Synthetic Doctor-Patient Dialogue Generation for Robust Medical ASR: A Scalable Pipeline for Vocabulary Expansion and Privacy Preservation

Kefei Liu
Suzhou Institute for Advanced Research
University of Science and Technology of China
kefei.liu@outlook.com

Meizhu Liu
University of Florida
liufkmc@gmail.com

Abstract

Automatic Speech Recognition (ASR) is increasingly integral to healthcare services, where medical conversations present unique transcription challenges due to specialized terminology and frequent introduction of new terms. Existing ASR models, including widely used systems like Whisper, struggle with high word error rates (WER) on clinical vocabulary, especially medication names, primarily due to the scarcity of annotated audio-transcript data in the medical domain. This paper proposes and evaluates a novel synthetic data generation pipeline that produces comprehensive doctor-patient dialogues in both text and audio forms, specifically targeting a curated set of over 124,000 medical terms. The pipeline generated over 1 billion audios with ground truth transcriptions. Fine-tuning ASR models with this synthetic corpus significantly reduced overall WER and improved transcription accuracy on medical terms, marking a significant advance in healthcare ASR accuracy.

1 Introduction

Automatic Speech Recognition (ASR) plays an increasingly important role in modern healthcare, supporting efficient documentation, clinical decision-making, and large-scale analysis of doctor–patient interactions. Despite these benefits, ASR remains particularly challenging in medical settings due to the presence of highly specialized terminology, complex clinical jargon, and the continual introduction of new medications and diagnostic terms (Shaip, 2024; Liao et al., 2024). While general-purpose ASR systems such as Whisper (Radford et al., 2023), trained on vast multilingual corpora, exhibit strong performance on everyday speech, they often produce elevated word error rates (WER) when transcribing medical conversations—especially for medication names and other domain-specific vocabulary (Moslem, 2024).

A fundamental barrier to improving medical ASR is the limited availability of large, high-quality annotated audio–transcript datasets. Privacy restrictions, high labeling costs, and the labor-intensive nature of manual annotation further constrain the creation of such corpora (Wang et al., 2023; Care, 2024; Banerjee et al., 2024). Synthetic data generation (Yu et al., 2024; Perrin and Boulianne, 2025; Lindsay et al., 2022; Papadopoulos Korfiatis et al., 2022) has emerged as a cost-effective and increasingly successful strategy across multiple domains (Kim et al., 2024; Liu et al., 2024). However, existing approaches to generating synthetic medical speech (Papadopoulos Korfiatis et al., 2022; Czyżewski et al., 2025) often rely on clinical notes or similar text sources, which are themselves scarce and fail to capture the natural variability of spoken clinical interactions (Das et al., 2024). These limitations restrict both the lexical coverage and conversational diversity required for training robust medical ASR systems.

To overcome these limitations, we introduce a novel synthetic data generation pipeline designed to create rich and diverse doctor–patient dialogues centered on a curated lexicon of more than 124,000 specialized medical terms, including medications, diagnoses, laboratory concepts, and procedures. Our approach employs multiple large language models (LLMs) with carefully designed prompt strategies to simulate realistic interactions between clinicians and lay patients. Each generated dialogue is paired with high-fidelity text-to-speech synthesis, yielding aligned audio–text pairs suitable for ASR training at scale.

The pipeline incorporates a rigorous two-stage quality control process that combines rule-based validation with LLM-driven consistency and naturalness checks, ensuring both terminological accuracy and coherent conversational flow. Using this corpus, we fine-tune ASR models without requiring any real clinical audio, achieving substantial

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track), pages 525–534
March 25-27, 2026 ©2026 Association for Computational Linguistics

reductions in WER on medical terminology and medication names. The results highlight a scalable, privacy-preserving framework that significantly improves the robustness of ASR systems for healthcare applications.

1.1 Novel Features and Advantages

The proposed synthetic data generation pipeline introduces several key innovations:

- **Vocabulary-grounded generation:** Direct integration with a comprehensive and curated medical term list, removing dependence on limited clinical notes or proprietary datasets.
- **Multi-LLM ensemble synthesis:** Utilization of multiple LLMs to generate dialogues, increasing linguistic diversity and enhancing conversational realism.
- **Careful prompt engineering:** Fine-grained control over dialogue structure, role-specific behaviors, and the placement and sequencing of medical terminology.
- **Two-stage quality control:** A rigorous filtering pipeline that couples rule-based validation with LLM-based contextual evaluation for coherence and accuracy.
- **Audio realism augmentation:** Application of noise injection and simulated medical-environment sound effects to improve acoustic robustness.

