DentalVoice: EMR Grounded Synthetic Dental Speech Dialogues for AI Scribes

Anonymous ACL submission

Abstract

Dental AI scribes convert chairside dental speech into structured electronic medical records (EMRs). Public data for this task is scarce because dental visits include private speech, regional language, and noisy operatory audio. We introduce DentalVoice, a Chinese dental dialogue and speech dataset for AI scribe research. It contains 13,861 multi-turn dialogues, 162,612 turns, and approximately 274 hours of generated speech. Each case is grounded in a real dental EMR, guided by a dental knowledge graph, conditioned on patient profiles, and augmented with chairside acoustic conditions. We validate the generated dialogues using LLM judges and clinical experts, finding that clinical knowledge retrieval improves safety and professionalism, while persona constraints enhance interaction realism. Next, we benchmark dental automatic speech recognition (ASR) and Subjective-Objective-Assessment-Plan (SOAP) note extraction. Results reveal persistent errors in dialectal ASR, dental terminology recognition, treatment normalization, and distinguishing planned from completed procedures.

1 Introduction

Ambient voice AI and automated clinical scribes are rapidly emerging as foundational technologies for streamlining medical documentation, significantly alleviating the administrative burden of manual data entry during patient encounters (Bundy et al., 2024). The necessity for such voice-driven systems is particularly pronounced in dentistry. Operating in a highly constrained, hands-busy environment, dentists must wear surgical masks and gloves during intraoral procedures, rendering concurrent manual documentation fundamentally impractical. Consequently, clinicians are frequently forced to reconstruct patient records retrospectively. This reliance on post-visit recall is highly vulnerable to information loss. Therefore, deploying a specialized AI scribe tailored for dental workflows is essential to simultaneously alleviate the clinician's documentation burden and safeguard the rigorous quality of patient records.

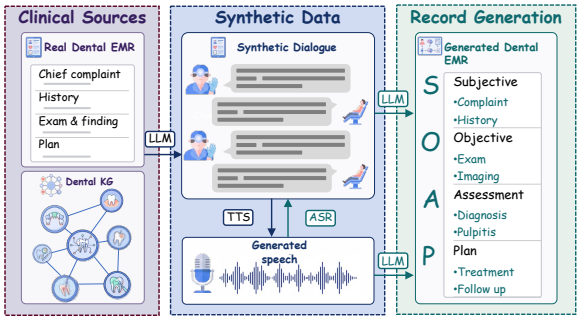

Figure 1: Overview of the DentalVoice dataset. The corpus is generated via a rigorous EMR-to-dialogue-to-speech pipeline and serves as a foundational benchmark for ASR-based SOAP note extraction.

However, the development of robust AI scribes is severely bottlenecked by a profound scarcity of high-quality audio-dialogue corpora. This deficit primarily stems from stringent privacy constraints and the historical absence of audio archiving in clinical workflows (Malo et al., 2026). Curating novel datasets requires navigating complex patient consent protocols, as authentic chairside encounters inherently embed sensitive, non-clinical information. Furthermore, generating precise ground-truth annotations for multi-turn, unstructured clinical audio is labor-intensive and challenging.

To circumvent these data acquisition barriers, prior studies have synthesized textual dialogues directly from existing EMRs or relied on simulated mock consultations (Korfiatis et al., 2022; Fareez et al., 2022). While standard medical speech corpora (Le-Duc et al., 2025; Le-Duc, 2024; Htun et al., 2024; Olatunji et al., 2023) offer isolated audio segments, they lack the interactive dynamics of multi-turn consultations. Furthermore, previous synthetic datasets were severely hindered by

a dual bottleneck: the tendency of earlier LLMs to generate medically imprecise, ungrounded dialogue (Kim et al., 2025), and the pervasive machine-like artifacts of legacy TTS systems lacking physical acoustic realism (Xie et al., 2025). Because these existing resources fail to jointly capture rigorous clinical accuracy and auditory authenticity, they remain insufficient for developing and evaluating real-world ambient dental AI scribes.

While earlier synthetic approaches fell short of clinical standards, recent advancements in generative AI have finally established synthetic data generation as a highly viable solution to this data scarcity. We address this gap with DentalVoice[1], a comprehensive Chinese dental audio-dialogue dataset designed specifically for AI scribe applications. Generated via a rigorous EMR-to-dialogue-to-speech pipeline, DentalVoice is grounded in authentic clinical records and comprises 13,861 multi-turn dialogues paired with ∼274 hours of high-fidelity audio across Mandarin and regional dialects. Because modern text-to-speech (TTS) systems can readily achieve human-like naturalness, the primary bottleneck in synthetic corpora has shifted to textual authenticity. Consequently, we rigorously validate DentalVoice's conversational and clinical realism through systematic evaluations by both automated LLMs and expert dentists.

Building upon this validated dataset, we further establish foundational downstream benchmarks to inspire future clinical NLP research. These experiments rigorously evaluate ASR performance and LLM-based SOAP note extraction efficacy across diverse state-of-the-art models.

Our contributions are as follows.

- We introduce DentalVoice, a Chinese dental audio-dialogue dataset for AI scribes, featuring 13,861 multi-turn EMR-grounded conversations and 274 hours of speech across multiple dialects.
- We propose a novel EMR-to-dialogue-to-speech synthesis pipeline, integrating longitudinal patient profiling, a script-writing agent, and an environment-aware TTS module to replicate clinical conditions.
- We rigorously validate the synthesized dialogues via LLM judges and expert dentists, confirming their high clinical safety, professionalism, and roleplay realism.
- We establish foundational benchmarks for a two-stage automated clinical scribing pipeline, providing robust baselines for ASR and LLM-based SOAP note extraction.

[1]Anonymous code, data, and pipeline scripts: https://anonymous.4open.science/r/DentalVoice-0307

2 Related Work

Clinical documentation resources and modality coverage. Clinical documentation resources cover only parts of the data needed for training ambient dental AI scribes. Chinese online consultation resources such as MedDialog (Zeng et al., 2020) and cMedQA (Zhang et al., 2017) support text-based medical consultation modeling, but lack chairside speech and structured dental records. Clinical note datasets such as ACI Bench (Yim et al., 2023) and PriMock57 (Korfiatis et al., 2022) support dialogue-to-EMR and SOAP note extraction research, but focus on English primary care. Medical ASR corpora expand spoken data for Chinese and multilingual settings (Chen et al., 2025; Le-Duc et al., 2025), but do not connect speech to dental record generation. No existing resource jointly provides dental case grounding, multi-turn doctor-patient-family dialogue, paired speech, dialectal variation, and structured record labels. DentalVoice targets this missing combination.

Synthetic medical dialogue generation. Current methods for synthesizing medical dialogues have evolved from early human-driven simulations (Korfiatis et al., 2022; Yim et al., 2023), which are hindered by scalability issues and high expert costs, to LLM-driven generation. This LLM-driven paradigm broadly encompasses three core strategies: dual-agent simulation for autonomous encounters (Ge et al., 2025), knowledge-grounded prompting via structural references (Liu and Liu, 2026; Wang et al., 2024), and model fine-tuning on authentic datasets to capture clinical nuances (Zeng et al., 2020; Lai et al., 2025; Chen et al., 2023). While effective for general clinical encounters, these approaches present critical shortcomings. Notably, they often fail to optimize for the cost-efficiency of large-scale generation. Furthermore, they tend to produce machine-like text that lacks the realistic conversational tone essential for speech synthesis, and they generally rely on broad medical knowledge rather than the specialized expertise required for disciplines like dentistry. Our pipeline addresses these gaps by optimizing for scalable, realistic, and domain-specific dialogue generation.

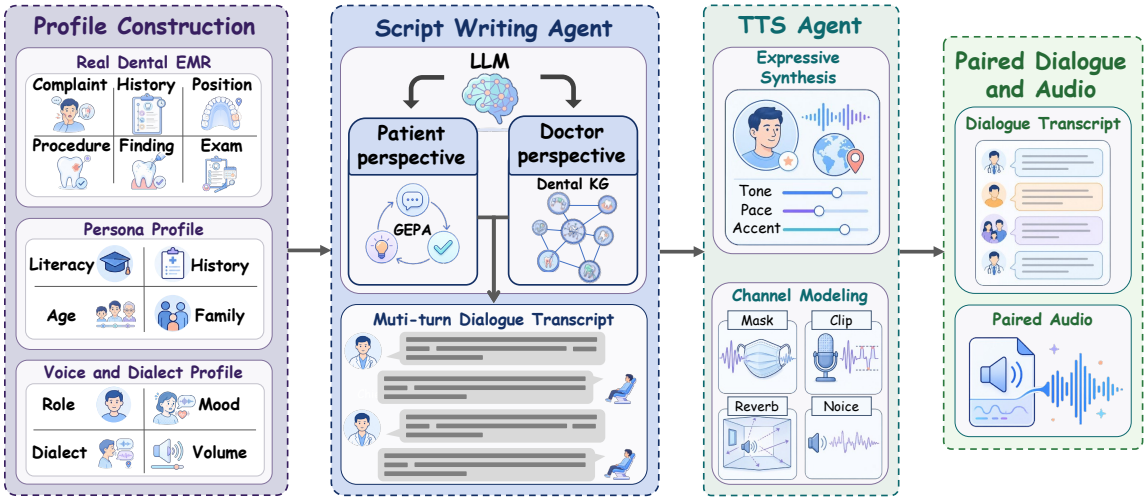

Figure 2: DentalVoice data synthesis pipeline.

3 Dataset Synthesis Pipeline

Figure 2 summarizes the DentalVoice synthesis pipeline, which maps real dental EMRs to multi-turn doctor-patient dialogues and paired audio. It contains three components that correspond to the following subsections. First, **Patient Profile Construction** builds longitudinal patient profiles with voice, dialect, education level, and medical history. These profiles control both dialogue style and speech rendering. Second, the **Script Writing Agent** converts each EMR into a dialogue script. It uses patient profile information, dental knowledge retrieval, and role specific prompting to generate patient responses and doctor questions. Third, the **TTS Agent** renders the script as audio. It combines expressive synthesis with speaker specific channel modeling to approximate dental operatory recording conditions.

