.96	**0.842**	**4.389**	1448.6
SpeakPatient	**14.06**	0.802	4.344	**1445.8**

tied at ≈1.45 s. Small SpkSim/UTMOS gaps match the shared CosyVoice2 path—conditioning mainly lifts text-side simulation (Table 3). UTMOS is an automatic speech-naturalness proxy; textual naturalness is Coher.

Against the cascaded Whisper→GPT-4o-mini→CosyVoice2 reference (Table 3, bottom), the cascade leads on automatic Med Acc / Coher / Halluc. and late Name/Disease, as expected from a much stronger LLM with explicit text state. SpeakPatient 0.5B remains ≈14.6× lower in first-audio latency on the same RTX 5090+FlashAttention v4 setup (362.8 vs 5283.7 ms TTFA) while staying fully speech-in/speech-out without an ASR transcript in the patient path. We present this as a latency–quality reference rather than a claim that 0.5B S2S matches GPT-4o-mini on Med Acc.

5 Conclusions

This work develops SpeakPatient, an end-to-end spoken virtual patient on LLaMA-Omni2. We adapt Omni2 into a speech-in/speech-out patient through patient-role SFT, preference alignment, and spoken stages, then improve it with patient-level conditioning so the agent better remains the same person and case over long encounters. On matched interactive evaluation under automatic simulation-quality metrics, SpeakPatient improves reply quality and late identity/case retention over the preference-aligned Omni patient; automatic speech metrics further support the S2S setting. Inference-time RAG/DST prompting on DPO Base can reach a similar aggregate Med Acc / Coher band, while SpeakPatient learns a persistent soft-prefix control without retrieval stuffing; we do not claim educational effectiveness or clinical realism from these automatic scores alone. Against a cascaded ASR–GPT-4o-mini–TTS reference (Table 3, bottom), SpeakPatient trades lower automatic Med Acc for substantially lower first-audio latency under the same 30-scenario protocol. Long spoken consultations still leave open challenges: absolute late name retention remains modest on $n{=}30$ scenarios, dual-state fusion is not uniformly best on every aggregate metric, and identity, onset/timeline, and earlier disclosures can still drift under extended questioning. Future work may pursue a same-data LM-only Stage 1a control; stronger optimization for long-dialogue accuracy and identity/case retention under extended history-taking; closing the cascade quality gap without cascade latency; and human listening / OSCE-style educational evaluation.

6 References

[1] Y. Li, C. Zeng, J. Zhong, R. Zhang, M. Zhang, and L. Zou, "Leveraging large language model as simulated patients for clinical education," *arXiv preprint arXiv:2404.13066*, 2024.

[2] D. A. Cook, J. Overgaard, V. S. Pankratz, G. Del Fiol, and C. A. Aakre, "Virtual patients using large language models: scalable, contextualized simulation of clinician-patient dialogue with feedback," *Journal of medical Internet research*, vol. 27, p. e68486, 2025.

[3] Y. Hicke, J. Geathers, N. Rajashekar, C. Chan, A. G. Jack, J. Sewell, M. Preston, S. Cornes, D. Shung, and R. Kizilcec, "Medsimai: Simulation and formative feedback generation to enhance deliberate practice in medical education," *arXiv preprint arXiv:2503.05793*, 2025.

[4] W. Cui, D. Yu, X. Jiao, Z. Meng, G. Zhang, Q. Wang, S. Y. Guo, and I. King, "Recent advances in speech language models: A survey," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 13 943–13 970.

[5] J. Peng, Y. Wang, B. Li, Y. Guo, H. Wang, Y. Fang, Y. Xi, H. Li, X. Li, K. Zhang, S. Wang, and K. Yu, "A survey on speech large language models for understanding," *IEEE Journal of Selected Topics in Signal Processing*, vol. 20, pp. 2–31, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:273549261

[6] Q. Fang, Y. Zhou, S. Guo, S. Zhang, and Y. Feng, "Llama-omni2: Llm-based real-time spoken chatbot with autoregressive streaming speech synthesis," in *Annual Meeting of the Association for Computational Linguistics*, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:278327164

[7] W. Yu, S. Wang, X. Yang, X. Chen, X. Tian, J. Zhang, G. Sun, L. Lu, Y. Wang, and C. Zhang, "Salmonn-omni: A standalone speech llm without codec injection for full-duplex conversation," *ArXiv*, vol. abs/2505.17060, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:278886216

[8] S. Chen, J. Wang, W. Chen, and Z. Wei, "Speechmedassist: Efficiently and effectively adapting speech language models for medical consultation," in *Annual Meeting of the Association for Computational Linguistics*, 2026. [Online]. Available: https://api.semanticscholar.org/CorpusID:284544191

[9] A. Maharana, D.-H. Lee, S. Tulyakov, M. Bansal, F. Barbieri, and Y. Fang, "Evaluating very long-term conversational memory of llm agents," *ArXiv*, vol. abs/2402.17753, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268041615

[10] D. Li and S. Lebai Lutfi, "Large language model–based virtual patient systems for history-taking in medical education: Comprehensive systematic review," *JMIR Medical Informatics*, vol. 14, p. e79039, 2026.

