

Generating Synthetic Doctor-Patient Conversations for Long-form Audio Summarization

Yanis Labrak¹, David Grünert², Severin Baroudi^{1,4}, Jiyun Chun³, Pawel Cyrta⁵, Sergio Burdisso¹, Ahmed Hassoon⁶, David Liu⁷, Adam Rothschild⁸, Reed Van Deusen⁹, Petr Motlicek¹, Andrew Perrault³, Ricard Marxer^{4,10}, Thomas Schaaaf^{11,12}

¹ Idiap Research Institute, Switzerland ² University of Zurich, Switzerland

³ The Ohio State University, USA ⁴ Université de Toulon, Aix Marseille Univ, LIS, CNRS, France

⁵ Stenograf, Poland ⁶ Johns Hopkins University Bloomberg School of Public Health, USA

⁷ Colorado School of Mines, USA ⁸ Allegheny Health Network, USA

⁹ University of Pittsburgh Medical Center, USA ¹⁰ ILLS, CNRS, France

¹¹ Solventum, USA ¹² Carnegie Mellon University, USA

Abstract

Long-context audio reasoning is underserved in both training data and evaluation. Existing benchmarks target short-context tasks, and the open-ended generation tasks most relevant to long-context reasoning pose well-known challenges for automatic evaluation. We propose a synthetic data generation pipeline designed to serve both as a training resource and as a controlled evaluation environment, and instantiate it for first-visit doctor-patient conversations with SOAP note generation as the task. The pipeline has three stages—persona-driven dialogue generation, multi-speaker audio synthesis with overlap/pause modeling, room acoustics, and sound events, and LLM-based reference SOAP note production—built entirely on open-weight models. We release 8,800 synthetic conversations with 1.3k hours of corresponding audio and reference notes. Evaluating current open-weight systems, we find that cascaded approaches still substantially outperform end-to-end models.

Index Terms: end-to-end spoken language understanding, long-form audio, multi-speaker environment

learning with verifiable rewards) and as a controlled evaluation environment.

We make the following contributions:

- A full synthetic pipeline—persona-based dialogue generation, audio synthesis featuring audio synthesis featuring overlap/pause modeling, sound event insertion, and room acoustic simulation and reference SOAP note production—built entirely with open-weight, permissively licensed models and developed in consultation with medical doctors.
- A dataset of 8,800 synthetic doctor-patient conversations with corresponding audio recordings and reference SOAP notes.
- An evaluation of current open-weight cascaded and end-to-end systems on the audio-to-SOAP-note task, finding that cascaded systems substantially outperform end-to-end approaches, with the cascaded pipeline achieving near-ceiling ASR performance.

We will release all data and code to support future work on long-context audio training and real-world validation.

1. Introduction

Toward the broader goal of human-level audio understanding, recent large audio language models (LALMs) [1, 2, 3, 4] have demonstrated impressive progress on benchmarks for audio processing and comprehension [5, 6, 7, 8]. However, these benchmarks focus predominantly on short-context tasks. Our understanding of LALM performance on long-context audio reasoning (> 5 minutes) remains limited, in part because LALMs capable of accepting long-context input only emerged in mid-2025, and in part because constructing meaningful benchmarks for such tasks is difficult.

Beyond data scarcity, evaluation itself is a bottleneck. Automatically constructed benchmarks risk covering only a narrow slice of audio understanding, and the tasks most relevant to long-context reasoning—such as summarization and note-taking—require open-ended generation, where multiple valid outputs exist for any given source, surface overlap metrics correlate weakly with human quality judgments [9], and fluent outputs can hallucinate content not present in the source [10]. Rather than contributing another benchmark, we propose a synthetic data generation pipeline that can serve both as a source of training signal (via supervised fine-tuning or reinforcement

2. Related Work

Recent advances have dramatically expanded the context windows of Large Audio Language Models (LALMs). While early attempts at end-to-end (E2E) speech summarization struggled with the quadratic memory complexity of processing long audio sequences [11], current systems can ingest continuous audio ranging from 40 minutes to over eight hours [1, 2]. However, the capacity to process long-form audio does not equate to the ability to reason over it. Recent evaluations, including MMAU-Pro [8] and BLAB [6], reveal severe degradation on long-sequence multi-hop reasoning, a “modality reasoning gap” [12] that manifests as representational drift and “lexical dominance” [7] when processing speech compared to text. E2E architectures still significantly trail cascaded ASR-to-text systems on complex open-ended generation tasks [13], a gap our dataset is designed to probe.