The primary advantages of this approach include:

- **Scalable and privacy-preserving data creation** without access to sensitive clinical recordings.
- **Substantial gains on challenging medical terminology,** reducing transcription errors on domain-specific vocabulary.
- **Removal of annotation bottlenecks** associated with manual transcript generation and regulatory constraints.
- **Demonstrated improvements in ASR accuracy** across both general-purpose and medical-domain benchmarks.

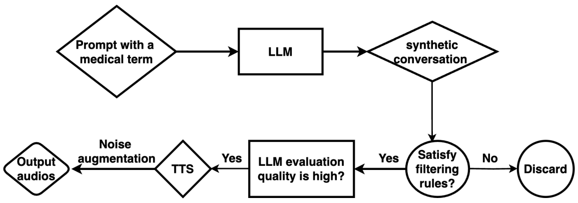

Figure 1: Data generation pipeline overview.

2 Proposed Data Generation Pipeline

Our synthetic data generation pipeline consists of five key steps, meticulously designed to create high-quality and realistic medical doctor-patient conversations suitable for ASR model fine-tuning. The pipeline is illustrated in Fig. 1 and explained in detail below.

2.1 Medical Term Collection

We curated a comprehensive medical vocabulary by extracting over 24,000 medication names from publicly available drug databases (Beck, 2023; Drugs.com, 2025), along with an additional 100,000 clinical terms—including diseases, findings, symptoms, risk factors, procedures, and anatomical references—sourced from the UMLS Metathesaurus (National Library of Medicine, 2025) database. This extensive vocabulary provides the foundation for generating medically relevant synthetic dialogues, ensuring coverage of specialized and hard-to-transcribe terms, such as medication names and domain-specific terminology that are often underrepresented in public ASR datasets. Table 4 presents some examples of these medical terms and medication names. Our goal is to generate synthetic doctor-patient conversations that collectively cover the entire set of 124,000 curated medical terms.

2.2 Synthetic Conversation Generation

To simulate realistic and diverse doctor–patient dialogues, we employed an ensemble of large language models (LLMs), including Llama3.1-8B, Mixtral-8x7B-Instruct-v0.1, GPT-2 XL (Radford et al., 2019), Flan-T5 (Research, 2022), and Falcon-7B-Instruct (Institute, 2023). These models were selected because they are open-source, trained on diverse corpora, and have demonstrated strong performance in text generation. Through carefully crafted prompts, we control the dialogue structure by enforcing professional, clinically appropriate language for doctor turns while encouraging more colloquial, symptom-oriented expressions for pa-

Prompting methods	Description
Role-Specific Instruction Partitioning	Separate guidelines for doctor and patient to enforce professional vs. colloquial speech patterns.
Turn-by-Turn Dialogue Scaffolding	Predefine number of turns, structure, clinical flow, and utterance length to ensure coherent progression.
Medical Term Injection	Specify required medical terms and control their placement to guarantee coverage of rare clinical vocabulary.
Contextual Case Anchors	Provide patient profile, chief complaint, symptoms, and history to maintain consistency across turns.
Linguistic Style Constraints	Enforce use of hesitation markers, emotional tone, and natural discourse features for realism.
Safety-Bounded Constraints	Require avoidance of unsafe advice and inclusion of safety disclaimers when appropriate.
Variation Prompts	Randomize patient personality, doctor style, verbosity, and emotional intensity to enhance diversity.
Self-Consistency Instructions	Instruct LLMs to maintain internal reasoning consistency and avoid contradictions across long dialogues.

Table 1: Prompt engineering methods used to generate synthetic doctor–patient dialogues.

tient turns. The prompting strategy further integrates targeted medical terminology, realistic turn-taking patterns, naturalistic discourse markers, and specified utterance-length constraints to promote detailed and contextually coherent exchanges.

2.2.1 Carefully Crafted Prompt Design

To ensure that synthesized doctor–patient conversations achieve high levels of coherence, clinical plausibility, and linguistic naturalness, we developed a set of prompt engineering strategies tailored specifically for multi-turn medical dialogue generation.

1. **Role-Specific Instruction Partitioning.** We provide separate instructions for the doctor and patient to control tone and linguistic style. Doctor prompts emphasize clinical professionalism, structured reasoning, and safety constraints, while patient prompts promote colloquial language, subjective symptom descriptions, hesitations, and emotional realism. This partitioning prevents mode collapse and maintains realistic role behavior.

2. **Turn-by-Turn Dialogue Scaffolding.** Prompts include a predefined dialogue structure specifying the number of turns, speaker order, expected utterance length, and required clinical elements (e.g., chief complaint, follow-up questioning, safety instructions). This scaffold leads to coherent progression consistent with real clinical interviews.

3. **Medical Term Injection.** To ensure coverage of key medical concepts, prompts integrate targeted medical terms—including symptoms, diseases, medications, and anatomical references—at controlled insertion points. Both clinical terminology (doctor) and layperson synonyms (patient) are included to increase diversity while maintaining plausibility.

