

Article

A Modular Evaluation of AI-Assisted Clinical Documentation

Julien Delaunay^{1,*} , Maissaa Sarkis¹ , Jordi Solé-Casals^{2,3}  and Jordi Cusido^{1,4,*} ¹ Top Health Tech, 08027 Barcelona, Spain² Data and Signal Processing Group, University of Vic—Central University of Catalonia, 08500 Vic, Spain³ Department of Psychiatry, University of Cambridge, Cambridge CB2 3EB, UK⁴ Departament de Projectes i Construcció, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain

* Correspondence: jdelahunay@tophealthtech.ai (J.D.); jordi.cusido@upc.edu (J.C.)

Abstract

Clinical documentation in Electronic Health Records (EHRs) remains a substantial source of administrative burden for clinicians. In this study, we evaluate a modular AI-assisted clinical documentation pipeline using two complementary approaches: (1) a controlled benchmark based on multilingual synthetic clinical dialogues, and (2) an observational analysis of real-world usage traces from routine deployments. The benchmark enables systematic comparison of ASR–LLM configurations under fully controlled conditions, using metrics for transcription accuracy (Word Error Rate and Medical WER), report-generation quality, and modeled processing cost. Within this benchmark setting, Voxtral showed the strongest ASR performance among the evaluated models, while GPT-4o and Gemini 1.5 Pro showed the strongest report-generation performance under the automated evaluation used in this study. The real-world trace analysis should be interpreted as descriptive evidence of operational use, not as prospective clinical validation or as a direct evaluation of any single benchmarked configuration. Taken together, the results support the use of this pipeline as a human-supervised draft-generation tool that still requires clinician review, local workflow evaluation, and prospective clinical validation before broader deployment.

Keywords: AI scribe; clinical documentation; large language models; automatic speech recognition; human-in-the-loop; synthetic clinical dialogues; digital trace analysis

1. Introduction

Clinical documentation in Electronic Health Records (EHRs) has become a central component of modern healthcare, improving data accessibility, standardization, and continuity of care. However, this shift has also introduced a substantial administrative burden for clinicians, often associated with increased documentation time, reduced patient interaction, and higher rates of professional burnout [1–3]. As a result, there is growing interest in technologies that can assist or partially automate clinical documentation workflows.

Recent advances in Automatic Speech Recognition (ASR) and Large Language Models (LLMs) have enabled a new class of AI-assisted documentation systems, commonly referred to as AI scribes [4–6]. These systems aim to capture clinician–patient conversations, transcribe them, and transform them into structured clinical notes. Early evidence suggests that such systems can support parts of the documentation process, but their effectiveness remains highly dependent on transcription accuracy, robustness to real-world variability, and the quality of generated reports [3,5,7].

Despite rapid progress, important challenges remain. ASR systems must handle specialized medical terminology, diverse accents, and noisy clinical environments, while



Academic Editors: Christophe Cruz,
Hocine Cherifi and Maria Alice
Bertolim Nicoleau

Received: 26 May 2026

Revised: 26 June 2026

Accepted: 30 June 2026

Published: 10 July 2026

Copyright: © 2026 by the authors.
Licensee MDPI, Basel, Switzerland.
This article is an open access article
distributed under the terms and
conditions of the [Creative Commons
Attribution \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

transcription errors may propagate into downstream summaries [8–10]. LLMs, although effective at structuring text, may introduce omissions, distortions, or fabricated content, raising concerns about reliability and safety [4,7,11]. For these reasons, current AI-assisted documentation systems are generally positioned as human-in-the-loop tools, where clinicians retain responsibility for verification and final approval rather than delegating decision-making to autonomous systems [12–14]. This perspective is consistent with emerging human-centered clinical AI literature, which emphasizes the role of AI as a supporting component within expert-driven workflows rather than a replacement for human judgment [11,12].

Recent work has already introduced datasets, benchmarks, and evaluation frameworks for ambient clinical documentation and clinical note generation, including ACI-BENCH, automatic note-generation evaluation work, and ambient digital scribing evaluation frameworks [15–17]. However, the evidence remains heterogeneous, and technical benchmarking, component-level ASR–LLM comparison, multilingual evaluation, and deployment-grounded usage analysis are often reported separately [18–20]. In this context, the present study should be read as an integrated evaluation study rather than as a new general evaluation framework.

The contribution of this paper is to bring together two complementary evidence streams. First, we compare several ASR and LLM configurations under controlled conditions using multilingual synthetic clinical dialogues with preserved ground truth. Second, we analyze finalized-export traces from routine deployments as descriptive evidence of operational use. These two parts answer different questions: the benchmark supports technical comparison under known conditions, whereas the deployment traces show observed use in routine settings. The traces do not measure adoption rates, editing effort, clinical safety, time savings, or patient outcomes. Therefore, the goal is not to claim full clinical validation, but to inform cautious pilot-stage deployment under clinician supervision.

This framing leads to the following research questions:

1. **RQ1.** How do contemporary ASR–LLM configurations differ in transcription accuracy, medical terminology preservation, and downstream report-generation performance in a controlled multilingual benchmark?
2. **RQ2.** What patterns of operational use are observed in finalized-export traces from routine, clinician-supervised deployments?
3. **RQ3.** What practical evaluation considerations emerge when controlled benchmarking and deployment-grounded usage traces are interpreted as complementary, but distinct, forms of evidence?

To answer these questions, we pursue three linked objectives: (1) compare ASR–LLM configurations in a controlled multilingual synthetic benchmark with preserved ground truth; (2) describe operational-use patterns from finalized-export traces in clinician-supervised deployments; and (3) identify practical interpretation limits when benchmark and deployment evidence are considered together. These objectives are intended to support cautious pilot-stage assessment, not to establish adoption rates, clinical effectiveness, workflow impact, or autonomous deployment readiness. We also frame AI-assisted documentation as a data-generation process, where reliable downstream information depends on the quality, structure, and governance of the clinical data produced at the point of care.

2. System Architecture

The AI scribe is a modular web application designed to automate the creation of structured clinical reports from doctor–patient conversations. Its architecture emphasizes flexibility, allowing for the benchmarking of different ASR and LLM models to find optimal configurations for various clinical use cases, languages, and specialties. The system

workflow is coordinated by a backend orchestration layer that connects its core functional modules, as illustrated in Figure 1.

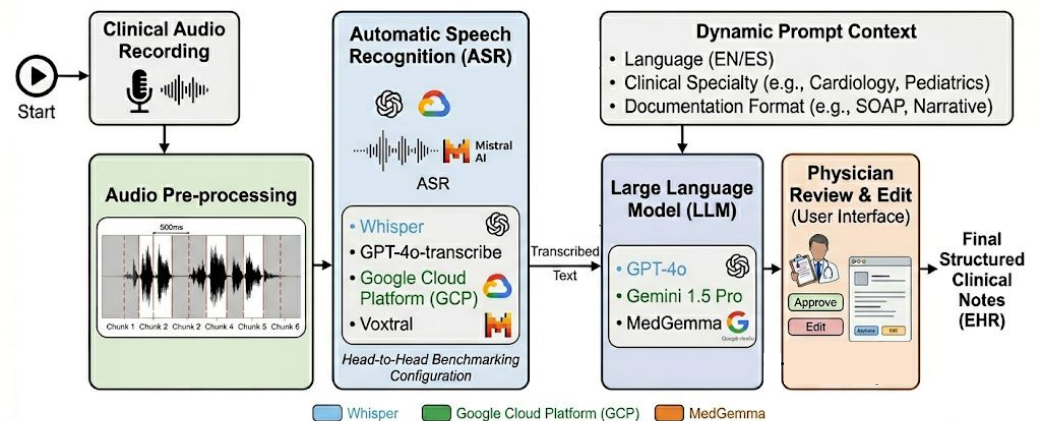


Figure 1. Overview of the specific clinical audio documentation workflow. The modular design allows for head-to-head benchmarking of different ASR (e.g., Voxtral, Whisper) and LLM (e.g., GPT-4o, Med-Gemma) configurations while enforcing dynamic, specialty-specific prompting, and mandatory human-in-the-loop review.

2.1. Conversion of Audio to Text

The AI scribe's workflow begins when a healthcare professional interacts with the web-based user interface, which allows for both direct audio recording of doctor–patient conversations and the upload of pre-existing audio files in various formats. Upon submission, the audio is securely transmitted to the backend server, where it undergoes a comprehensive pre-processing pipeline. This pipeline includes validation of audio format and quality, normalization of volume levels, and segmentation into shorter, more manageable chunks using advanced silence detection algorithms. This segmentation is not only necessary to adhere to the time limits of transcription modules, but also, as highlighted by several studies, to improve processing efficiency and the accuracy of downstream speech recognition [21,22].

Once pre-processed, each audio chunk is routed to the ASR module. The AI scribe supports a modular ASR backend, enabling the selection between top-performing models such as OpenAI's Whisper and GPT-4o-transcribe. The choice of ASR model can be configured per experiment or user preference, facilitating comparative analysis and optimization for different clinical scenarios, languages (including English, Spanish, French, and others), and audio qualities. The ASR module transcribes the spoken dialogue into raw text, which is then subjected to a series of post-processing steps. These include correction of common transcription errors, removal of non-speech artifacts (such as URLs or background noise), and optional speaker diarization if supported by the ASR model. These steps are important because they improve the readability and overall quality of the transcript before the subsequent LLM processing phase.

2.2. Generation of the Clinical Report

The cleaned transcript is then forwarded to the LLM module. This module is also highly configurable, supporting integration with leading LLMs such as Google's Gemini and OpenAI's GPT. The LLM receives a dynamically constructed prompt that includes the full transcript, the target language, the relevant medical specialty, and explicit instructions to structure the output as a formal clinical report. The prompt can be tailored to enforce specific report formats (e.g., SOAP: Subjective, Objective, Assessment, and Plan), a capability increasingly explored and validated for structured clinical note generation by LLMs [23,24]. It can also include or exclude certain sections and adapt to local clinical documentation

standards. The LLM processes the input and generates a comprehensive, well-structured clinical report, ensuring the use of precise medical terminology and adherence to the requested format.

After generation, the report is presented to the user in the web interface, where the healthcare professional can review, edit, and approve the content. This human oversight process is important in medical applications to support accuracy and safety [12]. The system provides tools for highlighting potential uncertainties or inconsistencies (e.g., flagged medication dosages or ambiguous terms), and supports exporting the final report in markdown, plain text, or PDF formats. Throughout the workflow, all core ASR and LLM processing is performed via secure cloud APIs, supporting scalability and access to current model versions. Importantly, the AI scribe is designed with privacy by default: no patient-identifiable information is required or stored, and all data processing is intended to comply with relevant data protection regulations, an important consideration in healthcare [25]. The system also supports multilingual and specialty-specific workflows, including evaluation with cross-specialty mismatch controls, and can be extended to integrate with EHR systems or other clinical IT infrastructure.

The AI scribe is distinguished by several design principles aimed at ensuring robust evaluation and adaptability. A core feature is its multi-model flexibility, which supports various ASR and LLM backends, thereby enabling rigorous benchmarking to identify the most accurate and cost-effective model configurations. This adaptability extends to different clinical contexts through dynamic prompts tailored for various medical specialties and languages, including English and Spanish.

3. Related Work

The automation of clinical documentation represents a significant area of research and development within health informatics, driven by the need to reduce clinician burden and improve efficiency [2,26,27]. This section reviews the evolution of documentation practices, the core technologies enabling automation, the current landscape of tools, evaluation methodologies, and the associated challenges.

3.1. Evolution of Clinical Documentation and AI Scribes

The evolution of clinical documentation reflects a continuous effort to improve record-keeping accuracy, accessibility, and efficiency in healthcare. Initially, clinical notes were primarily handwritten, a practice that, while personal and immediate, was prone to issues of legibility, completeness, and standardization [1]. The introduction of structured formats represented an early attempt to bring order and consistency to medical records, often utilizing pre-defined templates and checklists to ensure key information was captured. However, these structured approaches could sometimes feel rigid and fail to fully capture the nuances of complex patient encounters.

The digital revolution brought about the era of Electronic Health Records (EHRs), with widespread adoption being significantly accelerated by governmental initiatives such as the HITECH Act in the United States (<https://blogs.opentext.com/history-health-information-management-now/> (accessed on 26 June 2026)). EHRs promised numerous advantages, including enhanced data accessibility across different healthcare providers, improved data management, reduced errors associated with paper-based records, and the potential for data-driven insights. Despite these benefits, the implementation of EHRs often introduced unforeseen challenges. Clinicians frequently reported a substantial increase in the time and complexity associated with documentation, due to requirements for detailed data entry, navigation through complex interfaces, and adherence to standardized coding and billing practices [2,3]. This increased administrative burden has been widely recognized

as a significant contributor to clinician burnout and reduced job satisfaction, potentially impacting the quality of patient care.

In response to these challenges, interest has grown in AI-powered medical scribes, which use ASR and LLMs to convert clinician–patient conversations into draft structured clinical notes [28]. These systems are intended to reduce documentation workload and return clinician attention to direct patient care, although their practical value depends on transcription accuracy, report quality, and integration into clinical workflows.

3.2. Core Technologies: ASR and LLMs

Automated clinical documentation tools primarily rely on two core technologies: Automatic Speech Recognition to convert speech to text, and Large Language Models to structure that text into clinically relevant notes [4,6].

ASR models, such as OpenAI’s Whisper [9], Google Cloud Speech-to-Text [29], Amazon Transcribe, and Microsoft Azure Speech, extract audio from the doctor–patient conversation and convert it into a sequence of words. The performance of ASR systems is often quantified using the Word Error Rate (WER), defined as:

$$WER = \frac{S + D + I}{N}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of words in the reference transcription. While general ASR accuracy can be high, achieving low WER in the medical field presents unique challenges. These include specialized terminology, variations in accents and dialects, and noisy environments with overlapping speech [8,9,30]. Errors in transcribing medical terms are a critical concern, and specialized models are often employed to improve accuracy in these contexts [10,31,32].

The challenge of robust signal processing under variable acquisition conditions is not unique to clinical speech recognition. In other engineering domains, including structural health monitoring and non-destructive testing, recent studies have investigated methods for improving signal robustness under environmental variation through approaches such as long-term temperature compensation in ultrasonic guided waves [33] and machine learning-based multi-feature integration [34]. Although these methods are not directly transferable to clinical ASR, they reinforce a broader methodological principle: robust feature extraction, calibration, and multimodal or multi-feature analysis are often necessary when signal quality is affected by changing operational conditions. Similar considerations are relevant in healthcare environments, where accents, background noise, recording devices, and workflow variability can influence transcription performance.

The extracted text is then processed by LLMs such as GPT-4 [35], Gemini 1.5 Pro [36], and specialized medical models such as Med-Gemini [37] and ClinicalBERT [38]. These models transform transcripts into structured clinical notes, often in formats such as SOAP (Subjective, Objective, Assessment, and Plan) [6]. The quality of the generated notes can be evaluated using metrics like ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores, which measure the overlap of n -grams between the generated text and a reference summary. For example, ROUGE-1 focuses on unigram overlap:

$$ROUGE - 1 = \frac{\text{Number of overlapping unigrams}}{\text{Total number of unigrams in the reference summary}}$$

LLMs excel at understanding context and summarization but are prone to “hallucinations,” generating false or inaccurate information, necessitating validation and often clinician review [4,5]. Techniques like prompt engineering and task-specific fine-tuning are often used to improve reliability in medical documentation settings [6]. Research also

explores using LLMs to correct ASR errors [10,39], potentially by leveraging the LLM's contextual understanding to refine the ASR output.

3.3. Existing Systems and Landscape

Numerous commercial AI scribe solutions have emerged, each employing distinct techniques. Nuance DAX from Microsoft leverages ambient clinical intelligence to analyze the natural flow of conversations for detailed note production. Abridge (<https://www.abridge.com/> accessed on 26 June 2026) focuses on efficiently summarizing the most critical elements of these interactions, while DeepScribe uses proprietary AI models trained on medical dialogues to enhance accuracy. Other systems like Suki AI offer voice-enabled, hands-free interaction [40], and Augmedix often integrates a human-in-the-loop approach for quality assurance [40]. Nabla emphasizes real-time assistance and seamless integration with existing EHR systems for immediate documentation support [41].