Automating clinical documentation, such as SOAP note generation, is a historically challenging task traditionally addressed through cascaded ASR-to-text systems. The primary bottleneck for advancing end-to-end multi-modal models in this domain is the severe scarcity of conversational data. While pro-

proprietary systems have leveraged thousands of hours of real clinical audio [14, 15], strict privacy regulations like HIPAA prevent the distribution of these datasets to the open-source community [16, 17]. Consequently, public benchmarks have been heavily restricted to small-scale datasets like PriMock57 [18], which contains only 57 brief, actor-performed consultations that lack the organic complexity of genuine encounters, while ACI-Bench [19] provides 187 encounters in text form but no audio.

To circumvent these privacy barriers, recent research has successfully turned to Large Language Models (LLMs) to generate synthetic medical text. Multi-agent frameworks like NoteChat [17] and single-prompt LLM simulators [16] have demonstrated that models can produce highly realistic, medically accurate text transcripts of patient-physician interactions. However, these approaches remain strictly text-bound, failing to address the acoustic realities of medical settings, e.g., multi-speaker overlap, far-field microphone reverberation, and environmental noise [15]. Our pipeline extends these text-only approaches to full multi-speaker audio synthesis, advancing beyond earlier work combining LLMs and TTS for conversational ASR [20] by integrating persona-conditioned voices and acoustic simulation to cross the significant “sim2real” gap of the clinical domain. Given the inadequacy of surface metrics for clinical summarization [9], our evaluation adopts an LLM-as-a-judge framework [21, 22].

3. Transcript and Audio Generation

We now describe the three stages of our data generation pipeline, each targeting a specific gap identified above: (1) persona and context sampling, (2) persona-conditioned text dialogue generation, and (3) audio synthesis with acoustic simulation. Throughout, we use Gemma3-27B-IT [23] as the LLM, selected for its performance in early dialogue generation tests, and SDialog [24] for pipeline orchestration.

3.1. Sampling Personas

Directly prompting an LLM to generate a first-visit conversation results in low diversity: out of 2,800 dialogues, we found only three unique doctor-patient pairs, with most sharing the same reason for visit and interpersonal dynamics. To increase diversity, we sample structured personas — lists of discrete attributes — for the doctor and patient before generation. We selected attributes that are few enough for the LLM to follow reliably, impactful enough to alter register or clinical content, minimally conflicting, and samplable from external demographic distributions to avoid internal LLM bias.

Both the doctor and patient share the following attributes: name, age, height, weight, race, gender, forgetfulness, formality, hurriedness, and marital status. The patient has: fluency in English, occupation, insurance, and reason for visit. The doctor has years of experience. All attributes are sampled uniformly from predefined lists, using publicly available sources, except the reason for visit, which comprises 724 chief complaints compiled by a co-author with clinical expertise using an expert elicitation technique to collect a unique list of chief complaints across an entire healthcare system. These complaints cover typical and atypical primary care presentations that are clinical or administrative in nature (e.g., “hand swelling,” “unexplained bruises on legs”, “medical clearance”), contain no personally identifiable information, and will be released with the code.

Table 1: *Dialogue statistics comparing our generated synthetic conversations against baseline and reference datasets.*

Dataset	Doctor		Patient		Num Turns
	Turn Length	Fog Index	Turn Length	Fog Index	
Ours	49.9 ± 16.0	10.4	56.0 ± 15.2	6.9	28.4 ± 7.5
PriMock57	18.8 ± 4.7	6.6	12.3 ± 5.8	5.5	97.3 ± 17.7
Mocks	29.6 ± 6.3	7.0	13.6 ± 3.1	7.0	54 ± 6.2

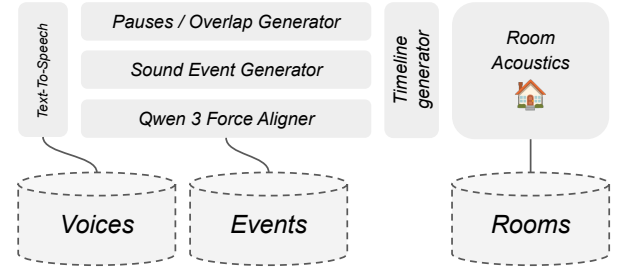


Figure 1: *Overview of the audio generation pipeline, integrating Text-To-Speech, sound event generation, temporal scene composition, and environmental acoustic simulation.*