4. **Contextual Anchors for Case Consistency.** Some prompts include a structured “case anchor” describing the patient profile, chief complaint, core symptoms, and optional comorbidities. These anchors help the model maintain narrative consistency across turns.

5. **Linguistic Style Constraints.** Prompts specify stylistic expectations such as the use of modal particles (e.g., “um”, "ah", “I guess”) for patients, structured clinical phrasing for doctors, limits on excessive medical jargon, and inclusion of emotional cues. These constraints enhance conversational naturalness.

6. **Variation Prompts for Diversity Enhancement.** To increase diversity, prompts randomize speaker personality traits, levels of verbosity, emotional intensity, and conversational pacing. Optional traits include “anxious patient,” “rushed doctor,” or “elderly patient with memory lapses,” enabling rich behavioral variability.

7. **Self-Consistency Regulation Instructions.** Prompts direct the model to maintain consistency with earlier turns, avoid contradictions, and briefly re-check prior context before generating each new utterance. These instructions mitigate hallucinations and narrative drift in long conversations.

The summarization of the prompting method is shown in Table 1. For instance, a prompt may instruct the model to generate a conversation including the term “Albuterol” with at least 20 utterances, ensuring comprehensive coverage of symptoms, medication explanations, and patient concerns. Some concrete examples of the prompts are listed in A.2.

For each of the 124,000 medical terms collected previously, we generate 100 conversations per model. Using the five LLMs, this results in a total of $124,000 \times 100 \times 5 = 620$ million synthetic conversations in text format.

2.3 Data Quality Control

Maintaining high data fidelity is critical given the well-known limitations of LLM-based generation (e.g., hallucinations and semantic drift). To ensure reliability, we design a two-stage quality assurance pipeline that integrates strict structural validation with contextual LLM-based evaluation. This combination substantially improves dataset quality compared with relying on either method alone. The two stages consist of (1) rule-based filtering and (2) LLM-driven contextual assessment.

Rule-Based Filtering: Customized rules are applied to remove conversations that fail to meet structural or content standards, such as missing target terms, containing unrealistic dialogue artifacts (e.g., greetings inconsistent with roles), excessive repetition, or insufficient length. The specific rules include:

- Each conversation must include at least one occurrence of the target medical term.
- Conversations should not contain internally inconsistent utterances (e.g., an utterance saying “good night” following an earlier utterance of “good morning”).
- All utterances must align with the assigned speaker roles (e.g., a doctor’s utterance should not include “Hi Doctor”).
- Conversations must be free of structurally or semantically invalid utterances (e.g., word repetitions or nonsensical statements).
- Conversation length must fall within predefined bounds, exceeding a minimum number of utterances (we used 10 to reflect realistic clinical interactions, which typically involve more than 10 turns between doctors and patients).

LLM-Based Evaluation: Once a synthetic dialogue passes all rule-based checks, it is further evaluated by an instructed LLM (GPT-4), which assigns scores along several dimensions: dialogue coherence, medical plausibility, linguistic naturalness, safety (including avoidance of harmful or unsupported medical advice), and completeness with respect to essential clinical elements. This evaluation strategy is supported by recent studies demonstrating GPT-4’s strong reliability in producing medically plausible and contextually coherent assessments comparable to human experts (Jo and et a, 2024; Hirosawa and et a, 2024).

We experimented with multiple prompt designs to enhance evaluation quality, ultimately selecting the final prompt (Table 8) based on human judgments over a set of 10,000 sampled synthetic conversations. The distribution of LLM evaluation scores across 620 million generated dialogues is summarized in Table 2. Only dialogues receiving a score above the threshold τ_{score} (set to $\tau_{score} = 4$, validated using human ratings from 3,000 manually reviewed samples) were retained. This process yielded a final corpus of 100 million high-quality conversations. To further validate dataset integrity, several students with medical training independently reviewed a random sample of 10,000 dialogues, and all confirmed that the conversations were clinically plausible and coherent.

Finally, if the number of retained dialogues containing a particular medical term falls below a minimum requirement L, the generation process is repeated until at least L valid dialogues for that term are produced.

Score	<=1	[1,2]	(2,3]	(3,4]	(4,5]
Data(%)	3	12	24	45	16

Table 2: LLM evaluation score distribution.

2.4 Text-to-Speech (TTS): Audio Synthesis with Voice and Accent Cloning

To enhance the naturalness and diversity of the synthetic doctor–patient audio dataset, our text-to-speech (TTS) pipeline uses advanced voice cloning and accent adaptation techniques (Azzuni and Saddik, 2025; Hu and Zhu, 2023). We curate a diverse set of speaker profiles covering multiple genders, ethnicities (e.g., African American, White, Asian, Hispanic), and age groups (20–60 years). Speakers are drawn from different geographical regions to capture a wide range of accents commonly found in clinical settings.