3.1 Patient Profile Construction

Rather than assigning randomized personas to isolated clinical cases, our generative pipeline is firmly anchored in longitudinal patient profiles. This generation process relies on persistent patient attributes to ensure consistency, specifically incorporating voice, dialect, educational level, and cumulative medical history.

To maintain clinical realism and legal compliance, we exclusively employ open-source voice models, meticulously filtering out stylized acoustics (e.g., anime or professional broadcasting) to preserve natural conversational dynamics. Each profile is subsequently parameterized with a specific educational background and linguistic preference (Mandarin or a regional dialect). These demographic anchors serve a dual purpose: they guide the **script writing agent** in aligning health literacy and syntactic complexity, while simultaneously instructing the **TTS agent** to synthesize congruent prosodic features and authentic accents.

3.2 The Script Writing Agent

The Script Writing Agent collects all the information, including the patient profile and knowledge graph and translates the EMRs into a dynamic, multi-turn conversational script. To maximize the authenticity of the interaction, the agent applies distinct prompt optimization strategies for both the patient and the doctor, ensuring the dialogue reflects real-world clinical dynamics.

3.2.1 Patient Perspective

A fundamental challenge in generating synthetic patient dialogue is mitigating the inherently pedantic and clinically omniscient tone typical of standard LLMs. Rather than manually engineering rigid stylistic constraints, we employ GEPA (Agrawal et al., 2025), a Generative Prompt Optimization framework, to automatically distill natural conversational dynamics from authentic clinical encounters. To this end, we curated 14 real-world dental consultation transcripts with explicit informed consent. These were partitioned into 12 optimization samples and 2 validation samples. The raw audio recordings were processed via speaker diarization and ASR, from which high-quality conversational QA pairs were filtered to construct the final dataset.

Prompt Optimization. The prompt evolution is driven by a three-agent LLM architecture. First, a *Task LM* simulates the patient by responding

to authentic clinical queries. Operated at a high temperature, this model is explicitly configured to maximize colloquial diversity and spontaneous generation. Its outputs are subsequently scored by an *Evaluator LM*—a zero-temperature judge dedicated strictly to assessing linguistic realism. The evaluator operates across four core dimensions: penalizing machine-like artifacts, medical terminology abuse, and excessive formality, while rewarding the natural integration of colloquial granularity (e.g., hesitations and discourse markers). Crucially, the evaluator is instructed to disregard factual divergence from the source transcripts, isolating its focus entirely on stylistic authenticity. Finally, a high-reasoning *Reflection LM* analyzes the feedback on underperforming mini-batches to identify linguistic deficits. It then iteratively refines the prompt using Pareto-optimal selection, ensuring the system avoids local stylistic optima.

Optimization Outcomes. Starting from a rudimentary seed prompt, the framework autonomously converged on a highly sophisticated behavioral instruction set. As detailed in Appendix D, the finalized prompt strictly suppresses medical jargon in favor of colloquial symptom descriptions. Furthermore, it enforces authentic conversational dynamics by mandating the integration of everyday discourse markers, micro-hesitations, and context-aware patient psychology, such as treatment anxiety or spontaneous digressions.

3.2.2 Doctor Perspective

While the patient model simulates linguistic imperfection, the doctor role requires rigorous, evidence-based reasoning. To achieve this, our framework first establishes a dental knowledge base extracted from the *National Qualification Examination Guidelines for Licensed Dental Practitioners* (National Medical Examination Center of China, 2025). Rather than building an exhaustive encyclopedia, we target common clinical diseases. This yields a structurally asymmetric database where 8 broad chief complaints map to 36 specific diagnoses. To navigate this mapping, each chief complaint branches into multiple inquiry dimensions. Each dimension drives a targeted follow-up question, and the patient's responses systematically narrow down the candidate diagnoses.

Bidirectional Anchoring. Crucially, our objective is not to construct a rigid, forward-reasoning diagnostic decision tree. Instead, given a known chief complaint and a predetermined target diagnosis, our goal is to enhance the diversity of the doctor's clinical inquiries and plausibly simulate the patient's responses. To achieve this, we first establish a bidirectional matching mechanism to anchor the conversation. We observed that traditional vector similarity struggles with chief complaint classification; embedding models often conflate clinically distinct intents (e.g., gingival bleeding versus routine scaling requests) due to broad semantic overlaps, rendering threshold selection unreliable. Consequently, we deploy a lightweight LLM to accurately classify unstructured inputs into eight predefined complaints or an undefined category. Diagnostic mapping, in contrast, is highly deterministic: extracted clinical entities are standardized and efficiently mapped to potential diagnoses via vector search. If no confident matches are retrieved, the agent activates a universal fallback protocol.

Inquiry Path Sampling. Upon establishing the bidirectional anchors (chief complaint c and target diagnosis d^*), injecting all combinations of forward and backward inquiry pathways into the LLM context is both computationally inefficient and overly deterministic. To balance core diagnostic logic with real-world clinical randomness, we dynamically sample the intermediate inquiry path to serve as the prompt context.

The next inquiry dimension q is selected from a candidate set $\mathcal{Q}$ based on the expected entropy reduction $\Delta H(q)$ over the candidate diagnosis distribution. Let $\mathcal{B}(D \mid h)$ denote the diagnostic belief state given the current inquiry history h. The expected entropy reduction is formulated as:

$$\Delta H(q) = H(\mathcal{B}(D \mid h)) - \mathbb{E}_{a \sim P(a|q,d^*)} \left[H(\mathcal{B}(D \mid h, q, a)) \right] \tag{1}$$

To prevent strictly greedy, locally optimal dialogues, we sample the next question using a softmax distribution ($\tau = 0.7$) over the $\Delta H(q)$ values. This introduces occasional suboptimal selections typical of real clinical practice. Once the candidate diagnoses narrow down ($|\mathcal{B}(D)| \leq 2$), the system enters a directed verification phase to confirm d^*. The final sampled sequence serves as a rigorous clinical constraint for the script-writing agent.

3.2.3 Dynamic Context Generation

Beyond semantic generation, the agent analyzes each utterance in conjunction with the underlying patient profile to dynamically formulate auxiliary natural language instructions. Acting as conversational acoustic prompts, these instructions guide the downstream TTS agent in rendering highly contextualized vocal micro-expressions. Specifically, these context hints explicitly dictate real-time adjustments to speech pacing, emotional polarity, volume, and overall attitudinal resonance.

3.3 The TTS Agent

To transform textual responses into highly realistic audio, the final step employs an environment-aware TTS Agent. We adopt Doubao-Seed-TTS 2.0 (ByteDance, 2025) as the synthesis backbone, leveraging its enhanced emotional expressiveness and fine-grained vocal control. Furthermore, to authentically replicate the acoustic artifacts inherent to clinical environments, the agent operates through a two-stage architecture—functioning sequentially as both a speech synthesizer and an integrated physical signal processor.

First, the **Expressive Synthesis Module** generates dynamic, human-like speech from the conversational scripts. A base voice is initially parameterized by regional dialect to reflect specific patient demographics. To achieve highly realistic delivery, the module incorporates dynamic context hints generated by the upstream Script Writing Agent.

Second, the **Integrated Digital Signal Processing (DSP) Module** grounds the synthetic audio into a physical clinical space by simulating real-world channel degradations. This pipeline sequentially applies: (1) *surgical mask muffling* via a localized filter (Badh and Knowles, 2023) to simulate the acoustic impedance of dentists' masks; (2) *hardware bottlenecks* using non-linear soft clipping to mimic microphone dynamic range limitations; and (3) *spatial acoustics* by convoluting the signal with a room impulse response and layering ambient background noise. By integrating both semantic generation and targeted physical degradation, the TTS Agent outputs clinical audio sequences that closely mirror genuine real-world recordings.

4 The DentalVoice Dataset

4.1 Overview

DentalVoice is a large-scale, longitudinal audio-dialogue dataset engineered to capture the complexity of authentic dental consultations. Spanning approximately 274 hours of paired speech across 13,861 multi-turn dialogues, it establishes a new benchmark for scale in this domain. As shown in Table 1, while existing medical benchmarks predominantly lack domain-specific dental grounding, paired conversational audio, or demographic diversity, DentalVoice bridges these gaps through three foundational anchors: (1) **EMR Anchoring**, by deriving every dialogue directly from a ten-year archive of clinical records; (2) **Knowledge Grounding**, utilizing dental knowledge graphs to guide medically sound reasoning; and (3) **Acoustic Realism**, rendering speech under dental operatory conditions with diverse regional dialects and authentic acoustic effects.

4.2 Dataset Schema and Statistics

Each entry in DentalVoice comprises four core components: a **patient persona** (gender, age, dialect, education), **dialogue text** with speaker-labeled turns (doctor, patient, and optionally a family member), anonymized **EMR** mapped to standard SOAP fields, and 16 kHz mono **audio**.

The dataset contains 3,861 unique patients, reflecting authentic clinical demographics. The cohort maintains a balanced gender ratio and spans from pediatric to geriatric populations (ages 3–98, mean 41.2). To capture real-world communicative diversity, DentalVoice incorporates varying educational backgrounds and regional accents (Mandarin, Shaanxi, Sichuan, and Dongbei). Notably, approximately 20% of the dialogues involve accompanying family members, introducing a multi-party dynamic that is vital for pediatric and elderly care but rarely addressed in existing medical datasets. Clinically, the records encompass major categories of frontline dental visits, with an automated AI pipeline systematically annotating each case to quantify clinical complexity.

Furthermore, DentalVoice facilitates the modeling of *longitudinal clinical trajectories*. Unlike datasets limited to isolated conversational snapshots, our records integrate temporal histories. Patients average 3.4 documented cases, with extreme tails capturing up to 28 visits and treatment spans extending over six years. This longitudinal depth is instrumental for training medical LLMs capable of continuous patient monitoring and long-term diagnostic reasoning.

Table 1: Comparison of medical dialogue and speech corpora.