Despite these diverse approaches, several key features are common. EHR integration, particularly with widely adopted systems like Epic and Cerner, is crucial for embedding notes directly into patient records [3,40]. Real-time processing and support for various clinical note formats, such as SOAP notes, ensure compatibility with existing workflows. Recognizing the global nature of healthcare, some systems are also expanding their multilingual capabilities to cater to diverse patient populations [42]. The underlying techniques employed by these AI scribe solutions exhibit notable variations. The choice of Automatic Speech Recognition (ASR) models can differ, with each model possessing unique strengths and weaknesses in handling the complexities of medical language, diverse accents, and noisy clinical environments. Similarly, the Large Language Models (LLMs) utilized for structuring the transcribed text into clinical notes can range from general-purpose models to those specifically fine-tuned for medical text processing. Finally, the degree of human oversight varies significantly, with some systems relying solely on AI processing while others incorporate human review to enhance accuracy and reliability [40]. These differences in techniques and features highlight the ongoing innovation and diverse strategies within the rapidly evolving field of AI-powered clinical documentation.

3.4. Challenges and Research Gaps

The landscape of automated clinical documentation, while promising, faces substantial technical and ethical hurdles. From a technical perspective, the main unresolved issues for evaluation are ASR robustness to language, audio-quality, and specialty-specific terminology variation, together with the risk that LLM-generated notes may omit, distort, or hallucinate clinically relevant information [4,5,42–44].

Beyond these technical considerations, ethical challenges demand careful attention. These encompass ensuring stringent data privacy and compliance with regulations like HIPAA, addressing potential biases within AI models stemming from training data that may not adequately represent diverse linguistic or patient populations, defining accountability for system errors, and obtaining informed patient consent [11,45].

Recent literature has introduced important benchmarking frameworks for AI scribes, such as ACI-BENCH [15] and the SCRIBE evaluation framework [17], which combine automated metrics, human evaluation, and simulation testing. Similarly, datasets such as MedSynth [46] and MMedBench [47] have significantly advanced the use of synthetic and multilingual data for medical LLM evaluation. However, many existing studies still evaluate either isolated components, rely on monolingual datasets, or provide limited evidence linking controlled benchmark behavior with routine deployment. Building upon these foundations, our study complements existing frameworks by combining a highly modular ASR–LLM benchmark across English and Spanish with real-world deployment

telemetry. A recent systematic review of AI scribes also reinforces the need for such complementary evidence: only eight heterogeneous studies were identified, most from recent and largely U.S.-centric settings, and outcome heterogeneity prevented stronger synthesis [18].

Table 1 summarizes this positioning relative to prior evaluation studies and clarifies how the present study complements, rather than replaces, existing benchmarks and evaluation frameworks.

Table 1. Positioning of the Present Study Relative to Prior Evaluation Work.

Prior Work	What It Contributes	How Our Study Is Positioned
ACI-BENCH [15]	Public benchmark corpus for automatic visit-note generation from clinical dialogue.	We are not the first benchmark. We add modular ASR–LLM comparison, English–Spanish synthetic audio, and deployment-grounded usage traces.
Ben Abacha et al. [16]	Analysis of evaluation methods and metrics for automatic medical note generation.	We do not claim that automated metrics solve evaluation. We use them as technical indicators within one applied benchmark.
SCRIBE framework [17]	Broad ambient digital scribing evaluation framework combining simulation, computational metrics, human review, and LLM-based evaluation.	We do not present a new general framework. We provide an applied integrated evaluation of one modular documentation pipeline.
Digital-scribe and ambient-AI studies [3,7,48]	Measure documentation time, note quality, editing, or clinician-reported outcomes.	Our telemetry does not measure these outcomes. It only describes finalized-export usage patterns.
Systematic reviews [18–20]	Show that the field is promising but heterogeneous, with variable methods and outcomes.	Our study contributes one structured evaluation, but it cannot close the evidence gap alone.

The present study should therefore be interpreted as an applied, early-stage evaluation that combines controlled benchmarking and deployment-grounded operational observations. It complements existing benchmarks and evaluation frameworks rather than replacing them.

4. Experiments

This section outlines the experimental methodology used to evaluate the performance of Automatic Speech Recognition (ASR) and Large Language Models (LLMs) in generating structured clinical reports from medical dialogues. The experimental process involved generating synthetic medical dialogues, converting them to audio, transcribing them using different ASR models, and generating clinical reports using different LLMs. The results were evaluated using predefined metrics, and performance and cost data were recorded across multiple iterations. This benchmark setup allowed us to compare technical performance and modeled processing costs under controlled conditions, while leaving real-world workflow integration and clinical effectiveness for future prospective evaluation.

4.1. Dataset Generation

To rigorously evaluate the performance of our automated clinical documentation system, we generated a set of synthetic medical dialogues across a diverse range of specialties, including dentistry, ophthalmology, general practice, cardiology, and dermatology.

The use of synthetic data is an emerging approach to address data scarcity in specialized domains [49,50]. In the context of this hybrid study, the controlled benchmark is deliberately synthetic for early-stage assessment, not merely a fallback for unavailable real data. This enables controlled comparison across ASR–LLM configurations, consistent scenario coverage across specialties and languages, and the use of shareable benchmark materials without exposing patient-identifiable information. The complete list of medical specialties, their associated conditions, and clinical contexts is detailed in Appendix A.1 (Table A1). An illustrative example of a generated doctor–patient discussion is also provided in Appendix A.2. As illustrated in Figure 2, these dialogues were created using Google’s Gemini-2.0-flash model and were designed to mirror realistic doctor–patient interactions. Each simulated encounter followed a structured format, encompassing essential elements such as initial greetings, patient identification, detailed symptom description, the process of clinical examination, a thorough discussion of the diagnosis, proposed treatment recommendations, and follow-up instructions.

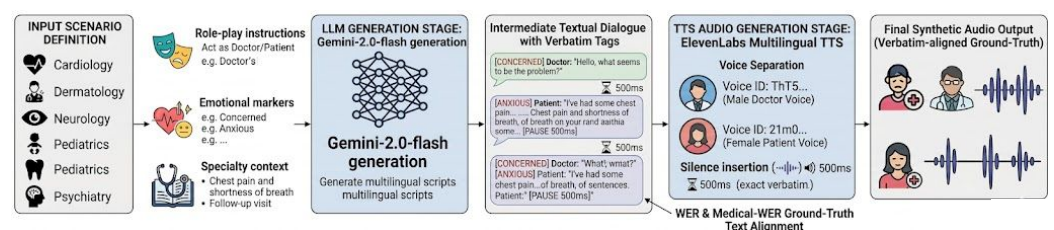


Figure 2. The process of generating synthetic medical dialogues. Gemini-2.0-flash is used to create textual discussions between doctors and patients across various medical specialties. These textual dialogues are then converted to audio using ElevenLabs’ text-to-speech service, resulting in a dataset of simulated doctor–patient conversations. This synthetic pipeline is crucial for the controlled benchmark, as it provides an exact, verbatim text ground-truth to accurately calculate Word Error Rate (WER) and Medical-WER, which is rarely possible with noisy real-world clinical recordings.

A key advantage of generating these synthetic dialogues is that we retain the original, verbatim text of the entire simulated discussion. This text serves as the ground truth against which the accuracy of the ASR component can be measured precisely. It also allows controlled evaluation of whether clinically salient entities and decisions are preserved through the full documentation pipeline. By comparing the original dialogue, the ASR-transcribed text, and the final LLM-generated report, we can quantitatively assess information flow and potential loss or distortion at each stage of the automated documentation pipeline. For these reasons, a synthetic evaluation component is well suited to controlled benchmarking and component-level comparison, even though it does not by itself establish prospective clinical validity.

This synthetic benchmark serves as the controlled benchmark within the study’s bench-to-field evidence framework. The controlled benchmark uses synthetic dialogues because private real-world consultations rarely provide shareable, perfectly aligned ground-truth transcripts and reference notes needed to calculate exact ASR error rates and downstream LLM-report discrepancies at the component level. In contrast, the deployment-grounded evidence stream uses routine deployment traces to test whether the same pipeline is operationally feasible under heterogeneous, messy environments in which microphones, room acoustics, conversational structure, specialty mix, and user behavior are not experimentally controlled. The two evidence streams therefore answer different but complementary questions: technical accuracy under known ground truth, and observed feasibility under real-world workflow conditions.

4.1.1. Dialogue Generation

Our synthetic dialogue generation process leveraged Google's Gemini-2.0-flash model, with parameters carefully tuned to achieve a balance between creative natural language generation and consistent adherence to the simulated clinical scenarios (temperature = 0.7, top_p = 0.95, top_k = 40). To thoroughly evaluate the multilingual capabilities of our system and its effectiveness in processing local medical terminology, each dialogue was generated in English and Spanish. The generation of these dialogues adhered to a structured JSON schema designed to enforce several key characteristics: clear delineation of speaker roles (doctor and patient), the inclusion of emotional context markers (neutral, concerned, reassuring, professional, anxious, relieved) to add realism to the interactions, the accurate incorporation of language-specific medical terminology, and a natural flow of conversation that mimics real-world clinical discussions. This structured approach ensured a degree of consistency and allowed for targeted analysis of the system's performance across different linguistic and emotional contexts.

4.1.2. Audio Generation

The textual dialogues generated in the previous phase were then converted into audio using ElevenLabs' sophisticated multilingual text-to-speech service. We implemented a detailed voice configuration system to further enhance the realism of the simulated interactions. For English dialogues, a professional male voice (identified by the ElevenLabs voice ID *ThT5KcBeYPX3keUQqHPH*) was consistently used for the doctor, while a neutral female voice (*21m00Tcm4TlvDq8ikWAM*) was used for the patient. Similarly, for Spanish dialogues, we selected professional male (*LIZr3QuzbW4WrPjgATHG*) and neutral female (*gD1lexrzCvsXPHUuT0s3*) voices, specifically chosen for their authentic accents, to ensure linguistic fidelity.

To further emulate natural conversation dynamics in the generated audio, we incorporated several key elements: consistent 500-millisecond silence intervals between speaker turns to provide realistic pauses, emotion-appropriate voice modulation based on the emotional context markers embedded in the dialogues, and language-specific prosody and intonation to ensure the audio sounded natural.

4.1.3. Dataset Characteristics

The descriptive statistics in Table 2 provide an overview of the characteristics of the synthetic medical dialogues and their corresponding audio files used in our experiments. The average audio duration is 226.28 s (approximately 3.8 min), with a standard deviation of 38.13 s, indicating moderate variability in dialogue length. These files contain approximately four minutes of near-continuous synthesized speech rather than natural consultations with long silences, pauses, interruptions, or background activity. Although this creates a dense benchmark signal, it is shorter than many real-world consultations, particularly in specialties such as psychiatry or general practice, where encounters may exceed 15–30 min. Longer consultations, silence patterns, background noise, and overlapping speech can pose additional challenges for ASR systems and downstream summarization, and noise has been shown to affect the quality of AI-generated clinical summaries [44]. These conditions were not directly studied in the present benchmark. Each discussion contains, on average, 573.88 words and 22.62 conversational turns, reflecting the controlled complexity and depth of the simulated clinical encounters.

A notable observation is the distribution of dialogue between the doctor and patient roles. Doctors contribute the majority of the content, accounting for 73.46% of the total words and 63.54% of the turns, while patients contribute 26.54% of the words and 36.46% of the turns. This asymmetry is consistent with real-world clinical consultations, where the

physician typically leads the conversation, provides explanations, and documents findings, while the patient responds to questions and describes symptoms.

Table 2. Descriptive statistics of the synthetic medical dialogues and audio files.

Metric	Mean	Std. Dev.
Audio duration (s)	226.28	38.13
Words per discussion	573.88	99.76
Turns per discussion	22.62	3.50
Doctor words	421.56	69.72
Patient words	152.31	38.87
Doctor turns	14.38	2.32
Patient turns	8.25	1.60
Proportion of doctor words (%)	73.46	—
Proportion of patient words (%)	26.54	—
Proportion of doctor turns (%)	63.54	—
Proportion of patient turns (%)	36.46	—

The standard deviations for both word counts and turn counts are relatively low compared to their means, suggesting that the dataset is reasonably balanced and does not contain extreme outliers. Overall, these statistics indicate that the generated dataset captures a diverse range of simulated medical dialogues and is suitable for evaluating the performance of ASR and LLM models in this benchmark.

4.2. Experimental Setup

The experimental setup was designed to compare a selection of ASR and LLM models that represent different paradigms: top-performing general-purpose models, a commercial service, and a specialized domain-specific model. A key aspect of this setup was to establish a lower bound on performance by evaluating the models' behavior when presented with semantically incongruent input, which serves as a benchmark for stress-testing robustness.

4.2.1. ASR Model Comparison

We conducted a comparative evaluation of four prominent ASR models: OpenAI's Whisper [9] and a specialized version of GPT-4 optimized for transcription tasks (gpt-4o-transcribe) were chosen as representatives of top-performing general-purpose ASR technology. Google Cloud Speech-to-Text [29] was included as a benchmark for a high-quality, widely used commercial service. Finally, we evaluated Voxtral [51], an open-source speech understanding model developed by Mistral AI. Released in July 2025 under the Apache 2.0 license, Voxtral is built upon the Mistral Small 3.1 backbone and features a 32k token context window. We specifically utilized the Voxtral Mini (3B parameters) variant, which is optimized for low-latency transcription and cost-efficiency, to assess the performance of a resource-efficient model designed for production-scale applications. To establish a lower-bound reference condition, we implemented a cross-specialty mismatch control. In what follows, this term refers to the concrete experimental setup, whereas "degraded-input negative control" refers to its analytical role. These labels are used here for clarity and are not intended as universal terminology. In this condition, Whisper transcribed each audio file under a deliberately mismatched specialty context, producing degraded transcripts that simulate strong domain mismatch. This control was intended as a stress test of the pipeline under semantically incongruent input rather than as a clinically realistic operating mode.

4.2.2. LLM Post-Processing

For the task of clinical report generation, we compared the performance of three advanced LLMs. GPT-4o [52] and Gemini 1.5 Pro [36] were selected as leading, general-purpose foundation models to set a high benchmark for performance. In contrast, MedGemma [53] was included as a specialized, smaller-scale (4B parameters) model designed specifically for medical applications. These models were evaluated both on standard ASR transcripts and on the degraded transcripts produced by the cross-specialty mismatch control. When the results tables refer to the cross-specialty mismatch control, this denotes the lower-bound condition built from cross-specialty Whisper transcripts, used as a reference condition for downstream report generation rather than as an additional stand-alone foundation model. This allowed us to assess how the pipeline behaves when the input transcript lacks semantic coherence with the target specialty.

4.2.3. Prompting and Structured Evaluation

The report-generation prompts were designed to keep the task clinically realistic while minimizing opportunities for hallucination. Across English and Spanish, the core instructions were the same: generate a concise physician-style report in the source language, preserve the supplied section headings, follow SOAP-like organization implicitly, include medications and diagnoses only when explicitly mentioned, and avoid adding information not present in the transcript. The Spanish variant additionally required all requested sections to be present and instructed the model to mark missing content explicitly when the dialogue did not support a section.

The automated evaluation prompt likewise used a fixed structured format. Gemini-2.0-flash returned JSON scores for factual accuracy, information coverage, clinical relevance, hallucination score, and terminology accuracy. This design follows the broader LLM-as-a-judge paradigm, in which strong language models can provide scalable approximate judgments for open-ended outputs, while also inheriting known evaluator biases such as position, verbosity, self-enhancement, and self-preference effects [54–56]. This allowed the benchmark to compare text quality, semantic preservation, and safety-oriented behavior within the same reproducible evaluation loop while keeping the scoring interface identical across ASR–LLM combinations.

A methodological limitation of this design is that the Gemini family was used in multiple roles across the study: synthetic dialogue generation, one of the evaluated LLM families, and the automated clinical evaluation framework described in Appendix A.4. Recent research suggests that this type of intra-family overlap may influence LLM-as-a-judge evaluations, because evaluators can favor generations that share similar styles, alignments, or post-training distributions [55,56]. We therefore interpret the resulting scores as an applied benchmark within a specific evaluation setup, not as a model-agnostic final ranking. In particular, the absolute Gemini 1.5 Pro scores should be read with caution, whereas cross-model trends, safety-oriented failure patterns, and the comparison between standard and degraded-input conditions remain useful for identifying pipeline behavior under a fixed scoring protocol.

4.2.4. Multilingual and Length Variations

To further assess the robustness and generalizability of the models, we extended our evaluation to include medical dialogues in English and Spanish reflecting the multilingual context of our research. Critically, the cross-specialty mismatch control ensures that we evaluated the lower bound of performance for both languages when the ASR models were presented with semantically mismatched audio. This comprehensive approach allowed us

to understand the fundamental limitations of the ASR and LLM pipelines when the initial audio transcription is significantly compromised due to domain mismatch.