3.2. Personas → Text Dialogue

We experimented with a direct generation approach, where the generator LLM receives both the doctor and patient personas as input and generates the entire dialogue in a single shot. We found that doing so led to unrealistically smooth dialogues, where the doctor quickly identified and handled the patient’s reason for visit. We thus opted for a multi-turn generation approach, where the generator LLM produces a single turn of dialogue at a time for the doctor (resp. patient), and receives the dialogue history so far as well as the doctor’s (resp. patient’s) persona. On a subset of 2,800 dialogues, the multi-turn approach increased Claude Sonnet 4¹ ratings of interpersonal challenge (1–5 scale) from 2.01 (CI: 1.98–2.04) to 2.50 (CI: 2.46–2.54). Persona adherence ratings likewise increased, from 3.10 (CI: 3.07–3.13) to 3.30 (CI: 3.26–3.34) (doctor) and from 3.60 (CI: 3.58–3.62) to 3.90 (CI: 3.88–3.92) (patient). We hypothesize this is because the generator LLM can focus on the active role more specifically. Multiple rounds of physician feedback were collected on the text dialogue generation, leading to iterative refinement of the personas and prompts.

As Table 1 shows, our generated doctors exhibit higher language complexity (fog index 10.4 vs. 6.6–7.0) and longer turns than either reference set, likely because Gemma 3 was trained primarily on written text. PriMock57’s telehealth setting further explains its higher turn count and shorter turn length relative to our in-person Mocks.² We used direct prompting to encourage conversational style, emphasizing the in-person spoken setting, which had a moderate effect and is used in the final system. Few-shot examples from Mocks were also tested, but they caused personality and topic leakage without meaningfully reducing complexity and were therefore excluded.

¹dev evaluation only; not in the final pipeline

²Three mock clinical encounter recordings (actor + real physician, realistic setting).

3.3. Text Dialogue → Conversation Audio

Using the generated personas (200 patients, 100 doctors), we synthesized 8,800 unique dialogues using a cross-product. The language models were prompted to incorporate natural conversational artifacts, such as colloquial speech patterns and implicit acoustic triggers (e.g., notations for door knocks or paper rustling). This design introduces a distinct “cross-register” modeling challenge: effectively mapping informal, spontaneous, and acoustically noisy spoken dialogue into formal, structured clinical documentation.

3.3.1. Persona-Conditioned Voice Synthesis

Translating these text dialogues into realistic audio requires a rigorous alignment of acoustic properties with the underlying linguistic personas. For each of the 300 unique personas, we perform conditioned voice cloning using the LibriTTS dataset [25]³ samples of 30s which have been normalized with peak normalization (−1.0 to 1.0). This alignment utilizes gender, which is known in the dataset, to match dialogue personas. For age, an initial attempt was made to estimate speaker age ranges using the Qwen3-Omni model [1]; however, a listening test revealed that the estimated ages were often inaccurate, particularly for elderly voices. Consequently, the initial random age assignment was retained. To maintain consistency and preserve the integrity of the training, development, and test subsets, this voice-to-persona matching is performed once and fixed for all experiments. Crucially, we ensure that the speakers designated for training, development, and testing are strictly disjoint.

3.3.2. TTS Engine Selection and Subjective Assessment

To establish a high-fidelity baseline, we considered nine TTS configurations (including *Chatterbox*, *Kokoro*, *Qwen3*, *XTTS*, and *IndexTTS*) using LibriTTS profiles. *XTTS* was excluded due to licensing restrictions. Qwen3-TTS-1.7B was selected for its balance of linguistic accuracy, expressiveness, and permissive licensing. The selected engine achieves a dry WER of < 2% against the source text; to ensure consistent scoring, we utilize Whisper text normalization⁴ preceded by a custom ASCII normalization step (e.g., standardizing curly quotes) on both the textual utterance ground truths and the hypotheses.

3.3.3. Acoustic Simulation and Flow Enrichment

Real-world clinical encounters occur in complex acoustic environments; we enrich the TTS output through six pipeline stages (Fig. 1): voice reference matching, neural speech synthesis, overlap & pause insertion, sound event insertion, scene timeline composition using the dialogue variant⁵ of scaper [26], and acoustic ray-tracing simulation with PyRoomacoustics [27].