Benchmark	Multi-Turn	EMR-based	KG-based	Specialty	Dur. (h)	#Dialogs	#Turns	Language
Speech-only medical corpora (audio + transcript, no EMR)								
MultiMed (Le-Duc et al., 2025)	×	×	×	General med	150	n/a	n/a	Multilingual
VietMed (Le-Duc, 2024)	×	×	×	General med	16	n/a	n/a	Vietnamese
myMediCon (Htun et al., 2024)	×	×	×	General med	11	n/a	n/a	Burmese
AfriSpeech-200 (Olatunji et al., 2023)	×	×	×	Med + general	≈123	n/a	n/a	Afr. English
Simulated medical consultations (multi-turn audio, actor scripts, no real EMR)								
PriMock57 (Korfiatis et al., 2022)	✓	×	×	Primary care	9	57	≈5,244	English
Fareez et al. (Fareez et al., 2022)	✓	×	×	Respiratory	55.15	272	n/a	English
DentalVoice (ours)	✓	✓	✓	**Dental**	**≈274**	**13,861**	**162,612**	**Chinese + 3 dialects**

5 Validation

This section validates the dialogue generation stage of DentalVoice. We assess whether the synthesized dialogues preserve Clinical Safety, Clinical Professionalism, and Roleplay Realism prior to speech rendering. Across a fixed set of dental cases, we evaluate different generation strategies, model capacities, and prompt configurations using structural statistics, LLM judges, and expert dentist evaluations.

5.1 Experimental Design

We validated generation quality using 15 anchor cases, encompassing a diverse range of diseases, treatments, and patient profiles.

Candidate LLM selection. We evaluate three Doubao-Seed-2.0 variants (ByteDance, 2026) to test how model capacity affects instruction following in dialogue synthesis. We choose this family because it produces fluent Chinese dialogues with relatively little artificial tone in our pilot generations. Mini, Lite, and Pro provide three capacity levels within the same model family. We also compare them with GPT-5.4 (OpenAI, 2026) as a strong general LLM and Baichuan-M3 (Dou et al., 2026) as a medical LLM. Additional results with other LLMs are reported in Appendix C.

Dialogue generation strategies. To evaluate clinical fidelity and procedural adherence, we compare two dialogue generation strategies. In the **Default Mode**, the model acts as a centralized "scriptwriter": it is fed the complete patient record and retrieved clinical context, generating the entire physician-patient dialogue in a single inference pass. In contrast, following the design in (Ge et al., 2025), the **Dual-Agent Mode** employs a multi-turn, interactive framework between distinct doctor and patient agents, driven by a stochastic intent graph. Rather than relying on a static global context, this dynamic approach incrementally updates each agent's context window with the evolving dialogue history and sampled intent states.

Ablation prompt configurations. We evaluate four prompt configurations to separate the effects of clinical knowledge and conversational style. **Baseline** asks the model to convert the raw EMR into a physician patient dialogue without auxiliary guidance. **Retrieval Only** adds retrieved dental inquiry paths to guide differential diagnosis. **Persona Only** adds conversational style constraints learned from real dialogues, but excludes clinical retrieval. **Full Realism** combines retrieved clinical knowledge with persona and style constraints.

5.2 LLM Judge Evaluation

To evaluate dialogue quality, three LLM judges (Gemini-3.1 Pro, Claude-Opus-4.7 Thinking, and GPT-5.5) independently score conversations on a 1-to-5 scale with qualitative rationales. The assessment covers three dimensions: (1) **Roleplay Realism** (fidelity of patient tone and emotion); (2) **Clinical Safety** (screening for allergies, comorbidities, etc.); and (3) **Clinical Professionalism** (depth and coherence of diagnostic reasoning). Table 2 summarizes all experimental results across different configurations.

Generation strategy. We first evaluated two generation strategies. The dual-agent approach incurs prohibitive computational costs, consuming up to ∼8,500 prompt tokens per generation due to iterative history accumulation. In comparisons, the default global script writer mode is highly efficient—requiring only ∼900 tokens, while simul-

Table 2: Evaluation results of generated dialogues across five LLMs and tested prompt configurations. Statistics include turns, prompt tokens and completion tokens. Realism, Safety, and Professionalism use a 1 to 5 scale.

Model	Prompt Version	Dialogue Statistics			LLM Evaluation			
		Turns	P.Tokens	Comp.Tokens	Realism	Safety	Prof.	Sum
Doubao-Seed-2.0-Mini	Baseline	7.9	322	3119	2.26	1.92	2.05	6.23
	Retrieval Only	10.1	658	3569	2.97	**4.18**	**4.72**	11.87
	Persona Only	14.3	526	3696	**3.74**	1.58	2.15	7.47
	Full Realism	12.0	808	3635	3.68	2.42	2.97	9.07
	Dual Agents	11.4	6776	3311	3.59	3.82	3.85	11.25
Doubao-Seed-2.0-Lite	Baseline	11.1	322	3660	2.82	2.17	2.45	7.44
	Retrieval Only	12.0	658	3783	3.03	4.07	4.68	11.78
	Persona Only	14.7	526	4288	4.53	1.73	2.76	9.02
	Full Realism	15.0	898	3848	**4.75**	**4.24**	**4.72**	**13.71**
	Dual Agents	9.1	8691	3631	3.95	3.53	3.95	11.43
Doubao-Seed-2.0-Pro	Baseline	10.0	322	4399	2.65	2.12	2.22	6.99
	Retrieval Only	14.4	658	4581	3.17	4.06	**4.57**	11.80
	Persona Only	18.5	526	4792	4.47	2.06	2.71	9.24
	Full Realism	16.7	898	5056	**4.67**	**4.10**	4.46	**13.23**
	Dual Agents	10.6	7590	3846	4.16	3.83	3.90	11.89
GPT-5.4	Full Realism	38.3	898	5920	**4.82**	4.25	**4.78**	**13.85**
Baichuan-M3-235B	Dental Role+Persona	21.1	661	4766	3.78	**4.53**	3.93	12.24

Note. Favorable results are highlighted in **blue**, whereas notable performance degradations are indicated in teal.

taneously achieving higher Realism, Safety, and Professionalism scores.

Model capacity. Breaking down the results by model scale, we observe that while model size has minimal impact on the utilization of external medical knowledge, the Mini model struggles significantly with persona adherence. Furthermore, this smaller model lacks the capacity to simultaneously integrate retrieved context, persona constraints, and safety requirements. Notably, the Lite model achieves a total score of 13.71, outperforming the Pro variant (13.23) and nearly matching GPT-5.4 (13.85). This suggests that once a model reaches a capacity threshold sufficient for robust instruction following, structured prompting diminishes the marginal returns of further parameter scaling. While GPT-5.4 produces significantly longer dialogues and the highest absolute Realism score, its marginal gain over Doubao-Seed-2.0-Lite does not justify the higher API cost. Thus, Doubao-Seed-2.0-Lite emerges as the optimal choice, delivering GPT-level performance with cost efficiency.

Prompt strategy. Finally, we compared the impact of different prompting techniques. While the **Persona Only** prompt successfully improves Role-play Realism, it compromises Clinical Safety, causing Doubao-Seed-2.0-Lite's safety score to drop from a baseline of 2.17 to 1.73. This indicates that prioritizing conversational mimicry often leads models to abandon procedural safety. In contrast, integrating retrieved knowledge boosts both Safety and Professionalism across the Doubao variants. Remarkably, our general Doubao LLM equipped with knowledge retrieval actually outperforms specialized medical models like Baichuan M3 in clinical professionalism (4.57 vs. 3.93), thoroughly validating the effectiveness of our curated dental knowledge base. By combining both persona and knowledge retrieval, the **Full Realism** approach offers an optimal overall performance.

5.3 Expert Evaluation

Five practising dentists provided 750 blinded scores through the questionnaire interface (Appendix H), scoring each on Clinical Safety and Professionalism. They consistently ranked the Retrieval Only condition highest on both clinical dimensions. Their condition-level means correlate with the LLM-judge ensemble at Pearson $r = 0.98$ (Safety) and $r = 0.91$ (Professionalism), supporting the use of LLM judges for large-scale ablation.

6 Experiments and Results

To demonstrate the dataset's utility, we establish baselines for a two-stage automated clinical scribing pipeline, comprising ASR transcription and subsequent SOAP note extraction.

6.1 Task 1: Dental ASR

To assess transcription capabilities for Chinese dental conversations, we primarily evaluate the Qwen3-ASR series (Shi et al., 2026) (0.6B and 1.7B parameters), which are explicitly optimized for Chinese and dialectal speech. For benchmarking, we compare them against Whisper Large-v3 (Radford et al., 2023) (1.55B parameters), a widely adopted multilingual baseline whose performance remains limited in complex Chinese contexts. Performance is measured using the standard Character Error Rate (CER) and a domain-weighted Dental Terminology Character Error Rate (DT-CER) that penalizes errors in core clinical terminology to assess semantic accuracy. As shown in Table 3, the Qwen3 series significantly outperforms Whisper on Chinese dental conversations. Qwen3-1.7B achieves the best overall performance (3.01% CER, 3.88% DT-CER), while the lightweight Qwen3-0.6B remains highly competitive (4.13% CER, 6.14% DT-CER).

Table 3: Overall ASR Performance.

Model	CER(%)	DT-CER(%)
Whisper L-V3	10.40	13.93
Qwen3-0.6B	4.13	6.14
Qwen3-1.7B	**3.01**	**3.88**

The gap between overall CER and DT-CER across models highlights the difficulty of transcribing specialized medical entities compared to general speech. However, scaling up model capacity effectively mitigates this issue; the larger Qwen3-1.7B leverages its enhanced semantic understanding to substantially reduce DT-CER to 3.88%.

As shown in Table 4, Qwen3-1.7B remains robust across regional dialects, whereas Whisper degrades severely on complex accents. Notably, all models struggle most with the Shaanxi dialect, likely due to its phonological divergence and underrepresentation in training corpora.

Table 4: Robustness Analysis across Dialects (CER %).