5. Results

We conducted a systematic evaluation of Automatic Speech Recognition (ASR) and Large Language Model (LLM) combinations for medical report generation across 230 experiments spanning 10 medical specialties in English and Spanish. Table 3 presents the experimental design matrix, showing balanced coverage across model combinations with particular emphasis on contemporary ASR–LLM pairings used in current clinical documentation research and practice.

This section presents the results of the controlled benchmark conducted under standardized experimental conditions. These analyses evaluate the relative performance of the investigated ASR–LLM configurations using synthetic multilingual clinical dialogues and should be interpreted as benchmark evidence rather than direct evidence of real-world clinical deployment.

Table 3. Number of reports generated by each ASR–LLM combination.

		LLM Model			
		Gemini 1.5 Pro	GPT-4o	MedGemma	Cross-Specialty Mismatch Control
ASR Model	GPT-4o Transcribe	20	20	20	0
	Voxtral	20	20	20	0
	Whisper	20	20	20	0
	GCP	10	10	10	0
	Cross-specialty mismatch control	0	0	0	20

The experimental design reflects the practical constraints and opportunities in clinical settings, where advanced ASR models are typically paired with sophisticated LLM systems, while the cross-specialty mismatch control functions as a degraded-input negative control. The reduced sample size for GCP combinations reflects the restriction to English language evaluation for this commercial solution. Here, the control-only pairing denotes the lower-bound condition derived from cross-specialty Whisper transcripts and is included as a reference condition rather than as an additional standalone foundation model.

5.1. Speech Recognition Results

Our analysis of ASR model performance centers on two critical metrics that capture different aspects of transcription quality in medical contexts. The Word Error Rate (WER) provides the standard measure of transcription accuracy, calculated as $WER = \frac{S+D+I}{N}$ where S , D , and I represent substitutions, deletions, and insertions, respectively, normalized by the total number of words N in the reference transcription. Additionally, we employed Medical WER (M-WER), a domain-specific variant designed to penalize errors in clinically relevant terminology. M-WER is calculated using a weighted Levenshtein distance alignment, where medical entities identified via the SNOMED CT ontology are assigned a penalty weight of $\lambda = 3.0$ for substitutions, deletions, or insertions, compared to $\lambda = 1.0$ for general vocabulary. This weighting scheme reflects the disproportionate patient safety risk associated with inaccuracies in drug names, anatomical terms, and clinical procedures.

Figure 3 presents the comparative performance evaluation across all ASR models and shows material differences within this benchmark.

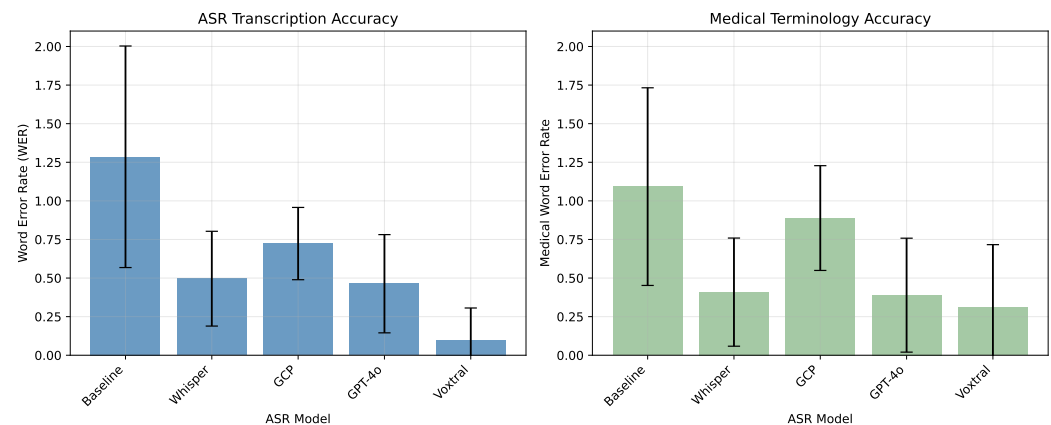


Figure 3. ASR model performance comparison. Lower values indicate better performance. **(Left):** Word Error Rate (WER). **(Right):** Medical Word Error Rate (M-WER).

Voxtral [51] achieved the lowest mean WER in this benchmark (0.099 ± 0.207), corresponding to an approximately 80% relative reduction compared with Whisper (0.496 ± 0.307). A similar pattern was observed relative to GPT-4o audio transcription (0.463 ± 0.318 WER), for which Voxtral showed an approximately 78% relative reduction. In this controlled benchmark setting, these results indicate that Voxtral was the strongest ASR model among the evaluated options.

The Medical WER results provide complementary evidence within the same benchmark, with Voxtral scoring 0.311 ± 0.406 compared with Whisper's 0.409 ± 0.350 . Because this metric upweights medically salient terminology errors, the lower value suggests more favorable handling of domain-specific vocabulary under the study conditions.

Statistical significance testing using the Wilcoxon signed-rank test detected a difference between Voxtral and Whisper ($p < 0.001$; exact values in Appendix A.6). The associated Wilcoxon effect size was large ($r = 0.877$ for the primary WER contrast), suggesting that the magnitude of the difference is unlikely to be trivial within this dataset, although direct workflow impact was not prospectively measured in this benchmark.

5.2. Language Model Results for Report Generation

The evaluation of LLM models for medical report generation encompasses multiple dimensions that capture both technical NLP performance and clinical utility. Our assessment employed ROUGE-L scores, which measure text similarity using longest common subsequences and are particularly effective for evaluating generated medical reports due to their ability to capture structural similarities in clinical documentation. We complemented this with BERT F1 scores using ModernBERT [57] embeddings, which provide semantic similarity measurements that go beyond lexical overlap to capture contextual understanding.

Figure 4 presents the performance comparison across all LLM approaches and shows measurable differences across the evaluated models.

GPT-4o and Gemini 1.5 Pro showed very similar performance profiles, both achieving ROUGE-L scores of 0.841 ± 0.106 . Within the evaluated benchmark setting, both models outperformed the cross-specialty mismatch control (0.278 ± 0.056) and the domain-adapted MedGemma model (0.397 ± 0.291). This convergence suggests that, within the constraints of our evaluation setup, the top-performing large language models may be approaching similar performance levels for this task.

The BERT F1 scores of 0.958 ± 0.035 for both GPT-4o and Gemini 1.5 Pro indicate high semantic similarity under the automated evaluation procedure. In this setting, these scores suggest that the models generally preserved the main medical content while transforming conversational speech into structured clinical documentation.

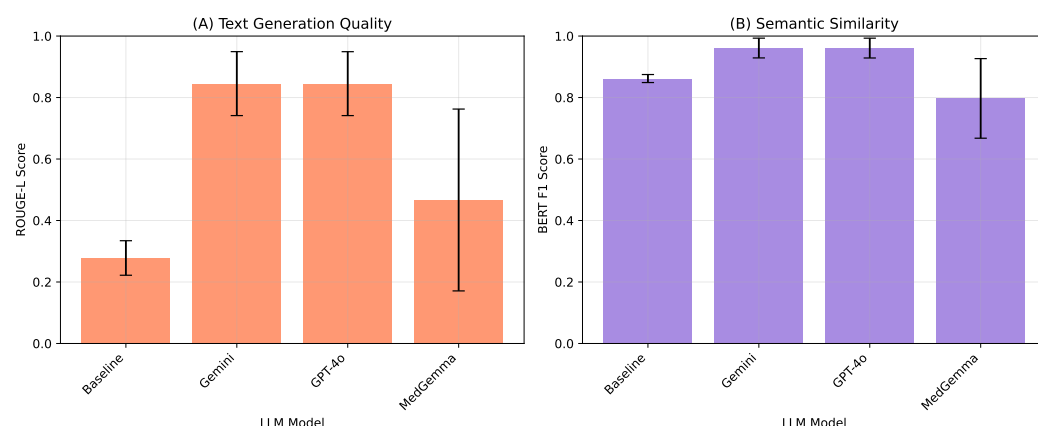


Figure 4. LLM text-quality metrics. (A) ROUGE-L scores. (B) ModernBERT F1 semantic similarity scores.

The clinical assessment dimensions include factual accuracy and clinical relevance, both evaluated using Gemini-2.0-flash as an automated assessor following established methodologies from recent literature [58,59]. Factual accuracy measures medical fact preservation on a 0–1 scale by comparing generated reports against reference transcriptions for factual consistency, while clinical relevance evaluates the utility of generated reports for actual clinical decision-making processes. We employed Gemini-2.0-flash for these assessments due to its demonstrated reliability in medical evaluation tasks [36,37], its consistency across large-scale evaluations, and its ability to provide objective, reproducible assessments without the variability and resource constraints associated with human expert evaluation. This automated approach has been validated in numerous studies as providing comparable accuracy to expert assessment while enabling scalable evaluation across diverse medical specialties [27,60].

A crucial aspect of our evaluation, particularly for deployment-oriented benchmarking, is the assessment of clinical quality using Gemini-2.0-flash as the evaluator, as shown in Figure 5. Both GPT-4o and Gemini 1.5 Pro achieved mean clinical relevance scores of 0.944 ± 0.022 within this automated evaluation setup. These scores indicate strong performance within the evaluation setup, but they do not by themselves establish readiness for unsupervised clinical use. Combined with factual accuracy scores of 0.960 ± 0.072 for GPT-4o and 0.958 ± 0.066 for Gemini 1.5 Pro, the results support continued evaluation of these systems as clinician-facing draft-generation tools when embedded within appropriate review and quality-assurance workflows.

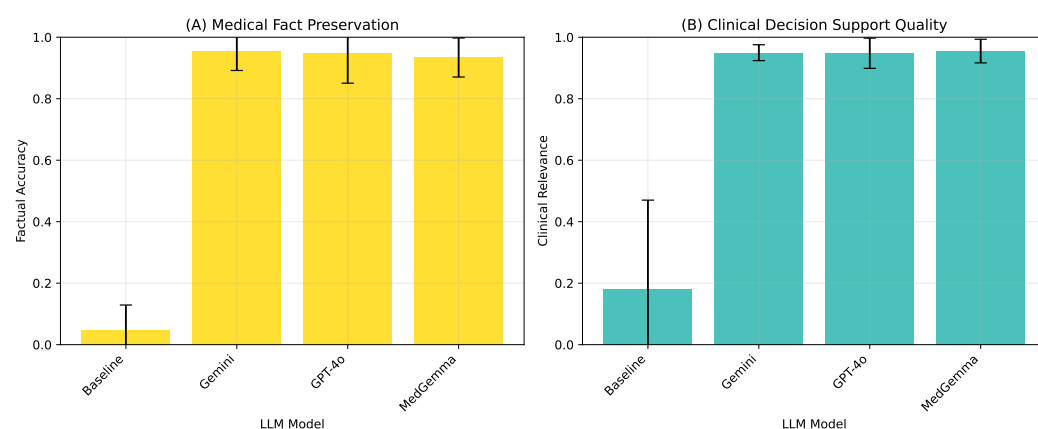


Figure 5. LLM clinical-quality metrics assessed with Gemini-2.0-flash. (A) Factual accuracy. (B) Clinical relevance.

The performance of MedGemma provides a useful counterpoint to expectations about domain-specific adaptation. Despite being specifically adapted for medical applications, MedGemma showed higher variability (0.714 ± 0.395 clinical relevance) and lower overall performance than the general-purpose models in this benchmark. This finding suggests that, under the prompting and evaluation conditions used here, larger general-purpose models may currently perform better than this smaller domain-specific alternative for complex medical report generation tasks.

This result highlights an important trade-off. We deliberately selected a very small model (4 billion parameters) that can be run locally on a personal laptop. Local deployment may offer privacy advantages because it can reduce the need to send sensitive information to external cloud providers. However, our findings suggest that while small models are relevant for privacy-sensitive applications, they may not yet match the performance of larger alternatives on all complex medical documentation tasks. This underscores the ongoing challenge of balancing privacy requirements with model performance in the development of clinical AI tools.

5.3. Safety-Oriented Quality Metrics

Because clinically useful draft notes must avoid fabricated content while preserving specialized vocabulary, we also evaluated hallucination score and terminology accuracy across the LLM configurations, as summarized in Table 4. These metrics complement the broader text-quality measures by focusing on two issues that are especially relevant for deployment-oriented benchmarking: whether the generated report stays anchored to the source dialogue and whether medical terms are reproduced appropriately.

Table 4. Safety-oriented quality metrics by LLM model. Higher values indicate fewer hallucinations and more accurate use of medical terminology. The highest values are highlighted in bold.

LLM Model	Hallucination Score	Terminology Accuracy
GPT-4o	0.980 ± 0.076	0.994 ± 0.024
Gemini 1.5 Pro	0.972 ± 0.069	0.993 ± 0.022
MedGemma	0.968 ± 0.068	0.981 ± 0.051
Cross-specialty mismatch control	0.010 ± 0.030	0.276 ± 0.360

The three evaluated LLM pipelines all achieved mean hallucination scores above 0.96 in this benchmark, whereas the lower-bound cross-specialty mismatch control scored 0.010, indicating that downstream report quality depends strongly on upstream transcript quality. GPT-4o and Gemini 1.5 Pro also achieved terminology-accuracy scores above 0.99, with MedGemma remaining high but somewhat lower (0.981 ± 0.051). These results suggest that once the transcript quality is adequate, the stronger LLM configurations generally preserve medical content without adding large amounts of unsupported information under the study conditions.

Pairwise comparisons were directionally consistent with these descriptive differences. For hallucination score, within the evaluated benchmark setting, GPT-4o outperformed MedGemma after Bonferroni correction (exact $p = 0.0039$; adjusted $p = 0.0117$), whereas the other hallucination comparisons were not significant after correction. For terminology accuracy, none of the pairwise comparisons among the modern LLM models remained significant after multiplicity adjustment. Together, these findings indicate that the main safety-oriented separation in this benchmark is between the modern LLM pipelines and the degraded-input negative control condition, rather than between GPT-4o and Gemini 1.5 Pro.

5.4. Economic Results Under Study Assumptions

The economic dimension of medical AI system deployment is an important factor for real-world feasibility alongside technical performance. Our cost analysis examines modeled cost components associated with different ASR-LLM combinations under the study assumptions.

The pricing structure used in our analysis reflects the published API prices from the respective providers at the time of evaluation (August 2025), as detailed in Tables 5 and 6. A notable exception is MedGemma, which with its 4 billion parameters was deployed locally without external API charges, representing a different economic model that trades upfront infrastructure investment for ongoing usage fees. This analysis enables healthcare administrators and clinical informaticists to make informed decisions about system deployment based on both performance requirements and budget constraints.

Table 5. ASR model pricing as implemented in the experimental code. The highest values are highlighted in bold.

Model	Role	Rate (USD)
Google Cloud STT	ASR	0.078/min
Whisper	ASR	0.006/min
GPT-4o Transcribe	ASR	0.006/min
Voxtral Mini	ASR	0.004/min

Table 6. LLM token pricing used in cost calculations. Input and output prices are for 1 million tokens.

Model	Usage	Input Price (\$)	Output Price (\$)
GPT-4o	Post-process	2.5	10.0
Gemini 2.0 Flash	Post-process	0.10	0.40
Gemini 1.5 Pro	Evaluation	0.075	0.30

Our cost analysis encompasses three primary components: ASR transcription prices, LLM inference token prices, and the resulting total AI-processing cost per generated consultation under the study assumptions. Additionally, we calculated efficiency ratios that normalize performance metrics by total cost, providing an internal comparative indicator for contrasting cost-performance trade-offs across different system configurations.

To provide an internal comparison that emphasizes clinical relevance, our efficiency ratio calculation uses a composite performance metric that combines ASR accuracy on medical terminology with LLM semantic understanding. In this study, the efficiency ratio is used as an internal comparative indicator for benchmarking configurations under a common costing assumption. The efficiency ratio is defined as:

$$\text{Efficiency Ratio} = \frac{\text{Clinical Performance Score}}{\text{Total Cost}} = \frac{((1 - \text{Medical WER}) + \text{BERT F1})/2}{\text{Total Cost}}$$

The efficiency ratio assigns equal weights to Medical WER and BERT F1 for simplicity, though we acknowledge that this choice is study-specific. The Medical WER penalty weight ($\lambda = 3.0$) was selected to reflect the higher clinical risk of errors in medical terminology (e.g., drug names, dosages) compared to general vocabulary, but alternative weighting schemes could be explored in future work and adapted to the objectives.