The TTS output is convolved with a Room Impulse Response (RIR) representative of a typical 8m² examination room, followed by the addition of clinical background noise (e.g., HVAC hum). Subsequently, 66 discrete sound event classes, including typing, paper rustling, and door knocks, are superimposed at predicted temporal offsets. These offsets are determined via Qwen3 forced alignment [28] based on acoustic triggers within the dialogue’s stage directions, which are mapped

³Modified version available on HuggingFace upon publication.

⁴github.com/openai/whisper/blob/main/whisper/normalizers

⁵github.com/dscaper/dscaper

to audio event classes using Gemma 3. Finally, conversational dynamics, including natural overlaps and pauses, are computationally injected using Gemma 3 to mimic genuine turn-taking behavior.

3.3.4. Signal Augmentation

To emulate realistic deployment constraints and acoustic conditions, we apply a comprehensive signal augmentation pipeline. We use the Opus codec at 16 kbps, yielding a compression ratio of approximately 14:1 and introducing realistic artifacts typical of real-world recordings. To enhance the perception of depth and compensate for the ray-tracing approach’s limitations on the low end of the spectrum, the patient’s audio amplitude is scaled down by a factor of 4 relative to the doctor’s. Evaluation of the UTMOS metric [29, 30] indicates that our generative framework achieves a score of 1.27, demonstrating a negligible performance gap when compared to our in-person Mocks baseline (1.28) and the DISPLACE-M [31] challenge reference data (1.29). These results indicate that our synthetic output closely approximates the perceptual fidelity of authentic recordings.

3.4. Summary of Dataset

The Synth-DoPaCo⁶ (Synthetic Doctor Patient Conversations) dataset comprises 8,800 synthetic doctor-patient dialogues totaling 1,329 hours of audio. Each dialogue contains on average 28 turns (~14 per speaker), 1,500 words, and is augmented with approximately 37 non-speech audio events (e.g., coughs, background sounds) to simulate realistic clinical encounters. Dialogues average 9 minutes in length (range: 2–47 min). The dataset is split into train, test, and development sets, as summarized in Table 2.

Table 2: Synth-DoPaCo dataset statistics.

	Train	Dev	Test	Total
Personas (Doc/Pat)	60/120	20/20	20/60	100/200
Dialogues	7,200	400	1,200	8,800
Hours	1,087	59	183	1,329
Words in dialogues	10.9M	582K	1.8M	13.3M
Turns/dialogue	28.4	28.7	28.0	28.4
Duration/dialogue (s)	544	530	548	544
Audio events/dialogue	37.7	36.7	36.5	37.5
Words per SOAP note	328.3	325.1	324.2	977.6

On the wet⁷ audio, Whisper Large V3 [32] achieves 2–3% WER while Qwen3-ASR [28] shows 10–14%, confirming the audio is intelligible yet acoustically challenging. The Qwen3-ASR gap is primarily driven by increased substitution and deletion rates, exacerbated by Opus compression.

TTS synthesis was performed on a single NVIDIA A100 40 GB GPU (approximately 2.5k GPU-hours); E2E or cascaded SOAP note generation using Qwen3-32B [33] ran on two NVIDIA A100 40 GB GPUs via Ollama (approximately 300 GPU-hours).

4. SOAP Note Generation and Evaluation

Using the speaker-attributed transcript produced by the pipeline, we generate reference SOAP notes and evaluate all

⁶Audio samples and example SOAP notes are provided as supplementary material for review.

⁷“wet” final augmented audio, optionally Opus-compressed

systems via two-stage processes, both using Kimi K2 Thinking [34]. (Kimi K2 inference was accessed via AWS Bedrock and is not counted in GPU-hours.)

4.1. Reference SOAP Note Generation

We found that even large open-weight reasoning LLMs are prone to hallucinations in long-context text summarization tasks. In particular, (1) they tend to invent natural extensions of what occurred in the dialogue, e.g., a more detailed physical exam or follow-up plan, and (2) they use medical terms that are unsupported, e.g., describing a cold (which could be viral or bacterial) as a “viral illness.” To mitigate these issues, we enforce strict grounding in a fact-extraction stage that produces a structured JSON fact table, in which each fact is linked to a supporting quote and its turn index in the transcript. This fact table serves as the sole input to the note generation, preventing access to the transcript and limiting opportunities for hallucination. In the generation phase, we allow terminology to be rewritten into standard clinical language, but increasing clinical specificity or introducing new clinical content beyond the extracted facts is prohibited. A round of clinical feedback was integrated, including strict separation of the history of present illness (HPI) and review of systems (ROS), tighter integration of the assessment and plan, and prevention of over-documentation by excluding non-clinically relevant administrative content.