Model	Mand.	Dong.	Sich.	Shaan.
Whisper L-V3	9.13	10.34	15.43	18.14
Qwen3-0.6B	3.79	4.44	4.54	6.83
Qwen3-1.7B	**2.75**	**3.46**	**3.35**	**4.90**

6.2 Task 2: SOAP Note Extraction

Building upon Task 1, the second stage of our pipeline addresses the ultimate objective: extracting structured SOAP notes from the speech-to-text outputs. We evaluate five mainstream LLM extractors (DeepSeek-V4-Flash, DeepSeek-V4-Pro, Qwen-Max, Baichuan-M3-Plus, and GPT-4) applied directly to these ASR outputs across 1,000 stratified cases. Performance is measured via a rule-based objective layer—assessing tooth position accuracy (TPA), safety concept recall (SCR), and treatment concept match (TxCM)—and an LLM-judged atom layer scoring standard SOAP fields. The full score matrix is detailed in the appendix.

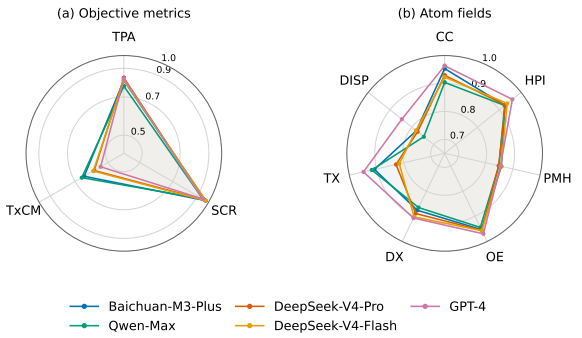

Figure 3: Performance of SOAP note extraction. SOAP fields: chief complaint (CC), present illness (HPI), past medical history (PMH), oral examination (OE), diagnosis (DX), treatment plan (TX), and disposition (DISP).

As shown in Figure 3, Baichuan M3-plus achieves the best overall SOAP note extraction performance, closely followed by GPT-4. However, all extractors exhibit a significant performance bottleneck in TxCM and DISP. This reveals that while LLMs effectively extract general clinical narratives, they struggle to standardize treatment terminology and resolve planned-vs-completed timelines.

7 Conclusion

In conclusion, this paper introduces DentalVoice, a Chinese audio-dialogue dataset designed for dental AI scribes, featuring approximately 274 hours of synthesized clinical conversations. Constructed via a novel EMR-to-dialogue-to-speech pipeline, the dataset leverages knowledge grounding and persona constraints to ensure high clinical safety, professionalism, and interactive realism. Benchmarking on downstream dental ASR and SOAP note extraction tasks confirms the dataset's utility, while also highlighting challenges in dialect processing, clinical terminology recognition, and timeline resolution for treatments. Ultimately, DentalVoice resolves a critical data bottleneck, empowering future advancements in ambient clinical intelligence.

Limitations

We identify three primary limitations regarding the scope of our work.

(a) Audio dialect. While we incorporate three regional dialects, acoustic diversity is constrained by the limited dialect support of current TTS systems. Nevertheless, our findings highlight the critical necessity of dialectal representation in real-world settings, making the expansion of dialect acoustics an essential direction for future work.

(b) Clinical scope. DentalVoice focuses on routine procedures typical of frontline dental clinics (e.g., implantology, orthodontics, and restorative dentistry). Complex, hospital-level workflows, such as maxillofacial surgery, are beyond the scope of this study.

(c) Conversational scope. The synthetic dialogues are strictly centered on the diagnostic and treatment-planning processes that map directly to SOAP notes. Extraneous elements common in real-world consultations, such as casual small talk and price negotiations, are intentionally omitted.

Ethics Statement

Data and consent. The ten-year EMR archive was anonymised before research-team access, with all direct identifiers removed. The 14 audio-recorded consultations used for GEPA prompt optimisation were collected with written informed consent under the practice's IRB-approved research protocol (protocol ID withheld for anonymous review). These recordings remain on-site at the collaborating practice and are not released. Five practising dentists volunteered for the blind rating study, with three identifying by initials and two requesting full anonymity.

Release and intended use. We release the 13,861 synthetic dialogues, the ∼274 hours of synthetic audio, the dental knowledge graph, the generation pipeline, and the evaluation scripts under a research-use-only license. The synthetic corpus contains no patient identifiers. We do not endorse deployment of derived models for clinical decision making without separate regulatory clearance. The anonymous repository for review is available at `https://anonymous.4open.science/r/DentalVoice-0307`.

Use of AI assistance. We used AI tools for language polishing, code assistance, and figure drafting. All scientific claims, experiments, and final text were reviewed and verified by the authors.

References

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, and 1 others. 2025. Gepa: Reflective prompt evolution can outperform reinforcement learning. *arXiv preprint arXiv:2507.19457*.

Gursharan Badh and Thea Knowles. 2023. Acoustic and perceptual impact of face masks on speech: A scoping review. *Plos one*, 18(8):e0285009.

Henry Bundy, Jay Gerhart, Sally Baek, Crystal Danielle Connor, McKenzie Isreal, Ajay Dharod, Casey Stephens, Tsai-Ling Liu, Timothy Hetherington, and Jeffery Cleveland. 2024. Can the administrative loads of physicians be alleviated by ai-facilitated clinical documentation? *Journal of general internal medicine*, 39(15):2995–3000.

ByteDance. 2025. Doubao-seed-tts-2.0: Text-to-speech model. Accessed: 2026-05-25; Volcengine Ark Platform.

ByteDance. 2026. Doubao-seed-2.0: Large multimodal language model. Accessed: 2026-05-25; API endpoint: Volcengine Ark.

Hongzhao Chen, XiaoYang Wang, Jing Lan, Hexiao Ding, Yufeng Jiang, MingHui Yang, DanHui Xu, Jun Luo, Nga-Chun Ng, Gerald WY Cheng, and 1 others. 2025. Mmedfd: A real-world healthcare benchmark for multi-turn full-duplex automatic speech recognition. *arXiv preprint arXiv:2509.19817*.

Siyuan Chen, Colin A Grambow, Mojtaba Kadkhodaie Elyaderani, Alireza Sadeghi, Federico Fancellu, and Thomas Schaaf. 2023. Investigating the utility of synthetic data for doctor-patient conversation summarization. In *INTERSPEECH*, pages 2338–2342.

Chengfeng Dou, Fan Yang, Fei Li, Jiyuan Jia, Qiang Ju, Shuai Wang, Tianpeng Li, Xiangrong Zeng, Yijie Zhou, Hongda Zhang, and 1 others. 2026. Baichuan-m3: Modeling clinical inquiry for reliable medical decision-making. *arXiv preprint arXiv:2602.06570*.

Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, and 1 others. 2022. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *Scientific Data*, 9(1):313.

Chengze Ge, Yu Xu, Qi Shao, and Shengping Liu. 2025. High-quality medical dialogue synthesis for improving emr generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2675–2687.

Hay Man Htun, Ye Kyaw Thu, Hutchatai Chanlekha, Kotaro Funakoshi, and Thepchai Supnithi. 2024. mymedicon: End-to-end burmese automatic speech recognition for medical conversations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12032–12039.

Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Chanwoo Park, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, and 1 others. 2025. Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint arXiv:2503.05777*.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. Primock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598.

Shuzhong Lai, Chenxi Li, Junhong Lai, Yucun Zhong, Chenyu Yan, Xiang Li, Haifeng Li, Gang Pan, Lin Yao, and Yueming Wang. 2025. Asd-illm: An intervention large language model for autistic children based on real clinical dialogue intervention dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8058–8079.

Khai Le-Duc. 2024. Vietmed: A dataset and benchmark for automatic speech recognition of vietnamese in the medical domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17365–17370.

Khai Le-Duc, Phuc Phan, Tan-Hanh Pham, Bach Phan Tat, Minh-Huong Ngo, Thanh Nguyen-Tang, and Truong-Son Hy. 2025. Multimed: Multilingual medical speech recognition via attention encoder decoder. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1113–1150.

Kefei Liu and Meizhu Liu. 2026. Synthetic doctor-patient dialogue generation for robust medical asr: A scalable pipeline for vocabulary expansion and privacy preservation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 525–534.

Marie-Françoise Malo, Sarah Bouhouita-Guermech, Hortense Gallois, Vardit Ravitsky, Bridge2AI Voice Consortium, and Jean-Christophe Bélisle-Pipon. 2026. Using voice and speech data in healthcare: a scoping review of the ethical, legal and social implications. *Frontiers in Digital Health*, 8:1750111.

National Medical Examination Center of China. 2025. *Practical Skills Guide for the 2025 Oral Medical Licensing Examination*. People's Medical Publishing House, Beijing, China.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure FP Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, and 1 others. 2023. Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr. *Transactions of the Association for Computational Linguistics*, 11:1669–1685.

OpenAI. 2026. GPT-5.4. OpenAI Blog. Accessed: 2026-05-25.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, and 1 others. 2026. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*.

Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2024. Notechat: a dataset of synthetic patient-physician conversations conditioned on clinical notes. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15183–15201.

Tianxin Xie, Yan Rong, Pengfei Zhang, Wenwu Wang, and Li Liu. 2025. Towards controllable speech synthesis in the era of large language models: A systematic survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 764–791.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific data*, 10(1):586.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, and 1 others. 2020. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 9241–9250.

Sheng Zhang, Xin Zhang, Hui Wang, Jiajun Cheng, Pei Li, and Zhaoyun Ding. 2017. Chinese medical question answer matching using end-to-end character-level multi-scale cnns. *Applied Sciences*, 7(8):767.

A Reproducibility

The full synthesis pipeline (§3) is released alongside the dataset, including the bidirectional dental knowledge graph as JSON, the GEPA-optimized patient prompt, and the DSP audio-rendering code.

Anchor EMR: 罗女士 / Ms. Luo, age 37; chief complaint: requested comfort scaling. **Persona**: ℓ_h = high, ℓ_v = Mandarin, ℓ_s = specific, ℓ_a = calm, ℓ_f = none.