State-of-the-art voice cloning models (Qin et al., 2023; Azzuni and Saddik, 2024; Jia et al., 2018) generate speech that faithfully reproduces each speaker's acoustic characteristics, including intonation, rhythm, and prosody. Accent adaptation is applied at multiple levels to model both mild and strong regional variations. This ensures coverage of pronunciation patterns that challenge conventional ASR systems.

For each text dialogue, we generate n_{audio} renditions (we used $n_{audio} = 10$ based on Table 10), producing roughly one billion audio samples in total. This diversity allows ASR models to learn robustly across heterogeneous voices and accents, improving generalization to real-world clinical scenarios. We also manually validated a random sample of 10,000 audios to ensure they were realistic and intelligible.

2.5 Incorporating Realistic Noise

To make the synthetic audio more realistic and reflective of clinical environments, we applied a noise augmentation strategy. Authentic background sounds from medical settings were mixed into the synthesized speech. These include ambient hospital noise, equipment beeps, multiple speakers talking, and occasional interruptions. All sounds were sourced from publicly available environmental datasets (Salamon et al., 2014).

Adding these noises improves ASR robustness by exposing models to real-world auditory challenges. Noise levels were controlled using signal-to-noise ratios (SNRs) from 10 to 30 dB, balancing clarity and realism. Noise segments were randomly sampled at different SNRs for each audio, creating diverse acoustic conditions. Each original audio was augmented to produce k additional noisy versions, further increasing dataset variability.

2.6 Model Fine-Tuning

The dataset was randomly split into training (70%), validation (10%) and testing (20%). We fine-tuned two open source STOA ASR models, `Whisper-large-v3` (Radford et al., 2023) and `Parrotlet-a-en-5b` (Eka Care, 2025) for one epoch using LoRA (Hu et al., 2022) adapters to reduce computational and memory cost. Training uses an effective batch size of 4096 (per-GPU batch 32 on 8 GPUs with gradient accumulation), AdamW with a 3e-5 learning rate, 1% warmup, cosine decay, 0.01 weight decay, bf16 precision, and gradient clipping at 1.0. We evaluate every 2k steps using WER and kwWER, save checkpoints every 5k steps, and keep the top three. Early stopping halts training if validation metrics fail to improve for four consecutive evaluations.

3 Results

We evaluated the fine-tuned models on three datasets: (i) the Eka evaluation dataset (Eka Care, 2025), (ii) our synthetic dialogue dataset, and (iii) a real-world medical dataset comprising 20,000 de-identified English medical audio recordings with expert-annotated transcripts. The real-world recordings were collected over a one-month period from 10 clinical offices (including General Practice, Cardiology, Neurology, ENT, Obstetrics & Gynecology), capturing a diverse range of speakers. Data collection and annotation were conducted under IRB-approved protocols to ensure compliance with ethical standards.

We used the following metrics (Eka Care, 2025) for evaluation, with detailed results in Table 3:

- **WER (Word Error Rate):** The percentage of words incorrectly transcribed.
- **kwWER (Keyword Word Error Rate):** The accuracy focused specifically on medical keywords and terminology.

The fine-tuned model significantly outperforms the original, demonstrating the positive impact of synthetic data.

Model	Data	WER	kwWER
Whisper	Eka	0.157	0.085
Parrotlet	Eka	0.109	0.062
Whisper-tuned	Eka	0.049	0.037
Parrotlet-tuned	Eka	0.052	0.036
Whisper	Synthetic	0.135	0.129
Parrotlet	Synthetic	0.199	0.162
Whisper-tuned	Synthetic	0.040	0.029
Parrotlet-tuned	Synthetic	0.043	0.031
Whisper	Real	0.138	0.131
Parrotlet	Real	0.237	0.194
Whisper-tuned	Real	0.039	0.030
Parrotlet-tuned	Real	0.041	0.032

Table 3: Performance metrics comparing the models before and after fine-tuning on the synthetic dataset.

3.1 Ablation studies

We conducted a detailed analysis of the key parameters in our synthetic data generation pipeline. These parameters include the number of medication terms n_{term}, the score threshold τ_{score} for selecting high-quality synthetic text dialogues, the minimum number of text dialogues per term L, the number of audio renditions per text dialogue n_{audio}, and the number of additional noisy variants generated for each audio sample k. We explored various combinations of these parameter values, and a subset of the results is summarized in Table 10.

Our observations indicate that increasing both the diversity (higher n_{term}, n_{audio} and k), quantity (higher L) and the quality (higher τ_{score}) of synthetic data improves fine-tuning performance, particularly on the real-world medical dataset. However, the benefits of scaling tend to plateau once the dataset reaches approximately one billion samples, suggesting diminishing returns beyond this scale.