This study-specific approach prioritizes Medical WER over standard WER for ASR evaluation, recognizing that errors in medical terminology (drug names, anatomical terms, clinical procedures) carry disproportionate consequences for patient care. The ASR performance score is calculated as the complement of Medical WER ($1 - \text{M-WER}$), representing

the proportion of medical terminology correctly transcribed. Combined with ModernBERT F1 scores that capture semantic meaning preservation in LLM-generated reports, this composite metric serves as an internal comparative indicator of benchmark performance per dollar invested, enabling side-by-side comparison of ASR–LLM combinations under the study assumptions.

Figure 6 illustrates the economic analysis across different ASR–LLM combinations, highlighting both absolute costs and efficiency-ratio values for internal comparison across deployment options.

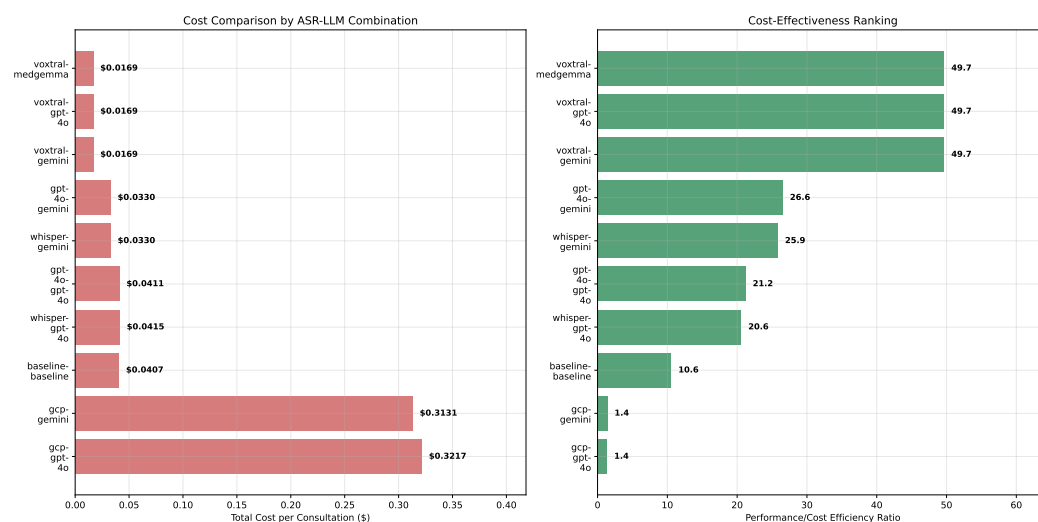


Figure 6. Cost-effectiveness analysis of ASR–LLM combinations. **(Left):** estimated AI-processing cost per generated consultation under the study pricing assumptions. **(Right):** study-specific efficiency-ratio ranking.

The cost breakdown analysis indicates that transcription-related costs typically dominate the modeled expense structure for most combinations. This suggests that improvements in ASR efficiency may have disproportionate effects on total system economics. GCP’s transcription costs exceed \$0.31 per generated consultation under these assumptions, making it economically challenging for large-scale deployment in cost-sensitive healthcare environments. In contrast, Voxtral’s pricing in this benchmark is associated with lower modeled AI-processing cost while maintaining strong technical performance.

Voxtral-based combinations showed favorable cost-effectiveness in this benchmark, with an estimated AI-processing cost of \$0.0169 per generated consultation under the study pricing assumptions while maintaining high performance on medical terminology preservation. Under this internal comparative indicator, Voxtral-based combinations yielded efficiency-ratio values of 49.7, compared with 26.6 for ASR (GPT-4o) + LLM (Gemini 1.5 Pro) combinations (\$0.0330) and 25.9 for ASR (Whisper) + LLM (Gemini 1.5 Pro) combinations (\$0.0330). These differences suggest that the modeled pricing and benchmark accuracy profile of Voxtral may translate into lower AI-processing expenditure, not lower overall consultation cost. Because these estimates depend on vendor pricing, they should be interpreted as pricing-contingent. Under a uniform $\pm 20\%$ variation applied to the August 2025 API prices, the relative efficiency-ratio ranking does not change.

Processing time analysis reveals additional trade-offs that affect both economic and operational considerations. MedGemma combinations achieve the fastest processing times (7–8 s) but with weaker quality metrics across multiple dimensions. Voxtral combinations require moderate processing time (31.7 s) while maintaining stronger quality metrics in this benchmark, suggesting a potentially favorable balance for workflows where both turnaround time and accuracy matter. This processing-time profile suggests that Voxtral-

based systems may be compatible with routine clinical workflows, an interpretation that can be read alongside the deployment-grounded evidence presented in later sections.

5.5. Performance Visualizations and Trade-Offs

Figure 7 summarizes ASR model performance across the evaluated dimensions, while Figure 8 complements this analysis with LLM performance comparisons.

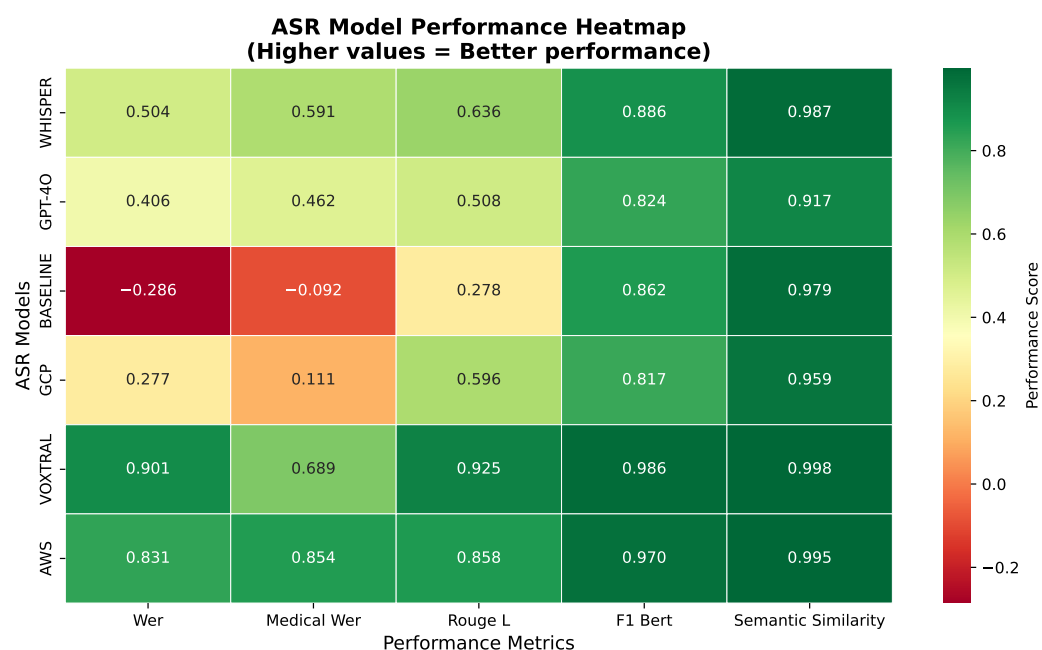


Figure 7. ASR model performance heatmap across the evaluated metrics.

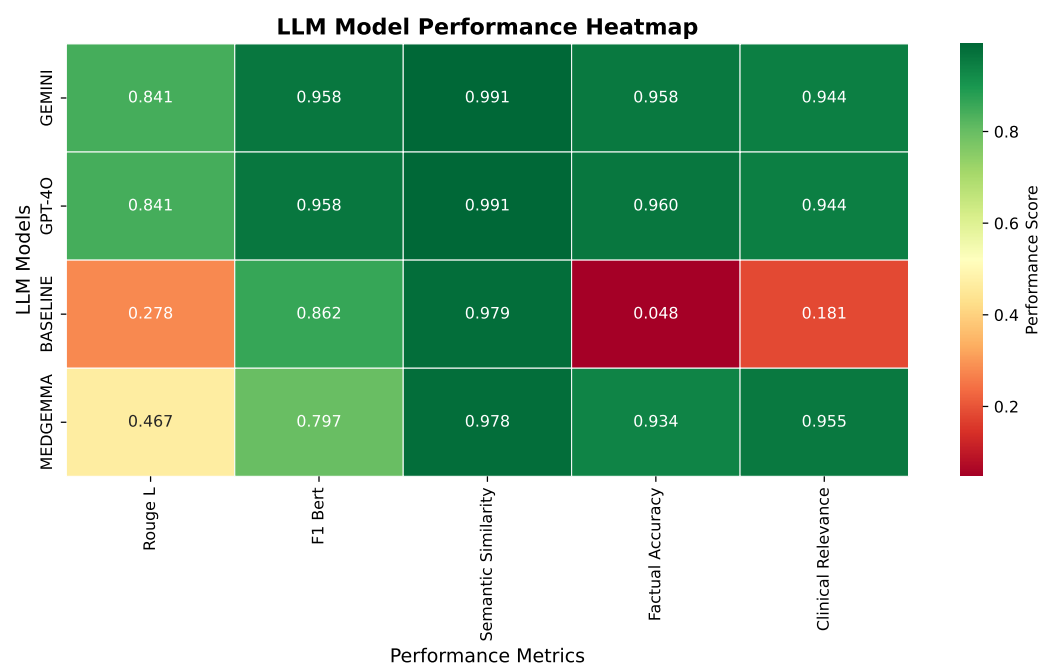


Figure 8. LLM model performance heatmap across the evaluated quality metrics.

The heatmap visualization shows performance patterns across multiple metrics rather than only individual scores. For ASR models, Voxtral performed consistently well across both transcription accuracy measures and downstream quality indicators in this benchmark, suggesting that improved transcription quality may benefit later stages of the documentation pipeline.

The LLM performance heatmap shows similar performance profiles for GPT-4o and Gemini 1.5 Pro across the evaluated medical quality dimensions, while distinguishing these models from both the cross-specialty mismatch control and MedGemma alternatives. This visualization is consistent with the interpretation that, under this benchmark, general-purpose models performed better than the domain-specific alternative for this report-generation task.

5.6. Cross-Language and Medical Specialty Results

Performance across different languages and medical specialties is an important consideration for healthcare systems that serve diverse patient populations. Our evaluation across English and Spanish consultations provides benchmark-level insight into the cross-linguistic stability of different model combinations, while analysis across 10 medical specialties examines performance across the sampled clinical domains.

Cross-language analysis reveals broadly consistent performance patterns across English and Spanish consultations. Voxtral maintained its performance advantage in both languages, achieving WER improvements of 79% for English and 81% for Spanish compared with Whisper in this benchmark. This cross-linguistic consistency is encouraging, but it should be interpreted as benchmark evidence rather than as proof of robust real-world generalization. The slight performance difference between Spanish and English may reflect differences in phonetic complexity or training data characteristics, but the present design does not allow strong inference about the cause.

Medical specialty analysis across cardiology, dermatology, neurology, oncology, and six additional specialties showed consistent performance improvements with Voxtral-based combinations within this benchmark, without clear evidence of specialty-specific degradation. This pattern suggests that the transcription improvements were not limited to a single vocabulary set or conversation type in the sampled synthetic dialogues. However, the present results do not eliminate the need for specialty-specific evaluation in real deployment settings.

The sample distribution of consultations per specialty per language supports the planned benchmark comparisons while maintaining practical constraints on evaluation scope. This balanced design improves coverage of the sampled conditions, but it does not by itself justify confident generalization to routine clinical environments.

5.7. Statistical Validation and Pilot-Stage Feasibility

The statistical foundation of our evaluation supports cautious interpretation of the benchmark results. With 230 total experiments distributed across balanced model configurations ($N = 20$ per condition for primary comparisons, $N = 10$ for GCP due to language constraints), the study has substantial sample coverage for the planned comparisons. We employed the non-parametric Wilcoxon signed-rank test for paired comparisons of WER and metric scores across identical audio samples, ensuring robustness against non-normal error distributions. To control for family-wise error rates in multiple pairwise comparisons, we applied the Bonferroni correction ($\alpha_{adj} = 0.05/m$).

Wilcoxon signed-rank tests [61] indicate statistically significant differences between Voxtral and competing ASR solutions ($W = 0.0$, exact p -values and Bonferroni-adjusted comparisons in Appendix A.6), with large Wilcoxon effect sizes ($r = 0.877$ for the two $n = 20$ contrasts and $r = 0.886$ for the $n = 10$ GCP contrast). These effect sizes suggest that the observed differences are unlikely to be trivial within this benchmark, although direct workflow effects were not measured prospectively.

Taken together, these benchmark results support cautious follow-up in deployment-oriented evaluation but do not establish real-world clinical effectiveness.

The modeled cost profile suggests that Voxtral-based solutions may be economically plausible enough to justify further pilot-stage evaluation. At an estimated \$0.0169 in AI-processing cost per generated consultation in this benchmark, the technology appears relatively inexpensive on a per-consultation basis under the study assumptions. However, this is not evidence of reduced physician administrative burden, improved documentation quality in practice, or cost-effectiveness after implementation costs and workflow constraints are considered.

6. Discussion

This evaluation of ASR–LLM combinations for automated clinical documentation provides evidence on the current performance of these systems under the study conditions. Our analysis across 230 experiments spanning 10 medical specialties in English and Spanish suggests potential implementation relevance while also highlighting important constraints for deployment in healthcare settings.

Taken together, the findings provide direct but bounded answers to the three research questions. For RQ1, the controlled benchmark shows measurable differences among ASR–LLM configurations, with Voxtral providing the strongest transcription performance and GPT-4o/Gemini 1.5 Pro providing the strongest report-generation performance under the study conditions. For RQ2, the deployment-grounded traces do not validate any single benchmarked configuration in isolation, but they show that the system was used routinely in real clinical environments and that observed usage was concentrated in documentation-heavy specialties. For RQ3, the combined evidence highlights practical evaluation considerations for future work, including the need to interpret synthetic benchmarks alongside operational data, clinician oversight, workflow impact, editing burden, and prospective clinical validation.

6.1. Key Findings and Clinical Implications

Our evaluation suggests that modern ASR–LLM combinations can achieve performance levels that may be relevant for supervised draft-generation use. In this benchmark, Voxtral achieved the lowest mean speech-recognition error, with a WER of 0.099 ± 0.207 compared with Whisper's 0.496 ± 0.307 . A similar pattern was observed for Medical WER, where Voxtral scored 0.311 ± 0.406 compared with Whisper's 0.409 ± 0.350 . These findings suggest more favorable handling of medically salient terminology under the study conditions, but they should not be interpreted as proof of equivalent performance in prospective clinical settings.

For language model performance, GPT-4o and Gemini 1.5 Pro showed very similar profiles, both achieving ROUGE-L scores of 0.841 ± 0.106 and clinical relevance scores of 0.944 ± 0.022 within our automated evaluation setup. This pattern suggests that the top-performing large language models may be converging for this task, at least within the scope of our benchmark and scoring procedure.

Our Efficiency-Ratio (ER) comparison suggests that Voxtral-based combinations may offer a favorable trade-off between modeled AI-processing cost and benchmark performance in this setting. Voxtral combinations achieve estimated AI-processing costs of 0.0169\$ per generated consultation under the study assumptions, with ER values of 49.7 compared with 26.6 for GPT-4o–Gemini 1.5 Pro combinations (0.033\$ under the same assumptions). This comparison should be interpreted cautiously because the ratio is intended as an internal benchmarking aid and because economic conclusions remain sensitive to vendor pricing, service availability, and local deployment constraints.

The relatively consistent performance across medical specialties is relevant for future implementation planning. Our analysis across cardiology, dermatology, psychiatry, and

other domains showed limited variation in the core performance metrics used here, suggesting that performance was reasonably stable across the sampled specialties within this benchmark. Even so, this interpretation remains bounded by the synthetic-data design and should be validated prospectively before stronger generalization claims are made.

6.2. Multilingual Capabilities and Global Healthcare Implications

The performance observed across English and Spanish suggests potential relevance for multilingual healthcare settings. Voxtral maintained its relative advantage in both languages, achieving WER reductions of 79% for English and 81% for Spanish compared with Whisper in this benchmark. This cross-linguistic consistency is encouraging, but it should be interpreted as evidence from synthetic multilingual dialogues rather than as confirmation of broad real-world generalization.

The multilingual evaluation also provides some support for the usefulness of the synthetic dataset methodology. The consistency of performance patterns across languages suggests that the generated medical dialogues capture at least part of the linguistic and clinical structure relevant to this task, providing a practical basis for future benchmarking work. Studying both English and Spanish allowed us to evaluate the system in the dominant operational languages across the markets examined here, providing a practically relevant multilingual benchmark. Extension to additional languages remains feasible, but further external validation would still be needed before strong generalization claims are made.