[11] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *Advances in neural information processing systems*, vol. 36, pp. 53 728–53 741, 2023.

[12] R. Wang, S. Milani, J. C. Chiu, J. Zhi, S. M. Eack, T. Labrum, S. M. Murphy, N. Jones, K. Hardy, H. Shen *et al.*, "Patient-{\Psi}: Using large language models to simulate patients for training mental health professionals," *arXiv preprint arXiv:2405.19660*, 2024.

[13] R. Wasenmüller, K. Hilbert, and C. Benzmüller, "Script-based dialog policy planning for llm-powered conversational agents: A basic architecture for an" ai therapist"," *arXiv preprint arXiv:2412.15242*, 2024.

[14] I. Steenstra, F. Nouraei, and T. W. Bickmore, "Scaffolding empathy: Training counselors with simulated patients and utterance-level performance visualizations," *arXiv preprint arXiv:2502.18673*, 2025.

[15] K. Lee, S. Lee, E. H. Kim, Y. Ko, J. Eun, D. Kim, H. Cho, H. Zhu, R. E. Kraut, E. Suh *et al.*, "Adaptive-vp: A framework for llm-based virtual patients that adapts to trainees' dialogue to facilitate nurse communication training," *arXiv preprint arXiv:2506.00386*, 2025.

[16] H. Yu, J. Zhou, L. Li, S. Chen, J. Gallifant, A. Shi, X. Li, W. Hua, M. Jin, G. Chen *et al.*, "Aipatient: Simulating patients with ehrs and llm powered agentic workflow," *arXiv preprint arXiv:2409.18924*, 2024.

[17] A. Aster, S. V. Ragaller, T. Raupach, and A. Marx, "Chatgpt as a virtual patient: written empathic expressions during medical history taking," *Medical Science Educator*, pp. 1–10, 2025.

[18] P. Staples, P. Clarke, C. Leininger, C. Principato, M. Shafiei, G. Goodwin, G. A. Ryslik, and R. F. Dougherty, "Building virtual patients for training mental health professionals," 2025.

[19] D. Li and S. L. Lutfi, "Smart virtual patients for tcm: Combining knowledge graphs and language models," in *2025 10th International Conference on Information Technology and Digital Application (ICITDA)*. IEEE, 2025, pp. 1–8.

[20] J. Wu, X. Wu, Y. Zheng, and J. Yang, "Medkp: Medical dialogue with knowledge enhancement and clinical pathway encoding," *ArXiv*, vol. abs/2403.06611, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268358622

[21] Y. Yu, H. Ying, H. Jin, W. Jiang, D. Xian, B. Wang, Z. Yang, and M. Wu, "Medkgeval: A knowledge graph-based multi-turn evaluation framework for open-ended patient interactions with clinical llms," *arXiv preprint arXiv:2510.12224*, 2025.

[22] R. S.-H. Yuen, T. T.-L. Tse, and J. Zhu, "Internalizing asr with implicit chain of thought for efficient speech-to-speech conversational llm," *arXiv preprint arXiv:2409.17353*, 2024.

[23] Q. Team, "Qwen3.5-omni technical report," 2026. [Online]. Available: https://api.semanticscholar.org/CorpusID:287607592

[24] Y. Shi, Y. Shu, S. Dong, G. Liu, J. Sesay, J. Li, and Z. Hu, "Voila: Voice-language foundation models for real-time autonomous interaction and voice role-play," *arXiv preprint arXiv:2505.02707*, 2025.

[25] Y. Wang, H. Liu, Z. Cheng, R. Wu, Q. Gu, Y. Wang, and Y. Wang, "Vocalnet: Speech llm with multi-token prediction for faster and high-quality generation," *arXiv preprint arXiv:2504.04060*, 2025.