4.2. SOAP Note Evaluation with LLM-as-a-Judge

We evaluate SOAP notes using a two-stage LLM-as-a-judge pipeline [35]: atomic claims are first extracted from the note into a structured JSON representation, then each claim is assigned a support label indicating whether it is grounded in the transcript, with explicit evidence required for supported or contradicted claims. We score SOAP notes along 12 dimensions. Four dimensions are scored on a 1–5 scale (5 = best): **Faithfulness** (claim support and contradictions), **Structure** (SOAP formatting and section placement), **Coverage** (completeness of documentation relative to transcript evidence), and **Conciseness** (redundancy and low-value content). Six dimensions are reported as counts: Over-medicalization (unjustified medical interpretation), Under-medicalization (loss of explicitly stated medical specificity), Over-specificity (adds unjustified specificity), Missed relevant facts (missing transcript-grounded key facts), Critical omissions, and Duplicated content across sections. We additionally evaluate two rates: unsupported claim rate and contradiction rate.

4.3. Comparison of Open-Weight LALMs

Using the generated conversation audio and SOAP notes, we compare the quality of the reference notes (using the reference-free LLM-as-a-judge pipeline) and the performance of current open-weight LALMs and cascaded ASR and LLM systems (Tab. 3, showing a subset of judge dimensions).

We compare these references to four baseline systems. Two of these systems are cascaded, using Qwen3-ASR [28] or Whisper Large V3 to first process the audio into a transcript (which lacks speaker attributions) and then provide that transcript as input to Qwen3-32B-Thinking [33], with the instruction to generate a SOAP note. We compare these to Qwen3-Omni-Instruct and Qwen3-Omni-Thinking [1], 32B parameter models that receive the audio conversation only and are instructed to produce a SOAP note directly. Because Qwen3-Omni is the same size and was constructed by combining Qwen3-ASR and Qwen3

during training, this comparison examines the gap between multi-modal and text-only systems.

First, we analyze the quality of the reference notes. They score highly using the reference-free LLM-as-a-judge and, in particular, are highly faithful to the transcripts (with an average faithfulness of 4.9/5). While coverage, structure, and conciseness are not perfect, they should act as a strong training and evaluation signals for current LALMs. Second, we find a large gap between end-to-end and cascaded Qwen3 systems across all three judged outcomes. End-to-end approaches achieve near-minimal faithfulness and coverage, and exhibit hallucination rates near 0.99–1.00 (vs. 0.21–0.23 for cascaded systems and 0.01 for the reference notes), indicating a widespread inability to process facts from the transcript.

Table 3: Comparison of different note generation methods. Metrics: (Faith)fulness, (Cov)erage, (Struct)ure, and (Conc)iseness

Architecture	Faith.	Cov.	Struct.	Conc.
-ASR+Thinking	3.1 (±0.8)	4.0 (±0.8)	3.9 (±1.0)	3.5 (±0.8)
Qwen3 -Omni-Instruct	1.0 (±0)	1.3 (±1.0)	4.5 (±0.7)	2.4 (±0.8)
-Omni-Thinking	1.0 (±0)	1.5 (±1.3)	4.4 (±0.8)	2.4 (±0.8)
Whisper Large V3+	3.2 (±0.8)	4.2 (±0.8)	4.0 (±1.0)	3.6 (±0.8)
Qwen3-Thinking				
Reference (Oracle + Kimi K2 Thinking)	4.9 (±0.3)	4.0 (±1.0)	4.6 (±0.8)	3.9 (±0.8)

4.4. Reference-Grounded SOAP Note Evaluation

In addition to the reference-free LLM judge, we evaluate generated notes against the reference using standard ROUGE [36] F1 metrics (R-2, R-3, R-L) and an Open Medical Concept metric⁸ inspired by [38], which extracts medical concepts from both reference and hypothesis notes and computes F1 over their overlap (Tab. 4). These metrics also show a large gap between cascaded and end-to-end performance.

Table 4: SOAP note evaluation on dev set (F1 scores in percent).