医生 [L3-social]: 罗女士，今天过来想做什么？
Doctor [L3-social]: *Ms. Luo, what brings you in today?*
患者 [L1-mechanism]: 想做个舒适洁牙。
Patient [L1-mech]: *I'd like a comfort scaling.* → *chief_complaint*
医生 [L1-mechanism]: 好的，先给您做个基础口腔检查。…… 口腔卫生整体尚可，少量软垢和外源性色素沉着，没有明显牙周炎症，符合洁牙指征。
Doctor [L1-mech]: *Let me do an oral exam first. … Hygiene overall acceptable, slight plaque and extrinsic pigmentation, no obvious periodontitis, scaling indicated.* → *exam, assessment*
患者 [L2-clarification]: 嗯…洗牙会不会很疼啊？
Patient [L2-clar]: *Hm… will it hurt much?* *anxiety-cued disfluency*
医生 [L3-social]: 不疼，可能有点酸。我们慢慢来。
Doctor [L3-social]: *Won't hurt, may feel a slight tingle. Take it slow.*
医生 [L3-social]: 接下来两天少喝咖啡浓茶，平时用巴氏刷牙法，每年定期洁牙一到两次就行。
Doctor [L3-social]: *Avoid coffee/strong tea for two days, use Bass brushing daily, scaling 1–2 times per year.* → *plan, instructions*

Figure 4: Example DentalVoice dialogue excerpt (Chinese + English gloss) with turn-level structural labels and EMR field alignment shown in gray.

Pre-registered evaluation scripts (DT-CER implementation, SOAP-extraction prompt templates, statistical tests, random seeds) are included in the release to enable third-party replication of the benchmarks in Sections 6.1 and 6.2. Real audio data collected from the collaborating practice remains on-site and is not part of the release.

B Visualization of the Medical Terminologies

To provide a more intuitive understanding of the dental data characteristics, we conducted a visual analysis of the constructed dental domain data and the dental knowledge graph. Figure 6 highlights the core concepts and primary focus areas within the dataset through a word cloud. Figure 7 details the structure of the dental knowledge graph.

C Validation Across Alternative LLMs

We extend the Section 5 evaluation matrix to two further open-source LLM families (Qwen3.5 and DeepSeek-V4) on the same 15-case anchor set and the same four prompt configurations. Table 5 reports the mean Realism, Safety, and Professionalism scores per cell, averaged over the three LLM judges (Gemini-3.1 Pro, Claude-Opus-4.7 Thinking, and GPT-5.5) on the 1–5 scale and over the 15 anchor cases. The Sum column is the row total over the three dimensions, the best Sum per model row is in bold, and cells marked † aggregate over 14 cases instead of 15 (one generation failed parsing).

Qwen3.5 series. Table 5 (top block) reports three Qwen3.5 mixture-of-experts variants (35B-a3B, 122B-a10B, 397B-a17B) with thinking mode enabled. Across all three sizes, the prompt-strategy ranking matches the Doubao and GPT-5.4 ranking: Retrieval Only > Full Realism > Persona Only > Baseline. Scaling Qwen3.5 from 35B to 397B activated parameters adds at most 0.6 points to the Full Realism sum (11.39 to 11.71), confirming that once a base capacity threshold is met, structured prompting dominates over parameter scaling.

DeepSeek-V4 series. Table 5 (bottom block) reports three DeepSeek-V4 configurations: Flash with internal reasoning disabled (Flash-nothink), Flash with the `reasoning_effort` API parameter set to `high` (Flash-maxthink), and Pro with the same `reasoning_effort=high` setting (Pro-maxthink). Thinking mode on Flash yields negligible movement (Full Realism: 12.34 versus 12.25), suggesting that internal reasoning offers limited benefit when the generation task is already heavily scaffolded by retrieval and persona prompts. Pro with maximum thinking reaches the best DeepSeek result on Retrieval Only (13.27) but does not surpass Doubao-Lite Full Realism (13.71).

Pattern stability across families. The B>D>C>A ranking holds for all eight tested model families (Doubao-Mini, Doubao-Lite, Doubao-Pro, GPT-5.4, Qwen3.5-35B, Qwen3.5-122B, Qwen3.5-397B, DeepSeek-V4 Flash-nothink, DeepSeek-V4 Flash-maxthink, DeepSeek-V4 Pro-maxthink). The Persona-only failure mode (high Realism, sharply lower Safety and Professionalism) reproduces in every family.

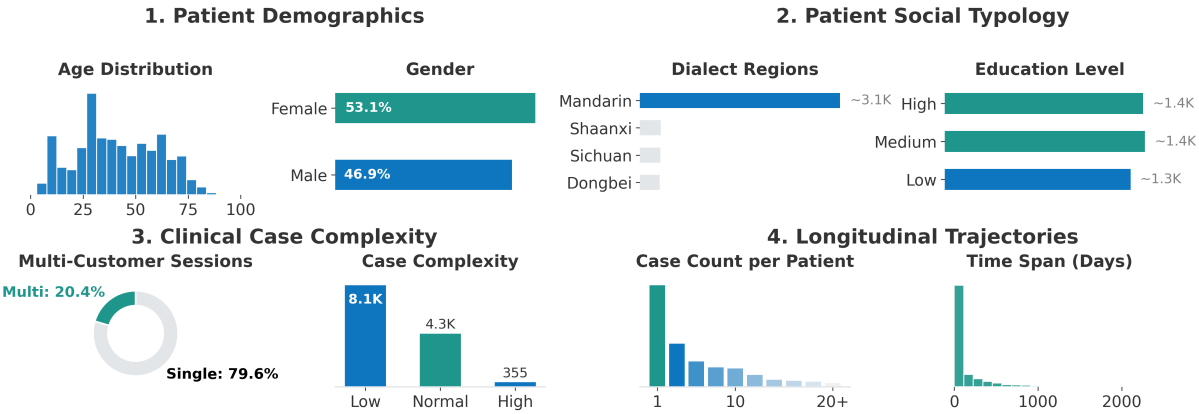

Figure 5: Dataset Statistics

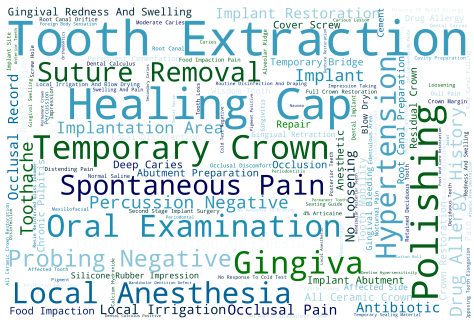

Figure 6: Word Cloud of the Dataset

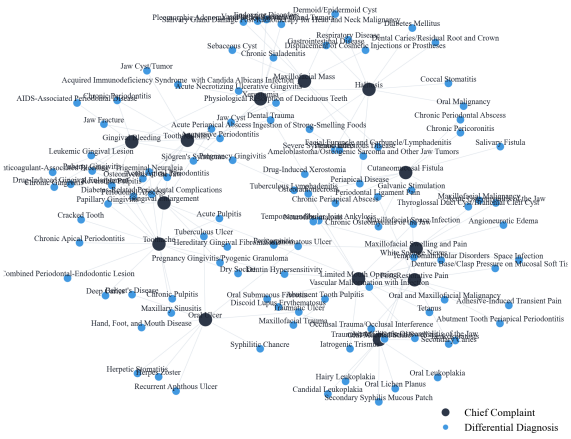

Figure 7: Dental Knowledge Graph

This is a prompt-driven effect rather than a model-specific artifact, and supports the choice of Full Realism as the production configuration in the released DentalVoice pipeline.

D GEPA Optimized Patient Prompt

This appendix summarizes the GEPA optimized patient persona prompt referenced in §3.2.1. The prompt evolved from a single sentence seed through reflective prompt evolution on real recording pairs. The complete prompt is released with the dataset.

GEPA Optimized Patient Prompt

Task positioning. Generate patient responses for Chinese dental consultations. The patient should answer from lived experience rather than clinical expertise.

Generation requirements. Avoid professional terminology. Use lay symptom descriptions. Add natural hesitation when appropriate. Calibrate response length to the doctor's question. Vary wording when the doctor repeats a question. Include patient psychology such as anxiety, uncertainty, or practical concerns.

Prohibitions. Do not use monosyllabic answers by default. Do not use overly polite formal phrases. Do not produce identity incongruent expressions. Do not introduce off topic detours. Do not add factually impossible content. Do not write in a polished written register.

E Expanded Formulation of Inquiry Path Sampling

Motivation for Dynamic Sampling. Upon establishing the bidirectional anchors (chief complaint c and target diagnosis d^*), the system faces a vast space of possible dialogue trajectories. Injecting all combinations of forward and backward inquiry pathways into the LLM context is both computationally intractable due to combinatorial explosion,

Table 5: Dialogue-generation scores for Qwen3.5 (top) and DeepSeek-V4 (bottom), extending Table 2.

Model	Prompt Version	Turns	Doctor Ch.	Patient Ch.	Realism	Safety	Prof.	Sum
Qwen3.5-35B-a3B	Baseline	9.4	311	98	2.77	1.64	2.24	6.65
	Retrieval Only	14.9	581	256	3.31	3.99	4.61	**11.90**
	Persona Only†	16.7	333	184	4.14	1.70	2.46	8.30
	Full Realism	19.3	434	216	4.51	2.79	4.09	11.39
Qwen3.5-122B-a10B	Baseline	12.3	367	136	2.85	2.03	2.42	7.30
	Retrieval Only	15.3	544	270	3.34	4.07	4.62	**12.02**
	Persona Only†	19.6	357	192	4.46	1.74	2.77	8.98
	Full Realism†	19.7	430	261	4.54	3.09	4.32	11.95
Qwen3.5-397B-a17B	Baseline	12.8	359	134	2.88	1.82	2.31	7.02
	Retrieval Only	16.1	592	275	3.39	4.01	4.61	**12.00**
	Persona Only	18.8	351	196	4.45	1.76	2.68	8.89
	Full Realism	19.7	421	235	4.55	3.14	4.02	11.71
DeepSeek-V4 Flash-nothink	Baseline	11.9	425	133	3.36	1.81	2.50	7.67
	Retrieval Only	12.9	508	333	3.61	4.17	4.71	**12.49**
	Persona Only	25.5	580	284	4.45	2.00	2.97	9.41
	Full Realism	26.5	637	307	4.67	3.27	4.40	12.34
DeepSeek-V4 Flash-maxthink	Baseline	14.9	473	162	3.46	1.89	2.74	8.09
	Retrieval Only	11.7	444	292	3.62	4.19	4.70	**12.51**
	Persona Only	24.7	571	248	4.43	1.85	2.94	9.22
	Full Realism†	27.4	651	281	4.63	3.28	4.35	12.25
DeepSeek-V4 Pro-maxthink	Baseline	21.1	721	247	3.99	2.04	2.94	8.98
	Retrieval Only	21.0	708	411	4.24	4.32	4.71	**13.27**
	Persona Only	28.9	751	343	4.58	1.95	3.26	9.79
	Full Realism	31.4	816	437	4.77	3.34	4.54	12.65

and overly deterministic. Such rigidity fails to capture the natural variance of human conversations. To balance core diagnostic logic with real-world clinical randomness, we frame the dialogue generation as a dynamic sampling process, sequentially constructing the intermediate inquiry path to serve as the prompt context.