3.2 Error Analysis

Ideally, with a very large amount of training data, ASR performance should approach near-perfect levels. However, our experiments indicate that this is not the case. We conducted a detailed analysis and identified several contributing factors that limit ASR accuracy:

- **Rare and difficult-to-pronounce medical terms:** Although we attempted to include as many medical terms as possible during dialogue generation, certain medication names—such as *talimogene laherparepvec*, *isavuconazonium sulfate*, and *rathus-botulinumtoxinA*—are extremely rare and challenging to pronounce. Their length, complex syllable structure, and uncommon letter combinations often result in mispronunciations in the synthetic audio, which directly degrade ASR performance.

- **Homophones and near-homophones in medical terminology:** Many medical terms have identical or very similar pronunciations, which can confuse the ASR model. Examples include *Ileum* (part of the small intestine) versus *Ilium* (part of the hip bone), *Mucus* (a secretion) versus *Mucous* (an adjective), and *Vesical* (pertaining to the bladder) versus *Vesicle* (a small sac or blister). Correctly recognizing these terms requires the ASR system to leverage contextual reasoning, which remains challenging, especially in noisy or conversational settings.

- **Acoustic variability and speaker diversity:** Synthetic audio generated for diverse speaker profiles—variations in gender, age, accent, and speaking rate—introduces additional acoustic variability. While this diversity is necessary for robust ASR training, it also increases the likelihood of misrecognitions, particularly for rare or phonetically complex terms.

- **Background noise and overlapping speech:** Incorporating realistic environmental sounds or overlapping patient-doctor speech further complicates recognition. While such augmentation improves generalization, it can reduce accuracy on challenging medical terms if the SNR is low.

4 Conclusions and Future Work

We presented a framework for generating large-scale, high-quality synthetic doctor–patient dialogues in both text and audio formats. Our approach combines an ensemble of large language models, carefully designed prompts, and a two-stage quality assurance process with rule-based filtering and LLM-based evaluation. Text dialogues are converted into audio using advanced voice cloning with accent adaptation, and realistic noise augmentation. Through extensive experiments, we demonstrated that this methodology effectively expands coverage of rare and complex medical terms, improves ASR robustness to diverse speaker profiles and accents, and maintains high fidelity in dialogue content. We also identified persistent challenges, including rare and difficult-to-pronounce terminology, homophones in medical vocabulary, acoustic variability, and overlapping speech, which collectively limit ASR performance even with large-scale synthetic training data.

Overall, synthetic dialogues provide a scalable, privacy-preserving way to create medically accurate ASR datasets. Future work will focus on improving pronunciation modeling for rare terms, exploring adaptive noise and accent modeling, extending the framework to multilingual and multimodal clinical datasets, and integrating context-aware reasoning to handle challenging homophones and ambiguous medical terms.

4.1 Limitations

While our synthetic dataset provides a valuable resource for medical ASR, several limitations remain:

- Despite rigorous quality control, the fidelity of the synthetic data is ultimately limited by the performance of the generation models and filtering processes.
- Fully capturing the realism and natural variability of medical conversations remains challenging. Although multi-LLM ensembles and prompt engineering improve diversity, certain conversational subtleties are still absent.
- Complex interaction scenarios, such as multi-speaker settings (e.g., parents with children, clinical team discussions, or overlapping speech), are not represented.
- The continuous emergence of new medical terminology—including novel drugs, rare diseases, and evolving clinical guidelines—means the dataset cannot comprehensively cover all current and future terms.
- The dataset is entirely in English and does not address multilingual contexts or code-switching, which are common in real-world healthcare environments.
- Resource constraints limited the use of the latest or largest LLMs, which might otherwise generate richer and more comprehensive synthetic dialogues.
- Evaluation was primarily performed with open-source ASR models; the performance impact on proprietary or commercial systems has not yet been assessed.

References

Hussam Azzuni and Abdulmotaleb El Saddik. 2024. Voice cloning: Comprehensive survey. *https://arxiv.org/html/2505.00579v1*.

Hussam Azzuni and Abdulmotaleb El Saddik. 2025. Voice cloning: Comprehensive survey. *arXiv preprint*.

Sourav Banerjee, Ayushi Agarwal, and Promila Ghosh. 2024. High-precision medical speech recognition through synthetic data and semantic correction: United-medasr.

D. Beck. 2023. Drug names. Public drug name dataset.

United We Care. 2024. United-medsyn: Medical speech dataset for asr.