Model	R-2	R-3	R-L	Open	#Wrd
Whisper+Qwen3	11.8	4.21	22.6	29.0	258
Qwen3-ASR+Qwen3	11.0	3.85	21.8	28.0	257
Qwen3-Omni-Thinking	4.9	1.39	13.0	16.6	336
Qwen3-Omni-Instruct	5.6	1.79	14.3	18.9	354

Several limitations warrant consideration. First, the reference SOAP notes are LLM-generated rather than authored by physicians, which may introduce systematic biases in both the content and the evaluation signal. Second, all dialogues are in English, two-speaker, primary-care first-visit encounters, limiting generalizability to other languages, clinical specialties, or multi-party settings. Third, the low WER achieved on wet audio (<3% for Whisper) suggests that the acoustic simulation may not fully replicate the difficulty of real clinical recordings, and a direct sim-to-real comparison remains for future work.

5. Conclusion

We presented a fully synthetic pipeline—persona-conditioned dialogue generation, multi-speaker audio synthesis, and ref-

⁸Medical concept F1: MeSH keyword matching + NER via `en_core_sci_md` (scispaCy [37]).

erence SOAP note production—built entirely on open-weight models, yielding 8,800 conversations and 1,329 hours of audio. Cascaded systems substantially outperform end-to-end models, which exhibit hallucination rates around 60% versus 20% for cascaded approaches. With Whisper Large V3 achieving 2–3% WER on wet audio, ASR is near ceiling; the primary challenge lies not in transcription but in reasoning over long, noisy conversations to produce faithful clinical documentation. Future work aims to close the gap between synthetic and real clinical audio — through multi-party scenarios (e.g., nurse, caregiver), accent-conditioned voices, and more natural conversational patterns — and to extend the pipeline to other professional domains. All data and code will be publicly released⁹.

6. Acknowledgement

The authors would like to thank Markus Müller (Amazon AGI) for his valuable discussions, leadership, and guidance throughout the duration of the workshop. The contribution by Markus Müller was made in his capacity as a workshop leader and does not necessarily reflect the official position of Amazon.

This work was supported by the 2025 Jelinek Memorial Summer Workshop on Speech and Language Technologies (JSALT 2025), hosted at the Brno University of Technology and organized by the Center for Language and Speech Processing (CLSP) at Johns Hopkins University. This work was supported by the 2025 Jelinek Memorial Summer Workshop on Speech and Language Technologies (JSALT 2025), hosted at the Brno University of Technology and organized by the Center for Language and Speech Processing (CLSP) at Johns Hopkins University. The Idiap employees were funded by European Union Horizon 2020 project ELOQUENCE (101070558).