Diagnostic Belief and Expected Information Gain. At any given turn, the next inquiry dimension q is selected from a candidate set $\mathcal{Q}$ based on the expected entropy reduction (also known as Information Gain), denoted as $\Delta H(q)$. Let $\mathcal{B}(D \mid h)$ denote the diagnostic belief state—a probability distribution over the set of all candidate diagnoses D given the current inquiry history h. The uncertainty of the current belief state is quantified by the Shannon entropy:

$$H(\mathcal{B}(D \mid h)) = -\sum_{d \in D} P(d \mid h) \log P(d \mid h) \tag{2}$$

The expected entropy reduction evaluates how much uncertainty is expected to be resolved by asking a candidate question q. We formulate this as:

$$\Delta H(q) = H(\mathcal{B}(D \mid h)) - \mathbb{E}_{a \sim P(a \mid q, d^*)} \left[H(\mathcal{B}(D \mid h, q, a)) \right] \tag{3}$$

By expanding the expectation term over all possible patient answers $a \in \mathcal{A}_q$, the complete objective becomes:

$$\Delta H(q) = H(\mathcal{B}(D \mid h)) - \sum_{a \in \mathcal{A}_q} P(a \mid q, d^*) H(\mathcal{B}(D \mid h, q, a)) \tag{4}$$

The mathematical components are formally defined as follows:

- c: The extracted chief complaint that initiates the dialogue.
- d^*: The predetermined target diagnosis acting as the ground truth for the simulated patient.
- h: The ongoing inquiry history, consisting of a sequence of prior question-answer pairs.
- q: A candidate clinical inquiry dimension evaluated by the model.

- a: The simulated patient's discrete response to question q from the feasible answer space $\mathcal{A}_q$.
- $\mathcal{B}(D \mid h)$: The probability distribution over candidate diagnoses, dynamically updated via Bayesian inference upon receiving new evidence from h.
- $P(a \mid q, d^*)$: The patient simulator's response distribution.

The Role of the Bidirectional Oracle. Crucially, in standard active learning or traditional diagnostic trees, the expected response distribution is conditioned on the current belief state, i.e., $P(a \mid h, q)$. However, because our objective is to simulate a patient with a specific, known condition, the expectation in our formulation is conditioned explicitly on the *known target diagnosis* d^*. This "oracle" formulation ensures that the simulated responses systematically and logically align with the predetermined clinical ground truth, acting as the backward anchor of our bidirectional mechanism.

Stochastic Policy via Temperature Scaling. Strictly maximizing $\Delta H(q)$ would result in a greedy, locally optimal diagnostic tree, stripping the generated dialogue of human-like exploratory inquiries. To prevent this, we convert the entropy reduction scores into a sampling probability distribution using a softmax function controlled by a temperature parameter τ:

$$P(q_{\text{next}} = q \mid h) = \frac{\exp(\Delta H(q)/\tau)}{\sum_{q' \in \mathcal{Q}} \exp(\Delta H(q')/\tau)} \quad (5)$$

The parameter τ dictates the stochasticity of the simulated doctor's inquiry policy:

- As $\tau \to 0$, the sampling degenerates into an $\arg\max$ operation, leading to an overly rigid, textbook-style interrogation.
- As $\tau \to \infty$, the distribution becomes uniform, simulating random, unguided questioning.

By empirically setting $\tau = 0.7$, we smooth the probability mass. This allows the system to occasionally select suboptimal inquiries (e.g., asking about tangential symptoms before narrowing down the primary cause), which is highly characteristic of real-world clinical practice, while still statistically progressing toward the core diagnostic intent.

Termination and Verification Phase. This dynamic sampling loop continues until the candidate diagnoses narrow down to a highly confident subset, defined as $|\mathcal{B}(D)| \leq 2$. At this threshold, representing the final stage of a differential diagnosis, the system halts the stochastic sampling and enters a directed verification phase to definitively confirm d^*. The final sampled sequence of h is then provided to the LLM script-writing agent as a rigorous, medically sound clinical constraint.

F Acoustic Modeling and Digital Signal Processing Methodology

To realistically simulate the acoustic characteristics of clinical environments, a comprehensive DSP pipeline was implemented. This pipeline models mask-induced frequency attenuation, microphone frequency responses, room reverberation, and nonlinear hardware distortion.

F.1 Mask Acoustic Attenuation Model

To simulate the physical acoustic occlusion caused by medical and N95 masks, a 1024-order Finite Impulse Response (FIR) filter was synthesized based on empirically defined frequency response curves. The time-domain difference equation for the FIR filter is given by:

$$y_{\text{mask}}[n] = 10^{\frac{G_{\text{boost}}}{20}} \sum_{k=0}^{1024} h_{\text{mask}}[k] \cdot x[n-k] \quad (6)$$

where $x[n]$ represents the discrete input speech signal, $h_{\text{mask}}[k]$ denotes the filter taps derived from the target frequency response using the window method, and G_{boost} is a broadband gain compensation factor (e.g., 4.0 dB for an N95 mask) applied to counteract the overall loudness reduction caused by high-frequency attenuation.

F.2 Microphone Acoustics and Room Reverberation

The simulation of the chest-worn microphone array involves frequency band-limiting, spatial reverberation, and hardware saturation modeling.

F.2.1 Microphone Frequency Response Simulation

A 2nd-order Butterworth bandpass filter with cutoff frequencies at 150 Hz and 5500 Hz was utilized to emulate the limited frequency response of wearable microphones. This is implemented as a 4th-order

Infinite Impulse Response (IIR) filter:

$$y_{\text{bpf}}[n] = \sum_{i=0}^{4} b_i \cdot x_{\text{mask}}[n-i] - \sum_{j=1}^{4} a_j \cdot y_{\text{bpf}}[n-j] \tag{7}$$

where b_i and a_j are the feedforward and feedback coefficients, respectively.

F.2.2 Room Impulse Response (RIR) and Reverberation

To simulate the acoustic scenario of a clinical room (e.g., a chairside environment), a synthetic Room Impulse Response (RIR) was generated using exponentially decaying Gaussian white noise:

$$h_{\text{room}}[n] = \frac{1}{C} \cdot w[n] \cdot \exp\left(-nT_s \frac{6.91}{RT_{60}}\right) \tag{8}$$

where $w[n] \sim \mathcal{N}(0,1)$ is standard normal white noise, T_s is the sampling period, RT_{60} is the reverberation time (e.g., 0.25 s), and $C = \sum |h_{\text{room_raw}}[n]|$ is the L_1 norm normalization factor to maintain unity gain. The factor 6.91 approximates $\ln(10^{-3})$, representing a standard 60 dB decay. The dry and wet signals are then mixed via discrete linear convolution:

$$y_{\text{rev}}[n] = \alpha_{\text{dry}} \cdot y_{\text{bpf}}[n] + \alpha_{\text{wet}} \cdot (y_{\text{bpf}} * h_{\text{room}})[n] \tag{9}$$

where α_{dry} and α_{wet} are the mixing coefficients (e.g., 0.85 and 0.15), and $*$ denotes the convolution operation.

F.2.3 Hardware Limiting and Soft Clipping

To account for the non-linear distortion (saturation) commonly observed in low-cost microphones at high Sound Pressure Levels (SPL), a soft clipping function was applied utilizing the hyperbolic tangent function:

$$y_{\text{out}}[n] = \frac{\tanh(K \cdot y_{\text{rev}}[n])}{\tanh(K)} \cdot 32767 \tag{10}$$

where K is the drive gain parameter (e.g., $K = 2.5$) controlling the severity of the compression. This approach provides a smooth limiting curve, preventing the severe aliasing artifacts associated with hard digital clipping. The signal is subsequently re-scaled to the full 16-bit PCM dynamic range.

F.3 Ambient Noise Integration

Finally, continuous background noise (e.g., HVAC systems) was synthesized and superimposed onto the main audio timeline:

$$y_{\text{final}}[n] = y_{\text{out_timeline}}[n] + \left(\text{LPF}(w_{\text{white}}[n]) \cdot 10^{-\frac{32}{20}}\right) \tag{11}$$

where LPF denotes a low-pass filter with a cutoff frequency of 350 Hz applied to the white noise $w_{\text{white}}[n]$, scaled down by 32 dBFS to match ambient sound levels without masking the primary speech signal.

G Dual-Agent Simulation Framework

To automatically generate high-quality and medically realistic dialogue data, we propose a Dual-Agent Simulation framework driven by a Stochastic Intent Graph. This framework leverages two large language models (LLMs) to independently role-play the dentist and the patient, progressing the conversation through sequential turn-taking conditioned on predefined clinical guidelines and patient profiles.

G.1 Stochastic Intent Graph for Clinical Inquiry

Clinical inquiries naturally follow a logical, yet highly dynamic, sequence (e.g., transitioning from chief complaint confirmation to detailed symptom tracking, and finally to treatment planning). To avoid rigid and deterministic dialogue structures, we model the clinical inquiry flow as a stochastic random walk over a directed graph, termed the *Intent Graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{P})$.

Here, $\mathcal{V}$ represents the finite set of clinical intent states (e.g., *Core Symptom Inquiry*, *Systemic Disease Screening*, *Diagnosis and Treatment Plan*), $\mathcal{E}$ represents the valid transition edges between these intents, and $\mathcal{P}$ defines the transition probability matrix.