[26] L. Mai, "Relays2s: A dual-path speculative generation for real-time dialogue," 2026. [Online]. Available: https://api.semanticscholar.org/CorpusID:286773939

[27] A. Emerson, K. Evanini, S. Somay, K. Frome, L. A. Ha, and P. Harik, "Conversational ai for virtual standardized patients using a speech-to-speech llm," in *International Workshop on Spoken Language Translation*, 2026. [Online]. Available: https://api.semanticscholar.org/CorpusID:289321200

[28] S. Young, M. Gašić, B. Thomson, and J. D. Williams, "Pomdp-based statistical spoken dialog systems: A review," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1160–1179, 2013.

[29] M. Henderson, B. Thomson, and S. Young, "Word-based dialog state tracking with recurrent neural networks," in *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, 2014, pp. 292–299.

[30] N. E. Ghazal, A. Caubrière, and V. Vielzeuf, "The speech-llm takes it all: A truly fully end-to-end spoken dialogue state tracking approach," *ArXiv*, vol. abs/2510.09424, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:282055902

[31] P. Feng, Z. Ma, W. Chen, Y. Li, S. Wang, K. Yu, and X. Chen, "Enhancing speech-to-speech dialogue modeling with end-to-end retrieval-augmented generation," *arXiv preprint arXiv:2505.00028*, 2025.

[32] Y.-P. Chen, N. Nishida, H. Nakayama, and Y. Matsumoto, "Post persona alignment for multi-session dialogue generation," *arXiv preprint arXiv:2506.11857*, 2025.

[33] C.-H. Hsu, Y.-J. Lin, and H.-Y. Kao, "From persona to person: Enhancing the naturalness with multiple discourse relations graph learning in personalized dialogue generation," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2025, pp. 187–198.

[34] H. Lee and Y. Ko, "Dynamic persona generation through commonsense inference in dialogues," *Computer Speech & Language*, p. 101896, 2025.

[35] V. V. Saley, G. Saha, R. J. Das, D. Raghu *et al.*, "Meditod: An english dialogue dataset for medical history taking with comprehensive annotations," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 16 843–16 877.

A Example Interactive Dialogues

We show bordered transcripts from interactive speech-in/speech-out rollouts on held-out MediTOD cases. The trainee asks as **Doctor**; the spoken virtual patient replies as **Patient** (decoded text). Excerpts illustrate opening identity, mid-history follow-up, and late re-ask under the same spoken protocol.

SpeakPatient: opening identity and chief complaint.

Doctor: Can you please confirm your name and date of birth?

Patient: Yes, my name is Oliver Smith and I am a student. My date of birth is 15 May 1998.

Doctor: What brought you in today?

Patient: I've been having this chest pain for the last two days.

Doctor: When did the chest pain start?

Patient: It started around 10 AM, and it's been getting worse since then.

Doctor: Were you doing anything when it started? Any trauma to the chest?

Patient: No, I haven't had any injuries or anything like that.

Doctor: Does the pain come on when you take a deep breath, or at other times?

Patient: It's just when I take a deep breath.

Doctor: On a scale of one to ten, how bad is the pain right now?

Patient: I'd say it's about a six or seven.

Doctor: Does the pain radiate anywhere else, like to your arm or jaw?

Patient: No, it stays in my chest.

Opening contrast: DPO Base vs SpeakPatient. On the same identity question, DPO Base may skip the name request while SpeakPatient answers it.

Doctor: Can I confirm your name and date of birth, please?

DPO Base: Hi, thanks for seeing me. I've been feeling really unwell.

SpeakPatient: Yes, my name is Aisyah binti Ahmad and I was born on the 15th of March this year.

SpeakPatient: mid-consultation symptom follow-ups.

Doctor: Uh, what can I do for you?

Patient: I've been feeling really unwell.

Doctor: When did your sore throat start? What about the chills?

Patient: It started last week.

Doctor: Have you had any difficulties swallowing?

Patient: Yeah, it hurts to swallow.

Doctor: Have you had any headaches?

Patient: No, I haven't had any headaches.

Doctor: Any stuffy nose or runny nose?

Patient: No, I don't have any stuffy or runny nose.

Doctor: Have you taken any medication for this?

Patient: Just some paracetamol, but it hasn't really helped much.

Doctor: Any fever that you noticed at home?

Patient: I think I had a mild fever yesterday evening.

Late identity re-ask. The name is retained, while age/date wording may still drift, motivating further work on long-horizon consistency.

Doctor: Can I confirm your name and date of birth, please?

Patient: Yes, my name is Aisyah binti Ahmad and I'm 32 years old.

Late chief-complaint re-ask.

Doctor: And just to make sure I have the story right—what is the main problem that brought you in today?

Patient: I've been having this chest pain for the last two days, especially when I take a deep breath.