6.3. Clinical Safety and Quality Assurance

The hallucination detection and terminology accuracy results summarized in Section 5.3 provide additional information relevant to safety-oriented evaluation. All modern LLM models achieved hallucination scores exceeding 0.96, with GPT-4o and Gemini 1.5 Pro reaching terminology-accuracy scores above 0.99 in the automated framework. These results suggest high internal performance on synthetic data, but they should be interpreted as benchmark indicators rather than as direct evidence of safety in clinical deployment.

Our evaluation included a degraded-input negative control, implemented here as a cross-specialty mismatch control in which semantically incongruent audio was fed to the models to simulate a worst-case scenario with major ASR errors. The markedly lower hallucination score of this control condition (0.010 vs. >0.96 for the modern LLM pipelines) suggests that downstream report quality depends strongly on upstream transcription quality. For implementation, this implies that ASR choice should be treated as an important safety-related design decision, although the direct relationship to patient outcomes was not measured here.

The Gemini-2.0-flash evaluation setup assessed factual accuracy, information coverage, clinical relevance, hallucination detection, and terminology accuracy, providing a practical automated proxy for comparing system outputs at scale. However, this setup introduces a dependency on an external API-based evaluator and its current model behavior, so the resulting scores should be interpreted as heuristic quality indicators rather than as substitutes for clinician review or independent expert adjudication [37].

6.4. Economic and Scalability Considerations

Our benchmark-level cost analysis suggests potential economic advantages for some AI-powered documentation configurations, with Voxtral-based combinations showing favorable modeled cost-effectiveness within the benchmark. With estimated AI-processing costs ranging from 0.0095\$ to 0.383\$ per generated consultation (mean: 0.062\$), these systems appear economically plausible under the study assumptions. Assuming a conservative estimate of 15 min saved per consultation and an average physician hourly rate of 50\$, the break-even cost threshold would be approximately 12.50\$ per consultation—much

higher than the modeled AI system costs reported here. However, our cost analysis focuses on AI-processing costs per consultation and does not account for additional deployment expenses, such as integration with EHR systems, data governance, cybersecurity, local infrastructure, clinician review time, procurement, training, maintenance, monitoring, or liability-related costs. Therefore, the economic conclusions should be interpreted as model-specific and benchmark-level, not as a comprehensive cost-effectiveness assessment for real-world deployment. This narrower benchmark-level view is consistent with broader arguments that the economic value of healthcare AI should be assessed not only in terms of implementation cost but also through longer-term organizational and system-level effects [62].

Voxtral-based combinations show the most favorable economic profile in this benchmark, with an estimated AI-processing cost of 0.0169\$ per generated consultation under the study assumptions, compared to 0.033\$ for GPT-4o–Gemini 1.5 Pro combinations. The cost breakdown analysis shows that transcription-related costs typically dominate the total modeled expense structure for most combinations, with GCP’s transcription-related cost exceeding 0.31\$ per generated consultation under the study assumptions, which may be economically challenging for larger-scale deployment in cost-sensitive healthcare environments.

The emergence of more efficient medical-domain models like MedGemma, while currently showing performance limitations compared to general-purpose models within this benchmark, represents an interesting economic trade-off. MedGemma’s local deployment eliminates ongoing API charges but requires infrastructure investment, achieving faster processing times (7–8 s) at the cost of reduced quality metrics across multiple dimensions.

For large healthcare systems processing thousands of consultations daily, these cost characteristics indicate possible economic headroom under the study assumptions. For example, a 500-consultation-per-day healthcare system would incur an estimated 8.45\$ daily in AI-processing cost using the lowest-cost Voxtral combinations in this benchmark. This is a modeled cost illustration, not evidence of realized time savings, return on investment, or operational benefit in practice.

6.5. Methodological Considerations for Medical NLP

Our study uses a modular ASR–LLM architecture that facilitates systematic evaluation of individual system components. This approach enables component-level analysis of performance, cost, and safety-related metrics that would be harder to isolate in monolithic systems. Furthermore, our evaluation setup, using Gemini-2.0-flash as an automated assessor, combines traditional metrics (e.g., WER, ROUGE) with domain-specific measures (e.g., Medical WER, clinical quality scores), providing a broader view of technical performance under the study conditions.

To address the practical issue of data privacy and access in healthcare research, we developed a synthetic data generation methodology using Gemini-2.0-flash. In this hybrid study, the controlled benchmark relies on synthetic dialogues for a specific scientific purpose: it enables controlled, reproducible, component-level comparison while preserving exact ground truth for the full dialogue. This is especially valuable in clinical documentation research, where access to real consultation audio is often heavily restricted and where aligned reference text is needed to measure ASR fidelity, hallucination, and downstream report-generation quality. The consistency of the benchmark patterns across diverse medical scenarios and two languages (English and Spanish) suggests that synthetic data can serve as a useful, privacy-preserving foundation for initial system development and benchmarking under controlled conditions. At the same time, because both the synthetic dialogue generation and parts of the automated evaluation depend on Gemini-based ser-

vices, the study remains exposed to external model behavior, API availability, and vendor updates. To facilitate further research and reproducibility, the dataset, including both text and audio files, is available upon request.

Just as importantly, the manuscript does not rely on the synthetic benchmark alone. It uses a complementary bench-to-field evidence framework: the controlled benchmark provides measurement of technical accuracy under known ground truth, while the deployment-grounded evidence provides operational evidence from routine use. Across a 120-day study period, the deployment-grounded evidence captured 3107 finalized clinical traces from 301 active users, corresponding to 29,535 min (~492 h) of authentic doctor–patient interaction across LATAM and Europe [5,14]. Although these data do not establish prospective clinical outcomes, they do show that the system was used in real clinical environments at a scale that would be difficult to capture in a small prospective pilot.

The comparable performance of general-purpose LLMs (GPT-4o and Gemini 1.5 Pro) relative to the domain-adapted alternative (MedGemma) complicates conventional assumptions about the necessity of medical domain adaptation for clinical NLP, while the present results suggest that scale and training diversity may currently offer an advantage for complex medical report-generation tasks, smaller domain-specific models remain important, especially for local deployment and privacy-sensitive settings. The performance variability observed in some MedGemma-generated reports may be addressable through further fine-tuning, such as adjusting temperature or modifying the structured output format. Accordingly, the current performance gap should be interpreted as a benchmark finding, not as evidence that smaller privacy-preserving models are unlikely to become competitive.

The evaluation also yields several observations that may inform clinical implementation. Within this benchmark setting, Voxtral and GPT-4o Transcribe outperformed Whisper, suggesting that more recent ASR systems may offer measurable gains in medical documentation tasks. The cost-effectiveness analysis indicates that some of these technical differences may also matter economically, with Voxtral-based combinations showing a 1.9× improvement in cost-effectiveness over selected alternatives under the assumptions used here. Even so, these findings should be interpreted with caution because both performance and cost may change as models, APIs, and local deployment constraints evolve.

7. Real-World Context and Implications

Unlike the controlled benchmark presented in Section 5, this section examines deployment-grounded operational observations derived from routine clinician-supervised use, providing descriptive evidence of operational use and implementation patterns rather than direct evaluation of the benchmarked ASR–LLM configurations or measures of clinical effectiveness.

This section situates the paper as a whole within its broader clinical and implementation context. Within the study’s bench-to-field evidence framework, the controlled benchmark defines technical performance under a known ground truth, whereas the deployment-grounded evidence shows how the same system was used in routine practice under uncontrolled conditions. A central motivation of this study was to examine whether technical benchmark patterns and operational implementation patterns provide consistent or complementary insights. Together, these complementary perspectives provide a broader understanding of how technical performance translates into real-world implementation while identifying priorities for future prospective clinical evaluation. Finalized exports can therefore be read as a bounded behavioral proxy for perceived usefulness: when clinicians export and finalize an AI-generated draft, they have judged it useful enough to refine and incorporate into routine documentation. That does not establish clinical correctness, safety, or efficiency, but it does add operational evidence beyond the synthetic benchmark and

avoids some of the recall limitations of self-reported workflow surveys [5,13,14,63]. The subsections that follow move from clinical motivation and evidence gaps to analytical framing, the study's specific contribution, observed usage patterns, and an overall interpretation of what these findings do and do not support.

7.1. Clinical Relevance and Motivation

Clinical documentation burden remains a significant contributor to clinician workload and burnout, and ambient-AI documentation tools are increasingly proposed as a potential response [3,5,13]. However, empirical evidence remains mixed, while some studies report perceived efficiency gains, systematic reviews continue to highlight limited and heterogeneous evidence regarding effects on documentation quality, workflow efficiency, burnout, and system-level outcomes [13,14,18].

Recent multisite evidence further tempers expectations. A study published in *JAMA* on 1 April 2026, reported modest reductions in EHR time (13.4 min) and documentation time (16.0 min) per 8 scheduled patient hours, small productivity gains, and no significant change in after-hours “pajama time” across five U.S. health systems [48]. Taken together, these findings suggest that the central question is not whether AI scribes are promising in principle, but how they should be evaluated and integrated cautiously in practice.

7.2. Evidence Gaps in Current Literature

One reason this question remains open is that existing evidence is often fragmented. Many studies emphasize controlled technical performance or self-reported usability, while fewer examine how these systems are used over time in routine clinical environments. This distinction is important because technical performance does not necessarily translate into sustained workflow value.

For example, a Dutch digital-scribe study (22 participants, 430 mock-consultation summaries) found that editing AI-generated drafts was modestly faster than fully manual summarization, but unedited automatic notes had substantially lower documentation quality and less diverse language despite higher word counts [7]. This suggests that any productivity gain depends on clinician editing, quality control, and workflow integration rather than autonomous note finalization.

7.3. Analytical Framework

This section connects controlled benchmarking with observed routine use. Rather than treating technical evaluation and field behavior as interchangeable, we consider them as complementary evidence streams. The controlled benchmark assesses ASR and LLM performance under controlled conditions with preserved ground truth. The deployment-grounded evidence examines how the system was used in routine deployment through retrospective log-mining when real-world consultations are noisy, heterogeneous, and not amenable to gold-standard transcript alignment.

The real-world traces were generated using a production ASR–LLM pipeline that evolved during the 120-day observation period, while the benchmark evaluates specific fixed configurations (e.g., GPT-4o, Gemini 1.5 Pro, MedGemma), the deployed system may include configuration updates, prompt refinements, or model substitutions over time. Therefore, the real-world evidence should be interpreted as complementary operational evidence rather than a direct evaluation of any single benchmarked configuration. Methodologically, digital trace analysis provides a non-interruptive alternative to recall-based survey instruments by relying on timestamped interaction data generated in situ rather than post hoc self-report [5,14]. This perspective linking benchmark results to field use is particularly relevant for systems intended to operate under human oversight, where technical performance and workflow integration jointly determine practical utility.

7.4. Contribution of the Present Study

Building upon recent ambient clinical documentation evaluation studies [15,17], this study is positioned as an early-stage evaluation rather than a purely methodological advance, while individual metrics such as Medical WER and the use of synthetic data generation have been explored previously [46], our primary contribution lies in a bench-to-field evidence framework: a controlled, multi-specialty synthetic benchmark is paired with real-world, human-in-the-loop deployment trace analysis to support research tracking, implementation insight, and interpretation of system use beyond laboratory conditions. These two evidence streams answer different questions. The controlled benchmark evaluates technical performance under standardized conditions, whereas the deployment traces describe operational use in routine clinical settings. Accordingly, they should be interpreted as complementary rather than interchangeable forms of evidence.

Under this framing, the primary contribution of the paper is the explicit combination of two complementary evidence streams within one study. The controlled benchmark is a modular ASR–LLM evaluation based on multilingual synthetic clinical dialogues, which makes exact measurement of WER, Medical WER, semantic preservation, and safety-oriented LLM errors possible. The deployment-grounded evidence stream is an analysis of 3107 finalized traces from LATAM and Europe, which adds complementary operational evidence that the system was used under routine, uncontrolled conditions.

Accordingly, the study should not be read as asking one evidence stream to replace the other. The synthetic benchmark addresses technical validity where private real-world data cannot provide exact ground truth, whereas the deployment-grounded evidence addresses ecological validity by showing sustained use under routine conditions. Together they provide a more informative hybrid evidence base than either a synthetic benchmark or an observational usage study alone, while still stopping short of claiming prospective clinical outcome validation [5,13,14,18,48].

7.5. Observed Usage Patterns During the 120-Day Deployment Period (November 2025–February 2026)

To complement the experimental benchmarks, we conducted a retrospective clinical log-mining analysis of routine system use. Because it relies on digital trace data captured during actual deployments rather than self-reported surveys, this analysis reflects observed physician behavior with minimal workflow disruption. It therefore provides evidence of operational use rather than prospective clinical outcomes [5,14]. It should also be read as product-level observational evidence, not as a direct comparison of real-world performance for a single fixed ASR–LLM configuration over the full study window.

Although the deployment analysis primarily focused on operational usage patterns rather than direct measurement of documentation quality, the broader motivation for this work is the generation of trustworthy clinical data. The value of AI-assisted documentation extends beyond note generation itself, as the resulting information may subsequently support clinical workflows, analytics, interoperability initiatives, quality-improvement activities, and future AI-enabled healthcare applications. Consequently, the quality, consistency, and trustworthiness of generated clinical information remain important considerations when evaluating AI-assisted documentation systems.

This perspective motivated our emphasis on transcription accuracy, medical terminology preservation, hallucination rates, and report quality within the benchmark component of the study, while modern AI systems can generate highly fluent documentation, fluency alone should not be interpreted as correctness. Large language models may occasionally produce outputs that appear plausible yet contain inaccuracies, reinforcing the importance of human review and appropriate governance mechanisms within clinical workflows.

The unit of analysis was the *clinical trace*, defined as a finalized and exported report. Between November 2025 and February 2026, we recorded 3107 traces across 120 days, corresponding to 29,535 min (~492 h) of authentic doctor–patient interaction. These traces were generated by 301 active users across Latin America (LATAM) and Europe. LATAM accounted for 2314 reports (23,190 min) and Europe for 793 reports (6345 min). Mean consultation duration was 10.02 min in LATAM and 8.00 min in Europe; in LATAM, the 75th percentile was 11.56 min, consistent with use during full consultations rather than only short tests. Usage distributions were also observed to be skewed across users in certain specialties, suggesting concentration effects typical of early-stage deployment [5,14].

These data should be interpreted as observational evidence of operational use, not as direct evidence of clinical effectiveness. They do not establish prospective clinical outcomes or exposure-adjusted adoption rates. Finalized exports are a proxy for perceived usefulness in routine practice, but they do not directly measure documentation quality, patient safety, time savings, or user satisfaction [5,13,14].

Across the observation period, the deployment logs showed continued operational use among active deployment sites, with 3107 finalized reports and 29,535 min of doctor–patient interaction processed. In the operational usage analysis, LATAM generated almost three times as many reports as Europe (2314 vs. 793), while the interpretation of broader uptake emphasized activation and habit formation rather than acquisition alone.

While this operational use should not be interpreted as evidence of adoption rates or clinical effectiveness, it provides complementary evidence that users incorporated the system into routine documentation workflows. We interpret these observations cautiously as indicators of perceived utility among active users rather than definitive measures of impact.

Importantly, clinician review remained part of the workflow throughout the observation period. The system was designed as a clinician-supervised documentation support tool rather than an autonomous documentation solution. This human-in-the-loop approach was considered essential given that highly capable AI systems may occasionally generate plausible but inaccurate outputs.

Table 7 summarizes the global in-field usage patterns, Table 8 highlights the main specialty-level usage patterns observed in LATAM, and Figure 9 provides a visual summary of specialty-level usage intensity in LATAM.

Table 7. Global in-field usage summary during the 120-day study period (November 2025–February 2026).

Region	Active Users (≥1 Report)	Reports Generated	Total Minutes Processed	Mean Minutes Per Report
LATAM	205	2314	23,190	10.02
Europe	96	793	6345	8.00
Total	301	3107	29,535	—

Table 8. Specialty-level usage highlights in LATAM.

Specialty	Active Users	Reports Generated	Total Minutes Processed
Psychology	14	120	3345
Traumatology & Orthopedics	15	161	2458
Psychiatry	7	70	1693
Cardiology (Electrophysiology) *	1	311	2482

* Single-user concentration effect; interpret as high-intensity individual use rather than broader specialty-level usage.