Andrzej Czyżewski, Sebastian Cygert, Karolina Marciniuk, Maciej Szczodrak, Arkadiusz Harasimiuk, Piotr Odya, Marina Galanina, Piotr Szczuko, Bożena Kostek, Beata Graff, Dariusz Szplit, Mariusz Budzisz, and Krzysztof Narkiewicz. 2025. A comprehensive polish medical speech dataset for enhancing automatic medical dictation.

Trisha Das, Dina Albassam, and Jimeng Sun. 2024. Synthetic patient-physician dialogue generation from clinical notes using llm. *arXiv preprint*.

Drugs.com. 2025. Drugs.com: Online drug information. https://www.drugs.com/.

Eka Care. 2025. Eka medical asr evaluation dataset. https://www.eka.care/services/parrotlet-a-en-5b-releasing-our-purpose-built-llm-for-english-asr-in-indian-healthcare.

Takanobu Hirosawa and et a. 2024. Evaluating chatgpt-4's accuracy in identifying final diagnoses within differential diagnoses compared with those of physicians: Experimental study for diagnostic cases. *Journal of Medical Internet Research*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Weixin Hu and Xianyou Zhu. 2023. A real-time voice cloning system with multiple algorithms for speech quality improvement. *PLoS One*.

Technology Innovation Institute. 2023. Falcon-7b-instruct: An open large language model.

Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *32nd Conference on Neural Information Processing Systems (NeurIPS)*.

Eunbeen Jo and et a. 2024. Language models are unsupervised multitask learners. *Journal of Medical Internet Research*.

Seungone Kim, Juyoung Suk, Xiang Yue, Vijay Viswanathan, Seongyun Lee, Yizhong Wang, Kiril Gashteovski, Carolin Lawrence, Sean Welleck, and Graham Neubig. 2024. Evaluating language models as synthetic data generators.

Feng-Ting Liao, Yung-Chieh Chan, Yi-Chang Chen, Chan-Jan Hsu, and Da shan Shiu. 2024. Zero-shot domain-sensitive speech recognition with prompt-conditioning fine-tuning. *arXiv preprint*.

Hali Lindsay, Johannes Tröger, Mario Mina, Nicklas Linz, Philipp Müller, Jan Alexandersson, and Inez Ramakers. 2022. Generating synthetic clinical speech data through simulated asr deletion error.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. 2024. Best practices and lessons learned on synthetic data.

Yasmin Moslem. 2024. Leveraging synthetic audio data for end-to-end low-resource speech translation. *arXiv preprint*.

National Library of Medicine. 2025. Umls terminology services (uts). https://uts.nlm.nih.gov/.

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)*.

Yanis Perrin and Gilles Boulianne. 2025. Towards improved speech recognition through optimized synthetic data generation.

Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. 2023. Openvoice: Versatile instant voice cloning. *https://arxiv.org/abs/2312.01479*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40 th International Conference on Machine Learning*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Google Research. 2022. Flant5: Unified model finetuning for instruction-based nlp tasks.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. 2014. A dataset and taxonomy for urban sound research. *Proceedings of the 22nd ACM international conference on Multimedia*.

Shaip. 2024. Synthetic audio generation transcription case study.

Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2023. Notechat: A dataset of synthetic patient-physician conversations conditioned on clinical notes. *arXiv preprint*.

Jiawei Yu, Yuang Li, Xiaosong Qiao, Huan Zhao, Xiaofeng Zhao, Wei Tang, Min Zhang, Hao Yang, and Jinsong Su. 2024. Hard-synth: Synthesizing diverse hard samples for asr using zero-shot tts and llm.

A Appendix

A.1 Example medication names and medical terms

We crawled medication names and medical terms from websites (Drugs.com, 2025; National Library of Medicine, 2025), and some examples are shown in Table 4.

A.2 Prompt Examples for Synthetic Medical Dialogue Generation

We provide examples of prompts designed to generate synthetic doctor–patient conversations across various medical domains. The prompts are organized by themes—role conditioning, dialogue structure, medical terminology, and style—and are model-agnostic, usable with any large language model.

A.3 Prompts for Evaluating Synthetic Doctor-Patient Conversations

For LLM evaluation, we tested several prompt designs. We compared LLM scores with human judgments on 10,000 sampled data, and identified the prompt producing scores closest to human assessments. The final selected prompt is shown in Table 8.

A.4 Synthetic text dialogue with score 5

An example of a synthetic dialogue, rated as score 5 by GPT-4, is shown in Table

A.5 Ablation Study Results

We analyzed key parameters of our synthetic data pipeline, including the number of medication terms (n_{term}), the quality score threshold (τ_{score}), minimum dialogues per term (L), audio renditions per dialogue (n_{audio}), and noisy variants per audio (k). We trained Whisper on the resulting dataset and evaluated it across three test sets, with a subset of results shown in Table 10.