7. References

- [1] J. Xu, Z. Guo, H. Hu *et al.*, “Qwen3-Omni technical report,” 2025. [Online]. Available: <https://arxiv.org/abs/2509.17765>
- [2] S. Ghosh, A. Goel, J. Kim, S. Kumar, Z. Kong, S. Gil Lee, C.-H. H. Yang, R. Duraiswami, D. Manocha, R. Valle, and B. Catanzaro, “Audio flamingo 3: Advancing audio intelligence with fully open large audio language models,” in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. [Online]. Available: <https://openreview.net/forum?id=FjByDpDVIO>
- [3] Amazon Artificial General Intelligence, “Amazon nova sonic: Technical report and model card,” *Amazon Technical Reports*, 2025. [Online]. Available: <https://www.amazon.science/publications/amazon-nova-sonic-technical-report-and-model-card>
- [4] Y. Li, J. Liu, T. Zhang *et al.*, “Baichuan-Omni-1.5 technical report,” 2025. [Online]. Available: <https://arxiv.org/abs/2501.15368>
- [5] S. Ghosh, Z. Kong, S. Kumar, S. Sakshi, J. Kim, W. Ping, R. Valle, D. Manocha, and B. Catanzaro, “Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities,” in *Forty-second International Conference on Machine Learning*, 2025. [Online]. Available: <https://openreview.net/forum?id=xWu5qpDK6U>
- [6] O. Ahia, M. Bartelds, K. Ahuja *et al.*, “BLAB: Brutally long audio bench,” 2025. [Online]. Available: <https://arxiv.org/abs/2505.03054>
- [7] J. Chen, Z. Guo, J. Chun, P. Wang, A. Perrault, and M. Elsner, “Do audio LLMs really LISTEN, or just transcribe? measuring lexical vs. acoustic emotion cues reliance,” in *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026.
- [8] S. Kumar, Šimon Sedláček, V. Lokegaonkar *et al.*, “MMAU-Pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence,” 2025. [Online]. Available: <https://arxiv.org/abs/2508.13992>
- [9] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau, “How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.
- [10] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, “On faithfulness and factuality in abstractive summarization,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Jul. 2020.
- [11] R. Sharma, A. Gupta, S. Kumar, and F. Metze, “End-to-end speech summarization using restricted self-attention,” in *Proc. ICASSP*. IEEE, 2022, pp. 8072–8076.
- [12] J. Xiang, S. Zhang, W. Zhou, and Y. Liu, “Closing the modality reasoning gap for speech large language models,” in *Proc. IEEE ASRU*, 2025.
- [13] J. Billa, “The cascade equivalence hypothesis: When do speech LLMs behave like ASR→LLM pipelines?” *arXiv preprint arXiv:2602.17598*, 2026.
- [14] I. Shafran, N. Du, L. Tran, A. Perry, L. Keyes, M. Knichel, A. Domin, L. Huang, Y.-h. Chen, G. Li, M. Wang, L. El Shafey, H. Soltau, and J. S. Paul, “The medical scribe: Corpus development and model performance analyses,” in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, May 2020, pp. 2036–2044. [Online]. Available: <https://aclanthology.org/2020.lrec-1.250/>
- [15] H. Soltau, M. Wang, I. Shafran, and L. E. Shafey, “Understanding medical conversations: Rich transcription, confidence scores & information extraction,” in *Interspeech*, 2021.
- [16] S. A. Haider, S. Prabha, C. A. Gomez-Cabello, S. Borna, A. Genovese, M. Trabelsy, B. G. Collaco, N. G. Wood, S. Bagaria, C. Tao, and A. J. Forte, “Synthetic patient–physician conversations simulated by large language models: A multi-dimensional evaluation,” *Sensors*, vol. 25, no. 14, 2025.
- [17] J. Wang, Z. Yao, Z. Yang, H. Zhou, R. Li, X. Wang, Y. Xu, and H. Yu, “NoteChat: A dataset of synthetic patient-physician conversations conditioned on clinical notes,” in *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, 2024.
- [18] A. Papadopoulos Korfiatis, F. Moramarco, R. Sarac, and A. Savkov, “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)*, 2022.
- [19] W.-w. Yim, Y. Fu, A. Ben Abacha, N. Snider, T. Lin, and M. Yetisgen, “ACI-BENCH: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation,” *Scientific Data*, vol. 10, no. 1, p. 586, 2023.
- [20] S. Cornell, J. Darefsky, Z. Duan, and S. Watanabe, “Generating Data with Text-to-Speech and Large-Language Models for Conversational Speech Recognition,” in *Synthetic Data’s Transformative Role in Foundational Speech Models*, 2024.
- [21] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, “Judging llm-as-a-judge with mt-bench and chatbot arena,” in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36.

⁹Train/Dev released before Interspeech 2026; Test data withheld until December 2026

Curran Associates, Inc., 2023, pp. 46595–46623. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and-Benchmarks.pdf

[22] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, “G-eval: NLG evaluation using gpt-4 with better human alignment,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds., 2023.

[23] G. Team, A. Kamath, J. Ferret *et al.*, “Gemma 3 technical report,” 2025. [Online]. Available: <https://arxiv.org/abs/2503.19786>

[24] S. Burdisso, S. Baroudi, Y. Labrak *et al.*, “Sdialog: A python toolkit for end-to-end agent building, user simulation, dialog generation, and evaluation,” 2026. [Online]. Available: <https://arxiv.org/abs/2506.10622>

[25] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, “LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech,” in *Interspeech 2019*, 2019, pp. 1526–1530.

[26] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, “Scaper: A library for soundscape synthesis and augmentation,” in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 344–348.

[27] R. Scheibler, E. Bezzam, and I. Dokmanić, “Pyroomacoustics: A python package for audio room simulation and array processing algorithms,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE Press, 2018, p. 351–355. [Online]. Available: <https://doi.org/10.1109/ICASSP.2018.8461310>

[28] X. Shi, X. Wang, Z. Guo *et al.*, “Qwen3-ASR technical report,” 2026. [Online]. Available: <https://arxiv.org/abs/2601.21337>

[29] J. Shi, H. jin Shim, and S. Watanabe, “Uni-VERSA: Versatile Speech Assessment with a Unified Network,” in *Interspeech 2025*, 2025, pp. 1798–1802.