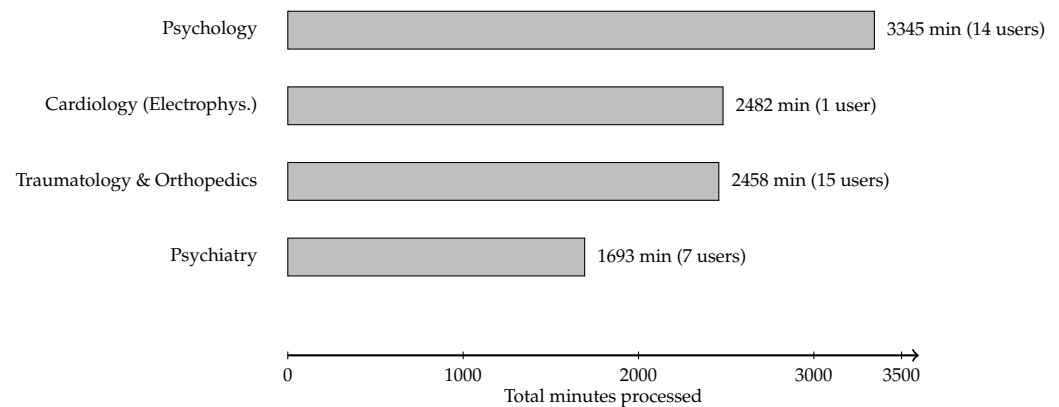


Figure 9. Specialty-level usage intensity in LATAM (derived from Table 8). Horizontal bars show total minutes processed; labels report minutes and active users.

7.5.1. Specialty-Level Usage Patterns

Usage was not evenly distributed across specialties, but concentrated in areas with distinct documentation demands. In LATAM, the highest activity was observed in Psychology (3345 min; 120 reports; 14 users), Traumatology & Orthopedics (2458 min; 161 reports; 15 users), and Psychiatry (1693 min; 70 reports; 7 users). These patterns are consistent with two possible task–technology fits: narrative synthesis settings (e.g., Psychology and Psychiatry), where long-form dialogue benefits from structured summarization, and motor-load-constrained settings (e.g., Traumatology & Orthopedics), where hands-free documentation may reduce workflow friction. These interpretations remain tentative because exposure, rollout strategy, and specialty-level user distribution were not controlled for [5,14].

7.5.2. Concentration Effects and High-Intensity Usage

A notable concentration effect appeared in Cardiology (Electrophysiology), where a single user generated 311 reports, or approximately 10% of total LATAM report volume. This is best interpreted as high-intensity individual use, suggesting that the system can become deeply integrated into a specific clinician’s workflow. However, it should not be treated as representative of broader specialty-level usage patterns without additional exposure and user-distribution data [5,14].

7.5.3. Sensitivity Analysis and Observational Bias

To assess potential selection and concentration effects, additional analyses were conducted at the user level. Usage distributions were found to be skewed, with certain specialties influenced by high-intensity users. For example, excluding the highest-contributing user in Cardiology reduced Cardiology’s observed share of LATAM exports from 13.4% to 0%, indicating strong dependence on a single high-intensity user and highlighting the potential for concentration effects in early-stage deployment. These analyses suggest that some specialty-level patterns are partially influenced by concentration effects. However, because exposure-adjusted denominators (e.g., total eligible consultations or clinicians with access) were not available, these findings should be interpreted as descriptive usage patterns rather than adoption rates.

7.5.4. Interpretation of Finalized Exports as a Behavioral Signal

Although we did not quantify editing effort (e.g., edit distance between draft and final report), the 3107 finalized exports still provide a behavioral signal of continued use. A finalized report can therefore be read as a pragmatic indicator of perceived utility: the draft was useful enough to refine and integrate into documentation. However, this should not be

interpreted as evidence of clinical correctness or completeness, since all outputs remained subject to physician verification and editing [5,14,63].

7.6. Summary Interpretation and Human-in-the-Loop Framing

Overall, the log-mining analysis adds an important but bounded source of evidence. It shows sustained operational use under routine conditions, with activity concentrated in specific specialties and, in some cases, among high-frequency users. It does not, by itself, demonstrate clinical quality, safety, time savings, burnout reduction, or adoption rates. The findings should therefore be read as descriptive evidence of observed usage patterns rather than definitive evidence of clinical effectiveness or workflow efficiency [13,14].

Recent work has emphasized the importance of human-centered or “human-allied” AI systems, in which the clinician remains the primary decision-maker and AI functions as a supporting component rather than an autonomous agent. In this context, our findings are most appropriately interpreted within a human-in-the-loop framework. The persistent need to verify generated reports, the variability observed across model configurations, and the observational nature of the deployment evidence all support viewing the present pipeline as a clinician-facing documentation tool rather than an autonomous system. This interpretation is consistent with positioning the system as a draft-generation tool integrated into existing workflows under human oversight [7,11–14].

More broadly, the findings highlight that trustworthy AI-assisted documentation depends not only on model performance but also on the quality of the data being generated, stored, and made available for later use. As healthcare systems increasingly rely on structured digital information for interoperability, analytics, decision support, organizational learning, and health-system governance, the generation of reliable, reviewable, and well-governed clinical data becomes increasingly important. Future research should therefore evaluate AI-assisted documentation systems not only as note-generation tools, but also as components of a broader healthcare data infrastructure. Prior work on converting clinical reports into FHIR bundles shows that this downstream structuring task is feasible with LLMs but remains sensitive to prompt design, iterative correction, hallucination control, and resource-mapping accuracy, reinforcing the need to treat documentation quality and data normalization as connected problems [27,64–66]. This perspective is consistent with the need to connect AI-generated documentation to high-quality data foundations, including interoperable representations such as FHIR, clinically meaningful terminologies such as SNOMED CT, robust storage and collection practices, and governance processes that support trustworthy downstream analysis. Accordingly, we do not interpret operational use as evidence of clinical impact on its own; rather, we view strong data quality and infrastructure as prerequisites for producing reliable information from which higher-level AI-enabled clinical, operational, or organizational value can later be assessed.

8. Future Work and Limitations

This study should be interpreted as an early, pilot-stage evaluation rather than as a definitive implementation study. It combines a controlled benchmark with observational usage traces, but it does not establish prospective clinical effectiveness, validated workflow benefit, or readiness for autonomous deployment.

Rather than proposing a validated implementation framework or positioning a “playbook” as a central contribution, we limit the practical implications of the present study to three cautious points. First, any adoption effort should begin with controlled pilots in clearly defined local settings before broader rollout is considered. Second, clinical verification should remain mandatory, with clinicians retaining responsibility for reviewing, correcting, and approving generated documentation. Third, deployment should be adapted

to specialty-specific documentation needs and to the realities of local workflow integration, since operational value is likely to depend on context rather than on a single uniform model of use. If any thresholds, staging logic, or rollout phases are inferred from these findings, they should be interpreted only as indicative, local, and not universally validated [12,62].

8.1. Future Research Directions

The future-work section should follow the same logic as the rest of the discussion: brief, specific, and aligned with what this study can actually support.

8.1.1. Technical Robustness and Real-World Validation

Future work should test the pipeline under more realistic clinical conditions, including diverse speakers, accents, recording devices, background noise, and longer consultations. This would clarify how acoustic variability affects both ASR performance and downstream report quality. The next step should also be prospective validation in routine care, with direct measures such as physician satisfaction, documentation quality, workflow fit, and unintended effects on clinical practice.

8.1.2. Advanced Model Development

Model development should focus on smaller and more specialized clinical models that can be fine-tuned for documentation tasks while remaining feasible for local, privacy-sensitive deployment. Our results suggest that domain-specific models still lag behind the strongest general-purpose systems within this benchmark setting, but they remain promising because they may offer better controllability, lower cost, and easier secure deployment. Another clear direction is multimodal documentation support that combines speech with structured EHR fields, physiological data, or relevant visual context.

8.1.3. Voice as a Clinical Signal Beyond Transcription

In the present study, speech is used mainly as a source of text for documentation. This is appropriate for the goal of the paper. However, speech also contains acoustic information that may have clinical relevance. Previous work has discussed vocal biomarkers for diagnosis, risk prediction, and remote monitoring [67], and recent reviews show that speech-signal analysis may support clinical decision-making in neurological and mental-health contexts [68]. In depression research, Vilenchik et al. showed that the way speech data are collected can affect the performance and generalization of voice-based detection models [69]. This is relevant because AI-scribe systems already capture clinical conversations. Future systems could therefore evolve from documentation tools into multimodal clinical assistants that support documentation, screening, and longitudinal monitoring, but only under explicit consent, appropriate governance, and rigorous clinical validation. This is only a future direction. The present study does not analyze vocal biomarkers and does not support diagnostic claims from voice.

8.1.4. Multilingual and Cross-Cultural Expansion

Future research should extend the benchmark to more languages, specialties, and healthcare settings. This is important because documentation structure, consultation style, and clinically relevant information vary across specialties and across health systems. Cross-cultural adaptation should therefore include local terminology, documentation norms, and mixed-language or code-switching scenarios, especially in multilingual regions.

8.2. Limitations and Methodological Considerations

This subsection summarizes the main constraints on both the benchmark and the deployment-based evidence, and therefore on the claims this study can support.

8.2.1. Dataset, Evaluation, and Generalizability Limits

The benchmark has three main limits. First, the synthetic dialogues support controlled and reproducible comparison, but they cannot fully capture the complexity of real-world clinical consultations, including spontaneous speech, interruptions, overlapping speakers, diverse accents, emotional variation, incomplete histories, complex comorbidity, difficult communication, ambient noise, and device-related recording variation [44,70,71]. Because these features are likely under-represented in TTS-generated dialogues, the benchmark may overestimate ASR and report-generation performance in actual clinical environments. Moreover, no direct comparison between synthetic and real-world dialogues was conducted to quantify this gap. The benchmark should therefore be treated as a reproducible test bed for component-level comparison, not as a substitute for prospective validation in real care; the large effect sizes reported here (e.g., $r \approx 0.88$) are specific to the controlled benchmark conditions and should not be assumed to generalize to noisy routine settings without further validation.

Second, automated evaluation cannot fully replace expert clinical review. Although Gemini-2.0-flash provides a standardized way to compare many model combinations, clinicians may still detect subtle errors, omissions, or contextual problems that automated metrics miss. The use of Gemini models for synthetic dialogue generation, one evaluated report-generation model, and automated scoring also introduces a potential self-preference or intra-family bias. Prior LLM-as-a-judge work supports automated judging as a scalable approximation to human evaluation but also documents systematic judge biases, and more recent studies show that models may recognize and favor their own or same-family generations [54–56]. Consequently, the favorable performance of Gemini 1.5 Pro may be slightly inflated by the Gemini-2.0-flash evaluator, even though the evaluated generator and judge were not identical models. Blinded physician review and, ideally, adjudication by evaluators from independent model families should therefore complement automated assessment in future work [72].

Third, the benchmark is limited in both linguistic and temporal generalizability. Results in English and Spanish may not transfer to other languages, dialects, or healthcare systems [73]. In addition, the API-based experiments were run from 5 August to 15 August 2025, so model behavior, pricing, and service quality may change over time [74]. Additionally, the reliance on cloud-based models (e.g., GPT-4o, Gemini 1.5 Pro) introduces reproducibility challenges, as their underlying versions and behaviors may evolve without public transparency, limiting the long-term durability of benchmark rankings. The benchmark should therefore be interpreted as a time-bounded comparative snapshot rather than a fixed statement of future performance.

8.2.2. Operational, Data, and Implementation Constraints

The observational results in Section 7 should likewise be read as complementary evidence of operational use, not as exposure-adjusted measures of adoption. We did not have denominators such as the number of eligible clinicians, eligible consultations, or potential opportunities for use. We also lacked direct measures of editing effort, recurrence depth, and workflow impact. Furthermore, the benchmark's reliance on clean, TTS-generated audio means that the observed performance—including the large effect sizes—may not reflect real-world conditions, where factors such as background noise, poor microphone quality, or overlapping speech could degrade ASR and LLM performance. For that reason, specialty-level differences are better interpreted as observed usage patterns than as direct measures of penetration or preference [13,14].

Operational logs also show only part of system value. They can reveal repeated use and specialty-level heterogeneity, but they cannot explain why clinicians returned, which features mattered most, or whether repeated use improved documentation quality. Some

specialty patterns were also shaped by high-intensity users, rollout effects, training, or local workflow constraints rather than product performance alone.

The present study did not evaluate editing frequency, correction rates, note acceptance rates, clinician satisfaction, workflow outcomes, or patient outcomes. As a result, operational use cannot be interpreted as direct evidence of documentation quality improvements, efficiency gains, or clinical effectiveness. Future studies should incorporate these measures to better understand the relationship between operational use and real-world value, including whether clinicians accepted AI-generated notes with minimal correction or substantially revised them before finalization.

The real-world analysis is also vulnerable to selection and measurement bias. Participating sites had already chosen to trial AI-assisted documentation, so they may over-represent digitally mature settings with stronger implementation support. Specialty labels may also be incomplete or inconsistently assigned across deployments. In addition, because the study evaluates the authors' own product, the implementation findings require extra caution even though the benchmark comparisons were standardized across multiple ASR and LLM configurations.

Finally, neither the benchmark nor the telemetry establishes clinical effectiveness. The real-world usage analysis provides a complementary observational evidence stream, but it does not directly measure routine-care documentation accuracy, time savings, note completeness, after-hours charting, clinician workload, user satisfaction, or patient outcomes. These outcomes would require a dedicated user-centered and prospective evaluation, such as clinical trials, structured physician surveys, or in-depth workflow studies. Future research should also examine potential unintended consequences of prolonged reliance on AI-assisted documentation systems, including changes in documentation practices, training, clinical reasoning, and possible deskilling effects.

9. Conclusions

This study provides an early-stage evaluation of AI-assisted clinical documentation systems through the combination of controlled benchmarking and deployment-grounded operational observations. The benchmark results identified meaningful differences across ASR and LLM configurations, demonstrating important trade-offs between transcription accuracy, terminology preservation, report quality, hallucination rates, and processing costs. The operational traces provide complementary descriptive evidence that the system was incorporated into routine documentation workflows across multiple healthcare settings, particularly within documentation-intensive specialties.

These findings support continued supervised pilot evaluation of AI-assisted documentation technologies. However, they should not be interpreted as evidence of clinical safety, clinical effectiveness, documentation time savings, adoption rates, workflow impact, or readiness for autonomous deployment. The present study provides benchmark and operational evidence rather than clinical validation.

Accordingly, AI-assisted documentation systems should currently be understood as clinician-facing draft-generation tools that require human review, appropriate governance, workflow integration, and prospective clinical evaluation. Future research should extend beyond technical performance toward broader assessment of interoperability, clinician burden, workflow outcomes, patient experience, economic value, and health-system impact.

Author Contributions: Conceptualization, J.D., J.C., M.S., and J.S.-C.; methodology, J.D., J.C., and J.S.-C.; software, J.D.; validation, J.D., J.C., and J.S.-C.; formal analysis, J.D.; investigation, J.D. and J.C.; resources, J.D. and J.C.; data curation, J.D.; writing—original draft preparation, J.D.; writing—review and editing, J.D., J.C., M.S., and J.S.-C.; visualization, J.D.; supervision, J.C. and J.S.-C.; project administration, J.C.; funding acquisition, J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the project “Optimización del tiempo de espera y visita de pacientes en clínicas privadas mediante compartición de datos” (Expedient No. TSI-100123-2024-31), funded by the Spanish Ministry for Digital Transformation and Public Function, within the framework of the Sectoral Data Spaces Programme (Second Call, 2024). The project is part of the Recovery, Transformation and Resilience Plan and is financed by the European Union—NextGenerationEU. J.C. work was supported by the Torres Quevedo grant “Asistente Virtual para la mejora del sistema de salud—Virtual Assistant for Better HealthCare—VA4BHC” (ref. PTQ2021-012147).

Institutional Review Board Statement: The real-world analysis used retrospective operational trace data generated during routine system use. No intervention was introduced for research purposes, and the analytical dataset did not include directly identifiable patient data. The analysis was therefore handled as a secondary observational study under local internal governance procedures for operational data review.

Informed Consent Statement: Not applicable.