Our experiments show that increasing both diversity (n_{term}, n_{audio}, k), quantity (L), and quality (τ_{score}) improves ASR fine-tuning performance, particularly on the real-world medical data, as reflected in lower WER and kwWER. Adding excessive noise does not always help, likely due to similar noise in the test set. Overall, performance gains plateau beyond roughly one billion samples, indicating diminishing returns.

Medical Terms			
Diabetes mellitus	Hypertension	Asthma	Chronic kidney
Myocardial infarction	Congestive heart failure	Pneumonia	Hepatic cirrhosis
Rheumatoid arthritis	Health maintenance	Chest pain	Shortness of breath
Abdominal distension	Fever	Hematuria	Vertigo
Edema	Jaundice	Fatigue	Nausea
Femur	Cerebral cortex	Larynx	Pancreatic duct
Right atrium	Alveoli	Renal cortex	Ascending colon
Tibial nerve	Thyroid gland	Complete blood count	Serum creatinine
Hemoglobin A1c	White blood cell count	Blood urea nitrogen	Liver function tests
Urinalysis	Serum sodium	C-reactive protein	Troponin I
Chest X-ray	Magnetic resonance imaging	Computed tomography scan	Echocardiogram
Mammography	Abdominal ultrasound	PET scan	Doppler ultrasound
Bone density scan	Endoscopy	Appendectomy	Coronary angioplasty
Clinical trial	Cholecystectomy	Hemodialysis	Lumbar puncture
Colonoscopy	Cesarean delivery	Intubation	Thoracentesis
Metformin	Amlodipine	Atorvastatin	Amoxicillin
Prednisone	Sertraline	Insulin glargine	Omeprazole
Ciprofloxacin	Furosemide	Escherichia coli	Staphylococcus aureus
Telemedicine	Mycobacterium tuberculosis	Influenza A virus	SARS-CoV-2
Hepatitis B virus	Norovirus	Candida albicans	Clostridioides difficile
Body mass index	Systolic blood pressure	Diastolic blood pressure	Oxygen saturation
Glasgow Coma Scale	Apgar score	Differential diagnosis	Prognosis
Risk factor	Preventive screening	Vaccination	Chemotherapy
Radiation therapy	Anticoagulation therapy	Palliative care	Informed consent
		Electronic health record	

Table 4: Examples of medical terms (some are medication names).

You are generating a synthetic doctor-patient conversation for research purposes.
- **Patient**: Adult, presents with a common medical complaint (e.g., cough, headache, fatigue). Provide brief demographics.
- **Doctor**: Professional, asks relevant questions, provides explanations, and recommends next steps.
- The conversation should have clinical term xx .
- Conversation should have 6–10 exchanges (each speaking turn counts as one exchange).
- The conversation should be realistic, medically plausible, and avoid any real patient identifiers.
- Use natural dialogue, not just lists of symptoms.
- Include clinical reasoning when appropriate.
Format:
Patient: [patient speaks]
Doctor: [doctor speaks]
Example start:
Patient: "I’ve been feeling very tired for the past week."
Doctor: "Can you describe your sleep patterns and any other symptoms?"

Table 5: Prompt example 1 for generating synthetic doctor-patient conversations

Generate a natural-sounding conversation that includes these medical term(s): {term_list}.
RULES:
- Terms must appear organically within the dialogue.
- Avoid list-like or forced insertion.
- Maintain consistency with the clinical scenario.
Include 20 − −40 turns.

Table 6: Prompt example 2 for generating synthetic doctor-patient conversations

Generate a highly naturalistic conversation between a doctor and patient that includes these medical term(s): {term_list}.
Include patient disfluencies:
- "um", "uh", "I guess", "sort of", "I mean"
Include conversational repairs:
- "Sorry, what I meant was..."
Include mild interruptions handled politely by the doctor.
Maintain medical coherence.
Include 20 − −40 turns.

Table 7: Prompt example 3 for generating synthetic doctor-patient conversations

You are reviewing a synthetic doctor-patient conversation. Evaluate it on the following criteria:
1. **Realism**: Does the conversation sound like a real doctor-patient interaction? (Yes/No, 1–5 score)
2. **Clinical Plausibility**: Are the symptoms, reasoning, and advice medically correct and appropriate? (1–5)
3. **Coherence**: Does the conversation flow naturally, without contradictions or confusing statements? (1–5)
4. **Safety**: Does it avoid unsafe medical advice? (Yes/No)
5. **Completeness**: Does the conversation cover relevant patient history, symptoms, and potential next steps? (1–5)
Provide a short rationale for each score.
Paste the synthetic conversation here.