[30] K. Baba, W. Nakata, Y. Saito, and H. Saruwatari, “The t05 system for the VoiceMOS Challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech,” in *IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 818–824.

[31] D. E. A. Meena, M. Nanivadekar, N. A. V. Azad, A. N. Shenoy, P. R. Chowdhuri, S. Banga, V. Chhabra, C. Bhat, S. babu Kalluri, S. R. Chetupalli, D. Vijayasanen, and S. Ganapathy, “Benchmarking speech systems for frontline health conversations: The displace-m challenge,” 2026. [Online]. Available: <https://arxiv.org/abs/2603.02813>

[32] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 28492–28518. [Online]. Available: <https://proceedings.mlr.press/v202/radford23a.html>

[33] A. Yang, A. Li, B. Yang *et al.*, “Qwen3 technical report,” 2025. [Online]. Available: <https://arxiv.org/abs/2505.09388>

[34] K. Team, Y. Bai, Y. Bao *et al.*, “Kimi K2: Open agentic intelligence,” 2026. [Online]. Available: <https://arxiv.org/abs/2507.20534>

[35] J. Glover, F. Fancellu, V. Jagannathan, M. R. Gormley, and T. Schaaf, “Revisiting text decomposition methods for NLI-based factuality scoring of summaries,” in *Proceedings of the Second Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, A. Bosselut, K. Chandu, K. Dhole, V. Gangal, S. Gehrmann, Y. Jernite, J. Novikova, and L. Perez-Beltrachini, Eds. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 97–105. [Online]. Available: <https://aclanthology.org/2022.gem-1.7/>

[36] Google Research, “rouge-score: A python implementation of rouge,” 2019. [Online]. Available: <https://github.com/google-research/google-research/tree/master/rouge>

[37] M. Neumann, D. King, I. Beltagy, and W. Ammar, “ScispaCy: Fast and robust models for biomedical natural language processing,” in *Proceedings of the 18th BioNLP Workshop and Shared Task*, D. Demner-Fushman, K. B. Cohen, S. Ananiadou, and J. Tsujii, Eds. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 319–327. [Online]. Available: <https://aclanthology.org/W19-5034/>

[38] L. Zhang, R. Negrinho, A. Ghosh, V. Jagannathan, H. R. Hassanzadeh, T. Schaaf, and M. R. Gormley, “Leveraging pretrained models for automatic summarization of doctor-patient conversations,” in *Findings of the ACL: EMNLP 2021*, 2021, pp. 3693–3712. [Online]. Available: <https://aclanthology.org/2021.findings-emnlp.313/>

8. Generative AI Use Disclosure

Generative AI tools were used in two distinct ways in this work.

Manuscript preparation. Large language models were used to assist with proofreading, improving conciseness, and formatting \LaTeX tables. All such use was directed and reviewed by an author; AI tools produced no significant portions of the manuscript without subsequent human revision. AI assistance was also used in the development of experimental code and scripts for running experiments on HPC clusters. Claude Sonnet 4¹⁰ was used for development-only evaluation of the text dialogue generation pipeline and is not a component of the released dataset or final system.

Research methodology. Generative AI models are integral components of the proposed pipeline and are described in full detail in the body of the paper. Specifically: Gemma3-27B-IT¹¹ [23] was used to generate all synthetic dialogues; Qwen3-TTS 1.7B [33] was used for speech synthesis; and Kimi K2 Thinking¹² [34] was used both for reference SOAP note generation and as the LLM-as-a-judge evaluator. Additionally, Qwen3-ASR¹³ [28], Qwen3-32B [33], and Qwen3-Omni¹⁴ [1] were evaluated as baseline systems and are subjects of investigation in this work. These uses constitute the research contribution of the paper and are not uses of AI for manuscript authorship.

All co-authors have reviewed and take responsibility for the full content of this paper.

¹⁰AWS Bedrock model

`global.anthropic.claude-sonnet-4-20250514-v1:0`

¹¹<https://huggingface.co/google/gemma-3-27b-it>

¹²AWS Bedrock model `kimi.moonshot.k2-thinking-v1:0`

¹³<https://huggingface.co/Qwen/Qwen3-ASR-1.7B>

¹⁴<https://huggingface.co/Qwen/Qwen3-Omni-30B-A3B-Thinking> and <https://huggingface.co/Qwen/Qwen3-Omni-30B-A3B-Instruct>