Data Availability Statement: The experimental dataset, including synthetic dialogue templates, audio files, transcriptions, generated reports, and evaluation metrics, is available from the corresponding author upon reasonable request for research purposes. Because the benchmark data are synthetic, sharing them does not involve patient-identifiable information. Real-world digital trace data used for the clinical log-mining analysis were collected from routine deployments in LATAM and Europe and are not publicly available due to privacy, contractual, and institutional governance restrictions; aggregated results supporting the findings are reported in Tables 7 and 8.

Acknowledgments: The authors would like to thank the anonymous reviewers for their detailed and helpful comments on the manuscript.

Conflicts of Interest: Authors were employed by Top Health Tech, the company that developed and deployed the documentation system evaluated in this study. The study therefore concerns the authors’ own product. The authors were involved in study design, software development, data analysis, and manuscript preparation. To mitigate potential bias, the study compared multiple ASR and LLM models rather than a single system configuration, used standardized quantitative metrics, and applied the same benchmarking framework uniformly across all evaluated combinations. These relationships should be considered when interpreting the findings.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
API	Application Programming Interface
ASR	Automatic Speech Recognition
BERT	Bidirectional Encoder Representations from Transformers
EHR	Electronic Health Record
FHIR	Fast Healthcare Interoperability Resources
GCP	Google Cloud Platform
HL7	Health Level Seven
IRB	Institutional Review Board
LATAM	Latin America
LLM	Large Language Model
M-WER	Medical Word Error Rate
ML	Machine Learning
ModernBERT	Modern Bidirectional Encoder Representations from Transformers model family
NLP	Natural Language Processing
ROUGE-L	Recall-Oriented Understudy for Gisting Evaluation–Longest Common Subsequence
SNOMED CT	Systematized Nomenclature of Medicine–Clinical Terms
SOAP	Subjective, Objective, Assessment, and Plan

TTS

Text-to-Speech

WER

Word Error Rate

Appendix A. Experimental Methodology and Implementation Details

This Appendix provides implementation details, prompts, and experimental configurations used in our evaluation of ASR–LLM combinations for medical report generation.

Appendix A.1. Dataset Composition and Medical Specialties

Our evaluation dataset comprises 230 experiments spanning 10 medical specialties across English and Spanish languages. Table A1 presents the specific medical conditions and clinical contexts used for each specialty.

Table A1. Medical specialties with associated medical conditions and clinical contexts. The dataset covers multiple medical domains with realistic clinical scenarios and synthetic doctor–patient interactions for each specialty.

Medical Specialty	Medical Condition	Clinical Context
Cardiology	Chest pain and shortness of breath	Heart health evaluation
Dentistry	Tooth sensitivity and bleeding gums	Regular dental checkup and cleaning
Dermatology	Unusual mole and skin rash	Skin cancer screening
ENT (Otolaryngology)	Chronic sinusitis and hearing loss	ENT evaluation
General Practice	Persistent cough and fever	Follow-up visit for respiratory symptoms
Neurology	Recurring migraines and dizziness	Neurological consultation
Ophthalmology	Gradual vision loss and eye strain	Annual eye examination
Orthopedics	Chronic knee pain after sports injury	Post-injury assessment
Pediatrics	Child with recurring ear infections	Pediatric checkup
Psychiatry	Anxiety and sleep disorders	Mental health consultation

Appendix A.2. Example Structure of a Generated Clinical Dialogue

To make the synthetic dataset more transparent, this subsection will provide an illustrative example of the generated doctor–patient dialogue structure used in the benchmark. The final example will include the ordered speaker turns, language, specialty context, and clinically relevant information that the ASR–LLM pipeline is expected to preserve during transcription and report generation:

Scenario:

Specialty: Cardiology

Condition: Chest pain and shortness of breath

Context: Heart health evaluation

Language: English

Dialogue:

[neutral] DOCTOR: Good morning, Mr. Davies. I’m Dr. Anya Sharma, your cardiologist. Please, have a seat. What brings you in today?

[concerned] PATIENT: Good morning, Doctor. I have been experiencing chest pain and shortness of breath for the past few weeks, and it’s getting worse.

[professional] DOCTOR: I understand. Can you describe the chest pain? Where exactly do you feel it, and what does it feel like? Is it sharp, dull, or a pressure?

[anxious] PATIENT: It's mostly in the center of my chest, like a heavy pressure. Sometimes it radiates to my left arm. It feels like someone is sitting on my chest. The shortness of breath comes on with even mild exertion, like walking up the stairs.

[professional] DOCTOR: Okay. Furthermore, how long does the chest pain typically last? Are there any specific activities that trigger it?

[anxious] PATIENT: It usually lasts for about 10-15 min. It seems to come on when I'm stressed or after I have eaten a big meal. Sometimes it happens just randomly, too.

[professional] DOCTOR: Do you have any other symptoms, such as palpitations, dizziness, sweating, or nausea?

[anxious] PATIENT: Yes, sometimes I feel a bit dizzy, and I have had some episodes of sweating, especially at night.

[professional] DOCTOR: Alright. What is your medical history? Do you have any pre-existing conditions, such as hypertension, diabetes, or high cholesterol? Furthermore, what medications are you currently taking?

[anxious] PATIENT: I was diagnosed with hypertension about five years ago, and I'm taking Lisinopril for it. My cholesterol is also a bit high, but I'm not on any medication for that yet. My father had a heart attack when he was in their early 60s.

[professional] DOCTOR: Thank you for that information. Now, I'm going to perform a physical examination. I will listen to your heart and lungs, check your blood pressure, and feel your pulse.

[neutral] DOCTOR: (Performing the examination) Okay, your blood pressure is a bit elevated, 150/90. Your heart sounds are regular, but I do hear a slight murmur. Your lungs are clear. There is no edema in your lower extremities. Your pulse is 88 bpm and regular.

[professional] DOCTOR: Based on your symptoms, medical history, and the physical examination, I suspect you may have angina, possibly related to coronary artery disease. However, we need further testing to confirm this diagnosis. I'd like to order an electrocardiogram (ECG), a stress test, and some blood work, including a lipid panel and cardiac enzyme levels.

[anxious] PATIENT: A stress test? What does that involve?

[reassuring] DOCTOR: The stress test will involve walking on a treadmill while we monitor your heart's electrical activity. This will help us see how your heart responds to exertion and identify any potential blockages in your arteries. The ECG is a simple test to record the electrical activity of your heart at rest.

[professional] DOCTOR: After we get the results of these tests, we'll discuss

the findings and potential treatment options. These may include lifestyle modifications like diet and exercise, medication to manage your blood pressure and cholesterol, or, in more severe cases, procedures such as angioplasty or bypass surgery.

[reassuring] DOCTOR: For now, I'd recommend that you avoid strenuous activity and try to manage your stress levels. Please follow a heart-healthy diet, low in saturated fat, cholesterol, and sodium. Furthermore, it's important to quit smoking if you are a smoker.

[professional] DOCTOR: I will have my nurse schedule you for the ECG, stress test, and blood work. We should have the results back within a week. Then, we can schedule a follow-up appointment to discuss them.

[anxious] PATIENT: Okay, thank you, Doctor. I appreciate your help.

[reassuring] DOCTOR: You're welcome, Mr. Davies. Try not to worry too much. We'll get to the bottom of this and develop a plan to manage your symptoms and protect your heart health. Please call the office if your symptoms worsen or if you have any questions before your follow-up appointment.

[neutral] DOCTOR: Take care, and I will see you soon.

[neutral] PATIENT: Thank you, Doctor. Goodbye.

Appendix A.3. Language Model Prompts for Medical Report Generation

The main prompt design is summarized in Section 4. For reproducibility, the report-generation prompts in both English and Spanish followed the same core constraints:

- Generate a concise clinician-style report in the source language;
- Preserve the provided section headings and maintain SOAP-consistent organization;
- Include diagnoses, medications, and management details only when explicitly supported by the transcript;
- Avoid inventing information, recommendations, or clinical findings not present in the dialogue; and
- In Spanish, include all requested sections and explicitly mark missing content when needed.

The automated evaluation prompt likewise used a fixed structured schema and returned JSON values for factual accuracy, information coverage, clinical relevance, hallucination score, and terminology accuracy. This Appendix therefore retains only the core prompt principles, while the main text carries the operational summary needed for manuscript readability.

Appendix A.4. Automated Evaluation Methodology with Gemini-2.0-Flash

Appendix A.4.1. Evaluation Model Configuration

Clinical quality assessment was performed using Gemini-2.0-flash with the following configuration:

- **Model:** Gemini-2.0-flash.
- **Temperature:** 0.1 (low temperature for consistent evaluation).
- **Response format:** Structured JSON output.
- **Response schema:** Pydantic BaseModel with enforced numerical outputs.

Appendix A.4.2. Clinical Quality Evaluation Prompt

The following prompt was used for automated clinical quality assessment:

Analyze and compare the original doctor-patient dialogue with the generated medical report. Provide numerical scores between 0.0 and 1.0 for each metric:

Original Dialogue:
{reference_text}

Generated Medical Report:
{medical_report}

Evaluate for:

1. **Factual Accuracy** - The extent to which all statements in the generated report accurately reflect the information in the original dialogue.
2. **Information Coverage** - The degree to which all key medical information points from the original dialogue are captured in the generated report.
3. **Clinical Relevance** - The degree to which the generated report focuses on medically important details and avoids irrelevant information.
4. **Hallucination Score** - A score representing the absence of added or fabricated information in the generated report. A score of 1.0 indicates no hallucinations.
5. **Terminology Accuracy** - The correctness and appropriateness of the medical terms used in the generated report.

Return as JSON with the score for each metric.

Appendix A.4.3. Structured Output Schema

The evaluation employed a structured output schema to ensure consistent numerical scoring:

```
class StructuredOutputSchema(BaseModel):
    factual_accuracy: float = Field(...,
        description = "Score between 0.0 and 1.0 indicating factual accuracy.")
    information_coverage: float = Field(...,
        description = "Score between 0.0 and 1.0 indicating information coverage.")
    clinical_relevance: float = Field(...,
        description = "Score between 0.0 and 1.0 indicating clinical relevance.")
    hallucination_score: float = Field(...,
        description = "Score between 0.0 and 1.0 indicating the absence of
        hallucinations (1.0 = no hallucinations).")
    terminology_accuracy: float = Field(...,
        description = "Score between 0.0 and 1.0 indicating terminology accuracy.")
```

Appendix A.5. Economic Analysis and Pricing Structure

Appendix A.5.1. Cost Components

Our economic analysis encompasses three primary cost components:

1. **Transcription Prices:** Provider price schedules applied per minute of audio processed;
2. **Processing Prices:** LLM inference prices applied per token for input and output;
3. **Infrastructure Costs:** For locally deployed models (MedGemma), amortized deployment costs.

Appendix A.5.2. Cost Statistics

Table A2 presents the benchmark cost analysis across all ASR–LLM combinations.

Table A2. Cost distribution statistics across all ASR–LLM combinations. The wide range reflects substantial cost variation between model combinations and is intended to inform comparative implementation planning under the study assumptions.

Cost Metric	Value (USD)
Mean cost per consultation	\$0.0620
Median cost per consultation	\$0.0297
Minimum cost per consultation	\$0.0095
Maximum cost per consultation	\$0.3830
Standard deviation	\$0.0891

Appendix A.6. Statistical Analysis Methodology

Appendix A.6.1. Statistical Tests Employed

Primary ASR inference used exact two-sided Wilcoxon signed-rank tests for the planned paired comparisons performed on identical audio clips. The three planned ASR contrasts reported in the main text were Voxtral versus Whisper, Voxtral versus GPT-4o Transcribe, and Voxtral versus GCP. Family-wise error for these planned contrasts was controlled using the Bonferroni procedure ($m = 3$, $\alpha_{adj} = 0.05/3 = 0.0167$). For Wilcoxon analyses, effect size was reported as $r = |Z|/\sqrt{n_{pairs}}$, where values of approximately 0.1, 0.3, and 0.5 are commonly interpreted as small, medium, and large, respectively.

To make the statistical reporting explicit, Table A3 reports the exact two-sided p -values, Bonferroni-adjusted p -values, and Wilcoxon effect sizes for the planned ASR pairwise contrasts. All three planned contrasts remained statistically significant after multiplicity correction, and the associated effect sizes were large ($r \approx 0.88$ for each contrast).

Table A3. Exact Wilcoxon signed-rank results for the planned ASR pairwise contrasts reported in the main text. Bonferroni-adjusted p -values use $m = 3$ planned comparisons.

Contrast	n_{pairs}	W	Exact p	Bonferroni p	r
Voxtral vs. Whisper	20	0.0	1.91×10^{-6}	5.72×10^{-6}	0.877
Voxtral vs. GPT-4o Transcribe	20	0.0	1.91×10^{-6}	5.72×10^{-6}	0.877
Voxtral vs. GCP	10	0.0	0.001953	0.005859	0.886

Power calculations were documented for the planned ASR contrasts using a conservative paired-sample large-effect approximation with the same Bonferroni-adjusted significance level ($\alpha_{adj} = 0.0167$). For the two primary $n = 20$ contrasts, the design achieved power of approximately 0.995 for a large within-pair effect ($d_z = 1.2$). For the smaller $n = 10$ GCP comparison, power was approximately 0.783 at $d_z = 1.2$ and exceeded 0.95 only for very large effects ($d_z \gtrsim 1.5$). Accordingly, the $n = 20$ ASR contrasts were highly powered for large effects, whereas the English-only GCP comparison should be interpreted somewhat more cautiously despite remaining significant after Bonferroni correction.

Appendix A.6.2. Illustrative Pilot-Stage Reference Thresholds

The following thresholds are included only as pragmatic reference points for interpreting benchmark differences during early pilot planning. They were defined by the authors for internal interpretation and are not statistically derived from the present study, nor should they be treated as universal deployment criteria.

- **WER Improvement:** 0.05 absolute improvement used as a practical reference point.

- **Medical WER:** 0.03 absolute improvement used as a practical reference point for medical terminology.
- **Clinical Quality Metrics:** 0.02 improvement on a 0–1 scale used as a practical reference point.
- **Cost-Effectiveness:** 10% improvement in the study-specific efficiency ratio used as a practical reference point.

Appendix A.7. Reproducibility and Data Availability

Appendix A.7.1. Experimental Reproducibility

All experiments were conducted with fixed random seeds and deterministic configurations where possible. Temperature settings were maintained at 0.1 for LLM evaluations to ensure reproducible results across multiple runs.

Appendix A.7.2. Code and Configuration Availability

Complete experimental code, configuration files, and analysis scripts are maintained in version control with documented dependencies and environment specifications to enable full experimental reproduction.

Appendix A.8. Clinical Safety Metrics: Hallucination Detection and Terminology Accuracy

The main text reports the safety-oriented summary metrics in Section 5.3. This Appendix retains only the supplementary multiplicity-adjusted pairwise comparisons that support those main-text summaries.

Supplementary Pairwise Comparisons

Statistical comparisons using Mann–Whitney U tests reveal significant differences in hallucination detection performance between specific model pairs, while terminology accuracy differences are less pronounced among the top-performing models. For each safety metric, three pairwise comparisons were performed and Bonferroni-adjusted p -values are therefore reported alongside the exact unadjusted values ($m = 3$, $\alpha_{adj} = 0.0167$), as summarized in Table A4.

Table A4. Exact and Bonferroni-adjusted pairwise comparisons for the safety-oriented evaluation metrics. Bonferroni adjustment was applied separately within each three-comparison metric family.

Metric	Contrast	Exact p	Bonferroni p	α_{adj}	Result
Hallucination score	GPT-4o vs. MedGemma	0.0039	0.0117	0.0167	Significant
Hallucination score	Gemini 1.5 Pro vs. GPT-4o	0.0543	0.1629	0.0167	Not significant
Hallucination score	Gemini 1.5 Pro vs. MedGemma	0.3501	1.0000	0.0167	Not significant
Terminology accuracy	GPT-4o vs. MedGemma	0.0741	0.2223	0.0167	Not significant
Terminology accuracy	Gemini 1.5 Pro vs. GPT-4o	0.2580	0.7740	0.0167	Not significant
Terminology accuracy	Gemini 1.5 Pro vs. MedGemma	0.4347	1.0000	0.0167	Not significant

These supplementary pairwise results are consistent with the main-text interpretation that the strongest separation in this benchmark is between the modern LLM pipelines and the degraded-input negative control condition, rather than between GPT-4o and Gemini 1.5 Pro.