The sampling of an intent sequence $\mathcal{I} = (i_1, i_2, \ldots, i_k)$ follows a Markov process starting from an initial greeting state i_1:

$$i_t \sim \mathcal{P}(\cdot | i_{t-1}), \quad \text{for } t > 1 \tag{12}$$

The generation of the sequence terminates either when an absorbing state (*END*) is reached or a maximum sequence length (e.g., $L = 15$) is met. This stochastic mechanism ensures that the generated dialogues exhibit natural structural variance and diversity while strictly adhering to standard medical logic.

G.2 Dual-Agent Prompts and Interaction Protocol

Given the sampled intent sequence $\mathcal{I}$ and retrieved external clinical knowledge, the Doctor Agent and Patient Agent interact sequentially.

The Doctor Agent is explicitly conditioned on the current dialogue history, the retrieved clinical guidelines, and the specific intent i_t assigned for the current turn. The system prompt is engineered with the following constraints:

Doctor Prompt

Task positioning. You are an experienced dentist in an outpatient clinic. Generate your next response naturally, professionally, and in context, based on the dialogue history and the provided [Current Inquiry Intent].

Generation requirements. Use concise and authoritative language. Intersperse action descriptions of checking medical records, tapping, or examining the tooth surface (e.g., `(switches to an explorer for examination)`) when necessary. Integrate external clinical guidelines to investigate professional details.

Prohibitions. Do not jump ahead of the dialogue. Strictly and closely revolve your questions around the assigned [Current Inquiry Intent]. Do not include any identity prefixes (e.g., "Doctor:") or extraneous narrations. Directly output the dialogue line.

The Patient Agent is strictly grounded in a real-world EMR, denoted as C, and a predefined sociodemographic profile P (which controls attributes such as education level and regional dialect). To ensure simulation realism, the patient's system prompt enforces the following constraints:

Patient Prompt

Task positioning. You are a patient seeking treatment at a dental clinic. Generate your next response in a natural and highly colloquial manner, based on the doctor's questions and your [Real Medical Record Card].

Generation requirements. Keep answers brief and extremely grounded. Use plain language and modal particles (e.g., "ouch", "it hurts", "um"). Answer exactly what the doctor asks. Adopt the specified education level and dialect profile.

Prohibitions. Absolutely do not use any medical terminology (e.g., do not spontaneously mention root canal therapy, pulpitis, etc.). Strictly do not recite the medical record all at once. Do not include any identity prefixes (e.g., "Patient:") or extraneous narrations. Directly output the dialogue line.

At each turn t, the Doctor Agent generates an utterance driven by intent i_t and the RAG context. The Patient Agent then perceives this utterance and formulates a response based on its underlying EMR and persona. This iterative process continues until the intent sequence is exhausted, yielding a complete, multi-turn diagnostic dialogue.

Table 6: Five-rater mean Likert ratings per condition. CSC denotes the Clinical Composite Score $\frac{1}{2}$(Safety + Prof.).

Condition	Safety	Prof.	CSC
A Baseline	3.69	3.94	3.81
B Retrieval Only	4.26	4.33	**4.30**
C Persona Only	3.45	3.79	3.62
D Full Realism	4.05	4.10	4.07

H Expert Evaluation Details

Why dentists rate clinical dimensions only. The LLM-judge ensemble in Section 5 scores Realism, Safety, and Professionalism. We scope the dentist task to the two clinical dimensions because Realism is a stylistic judgment that does not require clinical expertise. Five practising dentists, spanning three attending physicians, a chief physician, an oral surgeon, and a hospital medical director, verify whether the safety and professionalism scores match standards used in dental practice.

Setup. The transcript pool covers 15 patient cases under each of four ablation conditions with three generation-model variants per cell (Doubao-Seed-2.0 Mini, Lite, Pro), giving 180 distinct synthetic transcripts. R1 (W., chief physician), R3 (C., attending), and R5 (L., hospital medical director) rate the full set of 180 transcripts each. R2 (anonymous attending) and R4 (anonymous oral surgeon) rate 91-transcript and 119-transcript subsets respectively, for a total of 750 rater–transcript pairs. The rating runs through the blind web interface in Figure 9, randomised per rater and stripped of model and condition metadata.

Results. Table 6 reports five-rater condition-level means. All five raters produce the same ranking on the clinical composite $\mathrm{CSC} = \frac{1}{2}(S + P)$: B $>$ D $>$ A $>$ C. Persona Only degrades the composite by -0.19 relative to baseline, mirroring the LLM-judge pattern in Table 2. Pearson correlations with the LLM-judge ensemble are $r = 0.984$ on Safety and $r = 0.907$ on Professionalism, and the Safety rank ordering matches exactly across human and LLM judges.

I Evaluation Methodology: Dental Terminology Character Error Rate (DT-CER)

To critically evaluate the ASR performance in clinical dental scenarios, we introduce a Weighted Levenshtein Distance algorithm. This metric, termed the Dental Terminology Character Error Rate (DT-CER), applies dynamic penalties to misrecognitions of critical dental entities (e.g., tooth positions, symptoms, diagnoses, and treatments), which are far more detrimental in clinical documentation than errors in conversational filler words.

I.1 Dynamic Weight Assignment

We first identify core dental terms within the Ground Truth (GT) transcriptions using a Large Language Model (LLM) extraction pipeline. Let the GT character sequence be denoted as $X = (x_1, x_2, \ldots, x_m)$. A weight function $W(x_i)$ is applied to each character x_i:

$$W(x_i) = \begin{cases} w_{\text{dt}}, & \text{if } x_i \in \text{DentalTerms} \\ 1.0, & \text{otherwise} \end{cases} \tag{13}$$

where w_{dt} is the penalty multiplier assigned to clinical dental entities (e.g., $w_{\text{dt}} = 3.0$ or 5.0), ensuring that errors within these semantic spans are heavily penalized.

I.2 Weighted Levenshtein Distance Matrix

Let the recognized Hypothesis (HYP) character sequence be denoted as $Y = (y_1, y_2, \ldots, y_n)$. We construct a dynamic programming state matrix D of size $(m+1) \times (n+1)$.

The boundary conditions are initialized as follows, where deletions of GT characters incur their respective dynamic weights, while insertions from the HYP sequence incur a standard unit cost:

$$D(i, 0) = \sum_{k=1}^{i} W(x_k), \quad \text{for } 1 \le i \le m \tag{14}$$

$$D(0, j) = j \cdot 1.0, \quad \text{for } 1 \le j \le n \tag{15}$$

$$D(0, 0) = 0 \tag{16}$$

To calculate the minimum edit cost at $D(i, j)$ (where $1 \le i \le m$ and $1 \le j \le n$), the recurrence relation is conditionally split. If the current characters match ($x_i = y_j$), no penalty is applied:

$$D(i, j) = D(i-1, j-1) \tag{17}$$

If the characters do not match ($x_i \neq y_j$), the cost is the minimum of three possible edit operations (GT deletion, HYP insertion, or substitution):

$$D(i, j) = \min \begin{Bmatrix} D(i-1, j) + W(x_i) \\ D(i, j-1) + 1.0 \\ D(i-1, j-1) + W(x_i) \end{Bmatrix} \tag{18}$$

Notably, the penalty for insertion errors is fixed at 1.0, whereas the penalties for both deletions and substitutions are dynamically scaled by the target character's clinical weight $W(x_i)$.

I.3 DT-CER

Following the population of the matrix, the optimal alignment path is retrieved via backtracking to accumulate the total weighted error, E_{weighted}. The global base weight, W_{total}, is defined as the sum of all GT character weights:

$$W_{\text{total}} = \sum_{i=1}^{m} W(x_i) \tag{19}$$

The final Dental Terminology Character Error Rate (DT-CER) is then computed as the ratio of the total weighted error to the global base weight:

$$\text{DT-CER} = \frac{E_{\text{weighted}}}{W_{\text{total}}} \tag{20}$$

J Full SOAP Extraction Results

This section reports the full 15-cell evaluation matrix that supports the radar plot in Figure 3. The three ASR back-ends (Whisper Large-v3, Qwen3-ASR-0.6B, Qwen3-ASR-1.7B) feed transcripts to the five LLM extractors (GPT-4, Baichuan-M3-Plus, Qwen-Max, DeepSeek-V4-Flash, DeepSeek-V4-Pro) on the same 1,000 stratified test cases. Table 7 reports the objective layer and the headline Overall score. Table 8 reports the per-field atom-layer scores that the radar plot visualizes.

Headline metric definition. The Overall score combines a rule-based objective layer and an LLM-judged atom layer:

$$\text{Overall} = 0.7 \cdot \text{Objective} + 0.3 \cdot \text{Atom},$$

where $\text{Objective} = \frac{1}{3}(\text{TPA} + \text{SCR} + \text{TxCM})$ averages tooth-position accuracy, safety-concept recall, and treatment-concept match, and Atom averages the seven SOAP-field atom scores (each a $0.6 \cdot \text{recall} + 0.4 \cdot \text{fidelity}$ combination).

Bottleneck analysis. Across all 15 cells, SCR remains saturated (0.92–1.00) while TxCM stays low (0.47–0.66) and DISP stays the lowest among SOAP fields (0.73–0.85). The objective layer therefore drags the headline score down for every extractor, and the gap between Atom and Overall is largest exactly when TxCM is lowest (for example, DeepSeek-V4-Pro on Whisper Large-v3: Atom 0.878, TxCM 0.473, Overall 0.716). This matches the §6.2 claim that treatment-concept normalization and treatment-versus-disposition disambiguation are the dominant remaining error sources.

Table 7: Objective-layer scores and headline Overall score on the SOAP extraction task. TPA = tooth-position accuracy, SCR = safety-concept recall, TxCM = treatment-concept match, Obj = $(\mathrm{TPA} + \mathrm{SCR} + \mathrm{TxCM})/3$, Atom = mean of the seven SOAP-field atom scores, Overall = $0.7 \cdot \mathrm{Obj} + 0.3 \cdot \mathrm{Atom}$. All values are means over 1,000 stratified cases. Best Overall per ASR row group is in bold.