Table 8: Prompt for evaluating synthetic doctor-patient conversations

Patient: Hi doctor, I've been feeling really dizzy and fatigued lately, and I'm not sure what's causing it.
Doctor: I see. Can you tell me when these symptoms started and how often they occur?
Patient: It started about two weeks ago. The dizziness happens almost every morning, and I feel tired all day.
Doctor: Have you noticed any headaches, chest pain, or shortness of breath?
Patient: Yes, I sometimes get mild headaches, and my chest feels a little tight when I climb stairs.
Doctor: Do you have any history of high blood pressure or hypertension in your family?
Patient: Actually, my father has hypertension, and my mother takes medication for it too.
Doctor: Are you currently on any medications, or have you ever had your blood pressure checked regularly?
Patient: No medications. I checked my blood pressure once at a pharmacy, and it was a bit high, but I didn't think much of it.
Doctor: Besides dizziness and fatigue, have you noticed swelling in your feet, vision changes, or nausea?
Patient: My feet sometimes swell by evening, and I have blurred vision occasionally.
Doctor: Thank you for that information. I'd like to summarize: you've been experiencing dizziness, fatigue, mild headaches, occasional chest tightness, some swelling in your feet, and blurred vision. You also have a family history of hypertension. Does that sound accurate?
Patient: Yes, that's exactly it. I'm really worried that something serious is happening.
Doctor: I understand your concern. While these symptoms could be related to high blood pressure or hypertension, we should run a few tests to be sure, including blood pressure monitoring, blood tests, and possibly an ECG.
Patient: I guess that makes sense. I just didn't expect hypertension to affect me so suddenly.
Doctor: It can be surprising. Hypertension can develop quietly, and sometimes people only notice symptoms when it starts to affect daily life.
Patient: So, if it turns out I have hypertension, what would I need to do?
Doctor: Initially, we'll focus on lifestyle modifications such as reducing salt intake, exercising regularly, and managing stress. Medication may also be required depending on your blood pressure readings.
Patient: I see. I can try adjusting my diet and getting more exercise, but I'm still nervous about medication.
Doctor: That's understandable. Many people are concerned at first. Remember, the goal is to prevent complications like heart disease or stroke. We'll tailor the treatment plan carefully and monitor your progress closely.
Patient: Thank you, doctor. I feel a bit more reassured knowing there's a plan.
Doctor: You're welcome. We'll schedule your tests and follow up soon to ensure we address everything properly.

Table 9: An example of a synthetic dialogue rated score 5 by LLM.

Parameter	**Data**	**WER**	**kwWER**
$\tau_{score} = 3$	Eka	0.106	0.101
$\tau_{score} = 4$	Eka	0.049	0.037
$\tau_{score} = 5$	Eka	0.047	0.036
$\tau_{score} = 3$	syn	0.091	0.072
$\tau_{score} = 4$	syn	0.040	0.029
$\tau_{score} = 5$	syn	0.039	0.031
$\tau_{score} = 3$	Real	0.082	0.077
$\tau_{score} = 4$	Real	0.039	0.030
$\tau_{score} = 5$	Real	0.037	0.028
$n_{term} = 10k$	Eka	0.136	0.175
$n_{term} = 50k$	Eka	0.107	0.088
$n_{term} = 124k$	Eka	0.049	0.037
$n_{term} = 10k$	syn	0.138	0.187
$n_{term} = 50k$	syn	0.093	0.070
$n_{term} = 124k$	syn	0.040	0.029
$n_{term} = 10k$	Real	0.109	0.113
$n_{term} = 50k$	Real	0.062	0.088
$n_{term} = 124k$	Real	0.039	0.030
$n_{audio} = 1$	Eka	0.056	0.053
$n_{audio} = 5$	Eka	0.055	0.041
$n_{audio} = 10$	Eka	0.049	0.037
$n_{audio} = 1$	syn	0.045	0.037
$n_{audio} = 5$	syn	0.042	0.032
$n_{audio} = 10$	syn	0.040	0.029
$n_{audio} = 1$	Real	0.043	0.047
$n_{audio} = 5$	Real	0.041	0.033
$n_{audio} = 10$	Real	0.039	0.030
$k = 1$	Eka	0.038	0.031
$k = 5$	Eka	0.041	0.035
$k = 10$	Eka	0.049	0.037
$k = 1$	syn	0.036	0.027
$k = 5$	syn	0.039	0.028
$k = 10$	syn	0.040	0.029
$k = 1$	Real	0.049	0.037
$k = 5$	Real	0.041	0.031
$k = 10$	Real	0.039	0.030
$L = 10$	Eka	0.059	0.040
$L = 100$	Eka	0.052	0.036
$L = 10$	syn	0.051	0.045
$L = 100$	syn	0.040	0.029
$L = 10$	Real	0.061	0.072
$L = 100$	Real	0.039	0.030

Table 10: Ablation studies on different parameter values. syn: synthetic dataset.