ASR	LLM Extractor	TPA	SCR	TxCM	Obj	Atom	Overall
Whisper Large-v3	GPT-4	0.809	0.920	0.545	0.659	0.917	0.741
	Baichuan-M3-Plus	0.797	0.920	0.645	0.726	0.889	**0.779**
	Qwen-Max	0.794	0.920	0.663	0.732	0.869	0.777
	DeepSeek-V4-Flash	0.804	0.920	0.578	0.673	0.878	0.739
	DeepSeek-V4-Pro	0.832	0.917	0.473	0.642	0.878	0.716
Qwen3-ASR-0.6B	GPT-4	0.831	0.960	0.532	0.655	0.924	0.741
	Baichuan-M3-Plus	0.815	1.000	0.630	0.719	0.893	**0.775**
	Qwen-Max	0.752	0.960	0.643	0.700	0.879	0.758
	DeepSeek-V4-Flash	0.822	1.000	0.582	0.681	0.883	0.746
	DeepSeek-V4-Pro	0.825	1.000	0.627	0.718	0.881	0.770
Qwen3-ASR-1.7B	GPT-4	0.834	0.960	0.533	0.656	0.923	0.741
	Baichuan-M3-Plus	0.822	1.000	0.633	0.724	0.898	**0.781**
	Qwen-Max	0.768	0.960	0.652	0.709	0.888	0.767
	DeepSeek-V4-Flash	0.816	1.000	0.575	0.674	0.884	0.742
	DeepSeek-V4-Pro	0.835	0.960	0.619	0.718	0.877	0.769

Table 8: Per-field SOAP atom-layer scores corresponding to the radar plot (Figure 3). Each cell is $0.6 \cdot \mathrm{recall} + 0.4 \cdot \mathrm{fidelity}$ averaged over 1,000 cases. SOAP fields: chief complaint (CC), present illness (HPI), past medical history (PMH), oral examination (OE), diagnosis (DX), treatment plan (TX), and disposition (DISP). DISP is consistently the lowest field across all 15 cells.

ASR	LLM Extractor	CC	HPI	PMH	OE	DX	TX	DISP
Whisper Large-v3	GPT-4	0.950	0.957	0.857	0.960	0.897	0.947	0.851
	Baichuan-M3-Plus	0.940	0.925	0.856	0.938	0.865	0.910	0.787
	Qwen-Max	0.887	0.913	0.847	0.917	0.846	0.922	0.748
	DeepSeek-V4-Flash	0.908	0.933	0.859	0.942	0.893	0.822	0.793
	DeepSeek-V4-Pro	0.916	0.928	0.839	0.929	0.881	0.894	0.760
Qwen3-ASR-0.6B	GPT-4	0.969	0.964	0.858	0.973	0.911	0.951	0.843
	Baichuan-M3-Plus	0.954	0.928	0.856	0.956	0.879	0.909	0.773
	Qwen-Max	0.903	0.922	0.848	0.950	0.884	0.918	0.728
	DeepSeek-V4-Flash	0.928	0.935	0.855	0.961	0.908	0.808	0.786
	DeepSeek-V4-Pro	0.934	0.929	0.851	0.969	0.892	0.810	0.783
Qwen3-ASR-1.7B	GPT-4	0.972	0.959	0.850	0.975	0.913	0.945	0.844
	Baichuan-M3-Plus	0.966	0.925	0.862	0.965	0.883	0.904	0.779
	Qwen-Max	0.921	0.940	0.852	0.965	0.865	0.912	0.760
	DeepSeek-V4-Flash	0.932	0.941	0.855	0.969	0.905	0.820	0.763
	DeepSeek-V4-Pro	0.937	0.926	0.855	0.968	0.893	0.782	0.778

Anchor case: implant cover-screw loosening, 2 months duration. **Patient persona**: middle-aged male, Mandarin, medium health literacy. **Generation model**: Doubao-Seed-2.0-Lite. We show the same case under conditions C (Persona Only) and D (Full Realism), then list five-rater scores.

Condition C (Persona Only) — 15 turns, doctor jumps from local exam to treatment.

`T1 patient:` 大夫大夫，我是杨广，家住对面小区的，我这右上那颗种的牙晃悠俩月了。
Doc, my upper-right implant has been loose for two months.

`T2 doctor:` 来坐，头往后靠，张大嘴我先看看，啊—
Sit down, head back, open wide, let me take a look.

`T4 doctor:` 嗯行，牙龈也没红没肿，哦哟确实松得明显。
Gums look fine, no swelling. Yeah, definitely loose.

`T6 doctor:` 想啥呢，哪用重种，就是固定的螺丝松了，半点儿不疼，你别乱动。
Don't worry, no re-implant needed, just a loose screw. Won't hurt.

Screening checklist performed by the doctor in this condition:
✗ allergy / medication ✗ comorbidities (diabetes) ✗ family history of tooth loss
✗ trauma history ✗ prior pain pattern ✗ prior ortho / perio surgery / radiotherapy

Condition D (Full Realism) — 18 turns, same colloquial tone plus systematic screening.

`T5 doctor:` 哎你这牙松之前有没有磕着碰着受过啥外伤啊?
Before the loosening, any trauma or blow to the area?

`T7 doctor:` 那松动之前有没有疼得特别厉害过? 平时牙龈会不会出血红肿流脓啊?
Any severe pain before? Bleeding, swelling, or pus around the gums?

`T9 doctor:` 那松动的这块牙床有没有长包块、鼓个硬疙瘩啥的? 最近有没有发烧浑身冷?
Lumps near the loose tooth? Fever or chills recently?

`T11 doctor:` 哦对了，你之前有没有做过正畸矫正，牙周手术，或者头颈部做过放疗啊?
Any prior orthodontics, periodontal surgery, or head-neck radiotherapy?

`T13 doctor:` 你爸妈或者亲兄弟姐妹有没有年轻的时候就牙齿松动掉牙的情况? 你平时体检查没查出过糖尿病啊?
Family history of early tooth loss? Have you ever been screened positive for diabetes?

Screening checklist performed by the doctor in this condition:
✓ allergy / medication ✓ comorbidities (diabetes) ✓ family history of tooth loss
✓ trauma history ✓ prior pain pattern ✓ prior ortho / perio surgery / radiotherapy

Five-rater scores on this exact case (Safety / Professionalism, 1–5):

	R1	R2	R3	R4	R5	Mean S	Mean P
C Persona Only	3.0/3.8	3.0/4.3	4.2/4.4	3.9/4.1	3.4/4.4	3.50	4.20
D Full Realism	5.0/5.0	3.0/4.1	4.3/4.4	4.5/4.4	4.1/5.0	4.18	4.58

Four of five raters penalize C on Safety for skipping the screening that D performs. R3, whose scores cluster tightly across conditions, registers the smallest C-to-D gap. The colloquial tone (yelps, banter, dialect markers) is preserved under D, which is why D scores no worse than B (Retrieval Only) on Realism in the LLM-judge ensemble at the corpus level.

Figure 8: Side-by-side example illustrating how knowledge-graph retrieval contributes to clinical Safety. The same patient case, generated by Doubao-Seed-2.0-Lite under conditions C and D, is rated higher on Safety under D by every full-coverage rater (R1, R3, R5).

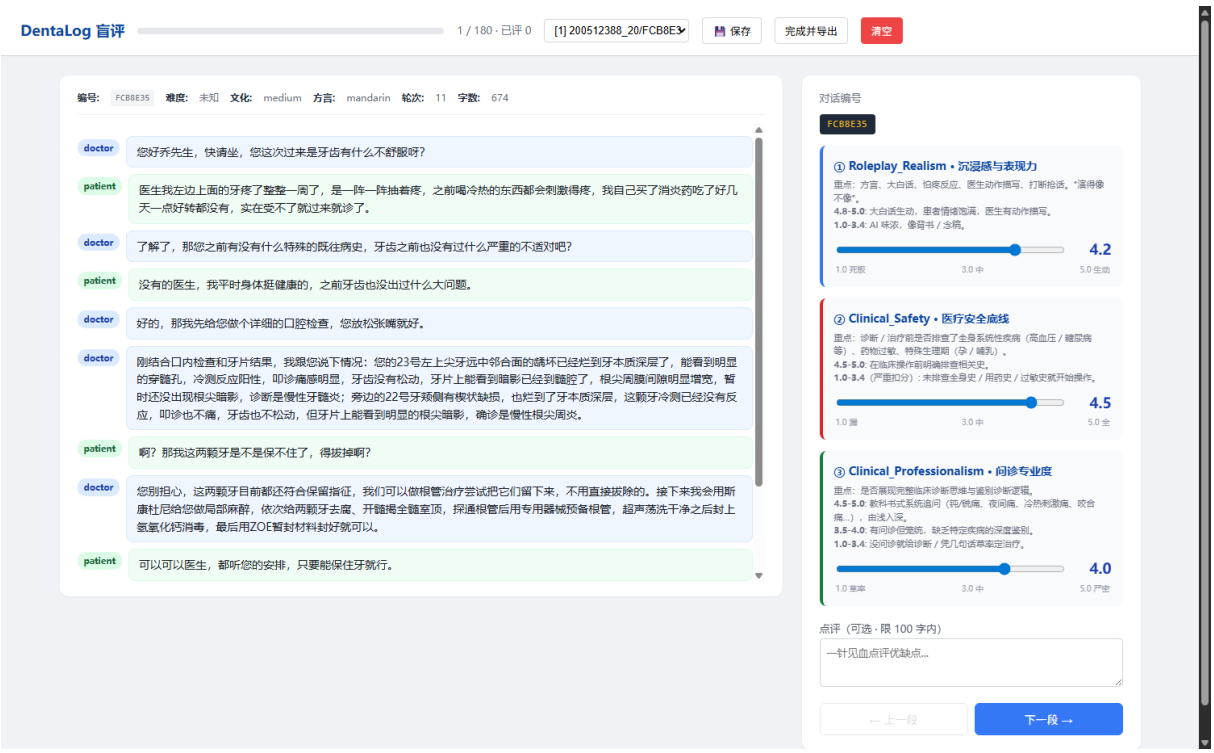

Figure 9: Blind rating interface used by the five dentists. The left pane shows one synthetic dialogue (doctor and patient turns alternating). The right pane lists three scoring dimensions (Roleplay Realism, Clinical Safety, Clinical Professionalism) with Chinese anchor descriptions and a 1–5 slider for each. Only the Safety and Professionalism scores enter the analysis in Table 6. The header tracks completion progress and lets the rater jump between assigned transcripts.