References

- Wenger, N.; Doyle, B.J. For Whom the Note Scrolls: A Brief History of the Medical Record's Transition from a Tool for Physicians to a Bill. *Ann. Intern. Med.* **2024**, *177*, 566–569. [[CrossRef](#)] [[PubMed](#)]
- Asch, D.A.; Asch, R.M.; Asch, J.; Asch, S.M.; Asch, S.M. Clinical Documentation in the 21st Century: Executive Summary of a Policy Position Paper from the American College of Physicians. *Ann. Intern. Med.* **2015**, *162*, 797–798. [[CrossRef](#)] [[PubMed](#)]
- Strong, P.; Attal-Singer, J.; Singer, D.R.; Marcus, A.; Carew, A.; Althoff, A.; Singh, H. Enhancing clinical documentation with ambient artificial intelligence: A quality improvement survey assessing clinician perspectives on work burden, burnout, and job satisfaction. *JAMIA Open* **2025**, *8*, ooaf013. [[CrossRef](#)] [[PubMed](#)]
- Saadat, S.; Khalilizad, M.; Qorbani, M.; Hemmat, A.; Hariri, S. Enhancing Clinical Documentation with AI: Reducing Errors, Improving Interoperability, and Supporting Real-Time Note-Taking. *InfoSci. Trends* **2025**, *2*, 1–13. [[CrossRef](#)]
- Lee, C.; Britto, S.; Diwan, K. Evaluating the Impact of Artificial Intelligence (AI) on Clinical Documentation Efficiency and Accuracy Across Clinical Settings: A Scoping Review. *Cureus* **2024**, *16*, e73994. [[CrossRef](#)] [[PubMed](#)]
- Leong, H.; Gao, Y.; Ji, S.; Zhang, Y.; Pamuksuz, U. Efficient Fine-Tuning of Large Language Models for Automated Medical Documentation. In Proceedings of the 2024 4th International Conference on Digital Society and Intelligent Systems (DSInS), Sydney, Australia, 20–22 November 2024.
- van Buchem, M.M.; Kant, I.M.J.; King, L.; Kazmaier, J.; Steyerberg, E.W.; Bauer, M.P. Impact of a Digital Scribe System on Clinical Documentation Time and Quality: Usability Study. *JMIR AI* **2024**, *3*, e60020. [[CrossRef](#)] [[PubMed](#)]
- Alsentzer, E.; Bergman, A.; McDermott, M.B.A.; Henry, P.; Ghassemi, M. Enhancing Clinical Documentation with Synthetic Data: Leveraging Generative Models for Improved Accuracy. *arXiv* **2024**, arXiv:2406.06569.
- Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv* **2023**, arXiv:2212.04356.
- Adedeji, A.; Sanni, M.; Ayodele, E.; Joshi, S.; Olatunji, T. The Multicultural Medical Assistant: Can LLMs Improve Medical ASR Errors Across Borders? *arXiv* **2025**, arXiv:2501.15310.
- Yu, S.; Lee, S.S.; Hwang, H. The ethics of using artificial intelligence in medical research. *Kosin Med. J.* **2024**, *39*, 229–237. [[CrossRef](#)]
- Ahuja, A.S.; Chen, E.P.; Hong, R.D.; Langlotz, C.P.; Lungren, M.P. Human-in-the-loop AI for clinical decision support: A review. *J. Am. Med. Informatics Assoc.* **2023**, *30*, 597–607. [[CrossRef](#)] [[PubMed](#)]
- Gottlieb, M.; Lin, S.; Strong, P.; Lee, J.; Singer, D.R.; Marcus, A.; Singh, H. Assessing Patient-Reported Satisfaction with Care and Documentation Time in Primary Care Through AI-Driven Automatic Clinical Note Generation: Protocol for a Proof-of-Concept Study. *JMIR Res. Protoc.* **2025**, *14*, e66232. [[CrossRef](#)] [[PubMed](#)]
- Strong, P.; Lin, S.; Gottlieb, M.; Singh, H. Benchmarking and Datasets for Ambient Clinical Documentation: A Scoping Review of Existing Frameworks and Metrics for AI-Assisted Medical Note Generation. *medRxiv* **2025**. [[CrossRef](#)]
- Yim, W.w.; Fu, Y.; Ben Abacha, A.; Snider, N.; Lin, T.; Yetisgen, M. Aci-bench: A novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Sci. Data* **2023**, *10*, 558. [[CrossRef](#)] [[PubMed](#)]
- Ben Abacha, A.; Yim, W.w.; Michalopoulos, G.; Lin, T. An Investigation of Evaluation Methods in Automatic Medical Note Generation. In *Proceedings of the Findings of the Association for Computational Linguistics*; ACL: Toronto, ON, Canada, 2023; pp. 2575–2588. [[CrossRef](#)]
- Wang, H.; Yang, R.; Alwakeel, M.; Kayastha, A.; Chowdhury, A.; Biro, J.M.; Sorrentino, A.D.; Handley, J.L.; Hantzmon, S.; Bessias, S.; et al. An evaluation framework for ambient digital scribing tools in clinical applications. *npj Digit. Med.* **2025**, *8*, 358. [[CrossRef](#)] [[PubMed](#)]
- Sasseville, M.; Yousefi, F.; Ouellet, S.; Naye, F.; Stefan, T.; Carnovale, V.; Bergeron, F.; Ling, L.; Gheorghiu, B.; Hagens, S.; et al. The Impact of AI Scribes on Streamlining Clinical Documentation: A Systematic Review. *Healthcare* **2025**, *13*, 1447. [[CrossRef](#)] [[PubMed](#)]
- Bracken, A.; Reilly, C.; Feeley, A.; Sheehan, E.; Merghani, K.; Feeley, I. Artificial Intelligence (AI)-Powered Documentation Systems in Healthcare: A Systematic Review. *J. Med. Syst.* **2025**, *49*, 28. [[CrossRef](#)] [[PubMed](#)]
- Ng, J.J.W.; Wang, E.; Zhou, X.; Zhou, K.X.; Goh, C.X.L.; Sim, G.Z.N.; Tan, H.K.; Goh, S.S.N.; Ng, Q.X. Evaluating the Performance of Artificial Intelligence-Based Speech Recognition for Clinical Documentation: A Systematic Review. *BMC Med. Inform. Decis. Mak.* **2025**, *25*, 236. [[CrossRef](#)] [[PubMed](#)]
- Shon, S.; Kim, H.K.; Kim, H.K.; Lee, H. Voice Activity Detection for Enhancing the Performance of Automatic Speech Recognition Systems. *Appl. Sci.* **2021**, *11*, 8146. [[CrossRef](#)]
- Ferrer, L.; Watanabe, S.; Gaur, Y.; Wiesner, M.; Delcroix, M.; Ogawa, A.; Nakatani, T.; Khudanpur, S. Studying the Effect of Silence on ASR Performance and How Its Usage Can Be Optimized in the ASR System. In Proceedings of the 23rd Annual Conference of the International Speech Communication Association (INTERSPEECH 2022), Incheon, Republic of Korea, 18–22 September 2022; pp. 1021–1025. [[CrossRef](#)]
- Lee, J.R.; Lungren, M.P.; Lee, P.H. Generating SOAP Notes from Doctor-Patient Conversations Using Large Language Models. *arXiv* **2023**, arXiv:2303.06488.

24. Agrawal, M.; Adams, M.; Brophy, E.; Chien, J.; Du, M.; Elibol, O.; Lin, C.C.; Logghe, P.; Maki, B.; Mehta, S.; et al. Medprompt: Cross-Modal Few-Shot Prompting for Ophthalmic Report Generation. In Proceedings of the Conference on Health, Inference, and Learning (CHIL), New York, NY, USA, 27–28 June 2024; pp. 1–25.
25. Dash, S.; Shakyawar, S.K.; Sharma, M.; Kaushik, S. Big data in healthcare: Management, analysis and future prospects. *J. Big Data* **2019**, *6*, 54. [\[CrossRef\]](#)
26. PHTI. *Adoption of AI in Healthcare Delivery Systems: Early Applications & Impacts*; Technical Report; Peterson Health Technology Institute: New York, NY, USA, 2025.
27. Delaunay, J.; Girbes, D.; Cusido, J. Evaluating the Effectiveness of Large Language Models in Converting Clinical Data to FHIR Format. *Appl. Sci.* **2025**, *15*, 3379. [\[CrossRef\]](#)
28. Nawab, K. Artificial intelligence scribe: A new era in medical documentation. *Artif. Intell. Health* **2024**, *1*, 12–15. [\[CrossRef\]](#)
29. Measure and Improve Speech Accuracy. Google Cloud Speech-to-Text Documentation. Available online: <https://cloud.google.com/speech-to-text/docs/speech-accuracy> (accessed on 23 April 2025).
30. Quiroz, J.C.; Laranjo, L.; Kocaballi, A.B.; Berkovsky, S.; Rezazadegan, D.; Coiera, E. Challenges of developing a digital scribe to reduce clinical documentation burden. *npj Digit. Med.* **2019**, *2*, 114. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Banerjee, S.; Agarwal, A.; Ghosh, P. High-precision medical speech recognition through synthetic data and semantic correction: UNITED-MEDASR. *arXiv* **2024**, arXiv:2412.00055.
32. Zhou, L.; Blackley, S.V.; Kowalski, L.; Doan, R.; Acker, W.W.; Landman, A.B.; Kontrient, E.; Mack, D.; Meteer, M.; Bates, D.W.; et al. Analysis of Errors in Dictated Clinical Documents Assisted by Speech Recognition Software and Professional Transcriptionists. *JAMA Netw. Open* **2018**, *1*, e183458. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Liu, W.; Hu, J.; Lv, F.; Tang, Z. A New Method for Long-Term Temperature Compensation of Structural Health Monitoring by Ultrasonic Guided Wave. *Measurement* **2025**, *252*, 117310. [\[CrossRef\]](#)
34. Li, H.; Liu, W. Bone Damage Detection Using Ultrasonic Guided Waves: Multi-Feature Integration and Machine Learning Approaches. *Nondestruct. Test. Eval.* **2026**, 1–34. [\[CrossRef\]](#)
35. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774. [\[CrossRef\]](#)
36. Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A.; Millican, K.; et al. Gemini: A Family of Highly Capable Multimodal Models. *arXiv* **2023**, arXiv:2312.11805. [\[CrossRef\]](#)
37. Saab, K.; Tu, T.; Weng, W.H.; Tanno, R.; Stutz, D.; Wulczyn, E.; Zhang, F.; Strother, T.; Park, C.; Vedadi, E.; et al. Capabilities of Gemini Models in Medicine. *arXiv* **2024**, arXiv:2404.18416.
38. Alsentzer, E.; Murphy, J.R.; Boag, W.; Weng, W.H.; Jindi, D.; Naumann, T.; McDermott, M. Publicly Available Clinical BERT Embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, MN, USA, 7 June 2019; pp. 72–78. [\[CrossRef\]](#)
39. Adediji, A.; Joshi, S.; Doohan, B. The Sound of Healthcare: Improving Medical Transcription ASR Accuracy with Large Language Models. *arXiv* **2024**, arXiv:2402.07658.
40. Sharma, A.; Sharma, R.; Sharma, M. Transforming Health Care with Artificial Intelligence: Redefining Medical Documentation. *Cureus* **2025**, *17*, e56965. [\[CrossRef\]](#) [\[PubMed\]](#)
41. Yadav, S. Embracing artificial intelligence: Revolutionizing nursing documentation for a better future. *Cureus* **2024**, *16*, e57725. [\[CrossRef\]](#) [\[PubMed\]](#)
42. Sharma, A.; Sharma, R.; Sharma, M. Inspired Spine Smart Universal Resource Identifier (SURI): An Adaptive AI Framework for Transforming Multilingual Speech Into Structured Medical Reports. *Cureus* **2025**, *17*, e64606. [\[CrossRef\]](#) [\[PubMed\]](#)
43. Hodgson, T.; Magrabi, F.; Coiera, E. Automatic speech recognition performance for digital scribes: A performance comparison between general-purpose and specialized models tuned for patient-clinician conversations. *JAMIA Open* **2023**, *6*, ooad020. [\[CrossRef\]](#) [\[PubMed\]](#)
44. Draper, T.C.; Leake, J.; Lamb-Riddell, K.; Cox, T.; McCormick, J.; Trowell, S.; Kiely, J.; Luxton, R. The Impact of Acoustic and Informational Noise on AI-Generated Clinical Summaries. *medRxiv* **2025**. [\[CrossRef\]](#)
45. Federation of State Medical Boards. *Navigating the Responsible and Ethical Incorporation of Artificial Intelligence into Clinical Practice*; Policy Document; FSMB: Euless, TX, USA, 2024.
46. Mianroodi, A.R.; Rezaie, A.; Todorov, N.G.; Rakovski, C.; Rudzicz, F. MedSynth: Realistic, Synthetic Medical Dialogue-Note Pairs. *arXiv* **2025**, arXiv:2508.01401.
47. Qiu, P.; Wu, C.; Zhang, X.; Lin, W.; Wang, H.; Zhang, Y.; Wang, Y.; Xie, W. Towards building multilingual language model for medicine. *Nat. Commun.* **2024**, *15*, 8384. [\[CrossRef\]](#) [\[PubMed\]](#)
48. Rotenstein, L.S.; Holmgren, A.J.; Thombley, R.; Sriram, A.; Dbouk, R.H.; Jost, M.; Aizenberg, D.; MacDonald, S.; Kanaparthi, N.; Williams, B.; et al. Changes in Clinician Time Expenditure and Visit Quantity with Adoption of Artificial Intelligence–Powered Scribes: A Multisite Study. *JAMA* **2026**, *335*, 1408–1417. [\[CrossRef\]](#) [\[PubMed\]](#)
49. Lee, K.C.K.; Clifton, D.A.; Chen, T.J.H. Generating synthetic medical data with large language models: Promises and pitfalls. *Lancet Digit. Health* **2023**, *5*, e753–e754.

50. Chen, H.; Shu, K.; Wang, S.; Wang, S.; Chen, T.; Skiena, S. Generating Synthetic Text Data for Natural Language Processing. *arXiv* **2021**, arXiv:2103.02762.
51. Liu, A.; Ehrenberg, A.; Lo, A.; Denoix, C.; Barreau, C.; Lample, G.; Delignon, J.M.; Chandu, K.; Platen, P.; Muddireddy, P.; et al. Voxtral. *arXiv* **2025**, arXiv:2507.13264. [[CrossRef](#)]
52. OpenAI. GPT-4o System Card. 2024. Available online: <https://openai.com/index/gpt-4o-system-card/> (accessed on 7 April 2026).
53. Sellergren, A.; Kazemzadeh, S.; Jaroensri, T.; Kiraly, A.; Traverse, M.; Kohlberger, T.; Xu, S.; Jamil, F.; Hughes, C.; Lau, C.; et al. MedGemma Technical Report. *arXiv* **2025**, arXiv:2507.05201.
54. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.P.; et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv* **2023**, arXiv:2306.05685. [[CrossRef](#)]
55. Panickssery, A.; Bowman, S.R.; Feng, S. LLM Evaluators Recognize and Favor Their Own Generations. *arXiv* **2024**, arXiv:2404.13076. [[CrossRef](#)]
56. Chen, Z.Y.; Wang, H.; Zhang, X.; Hu, E.; Lin, Y. Beyond the Surface: Measuring Self-Preference in LLM Judgments. *arXiv* **2025**, arXiv:2506.02592.
57. Warner, B.; Chaffin, A.; Clavié, B.; Weller, O.; Hallström, O.; Taghadouini, S.; Gallagher, A.; Biswas, R.; Ladhak, F.; Aarsen, T.; et al. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, Vienna, Austria, 27 July–1 August 2025; Volume 1, pp. 2526–2547. [[CrossRef](#)]
58. Hada, R.; Gumma, V.; Wynter, A.; Diddee, H.; Ahmed, M.; Choudhury, M.; Bali, K.; Sitaram, S. Are Large Language Model-based Evaluators the Solution to Scaling Up Multilingual Evaluation? In Proceedings of the Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, 17–22 March 2024; pp. 1051–1070.
59. Gupta, G.K.; Singh, A.; Manikandan, S.V.; Ehtesham, A. Digital Diagnostics: The Potential of Large Language Models in Recognizing Symptoms of Common Illnesses. *arXiv* **2024**, arXiv:2405.06712.
60. Delaunay, J.; Cusido, J. Evaluating the Performance of Large Language Models in Predicting Diagnostics for Spanish Clinical Cases in Cardiology. *Appl. Sci.* **2025**, *15*, 61. [[CrossRef](#)]
61. Rosner, B.; Glynn, R.J.; Lee, M.L.T. The Wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics* **2006**, *62*, 185–192. [[CrossRef](#)] [[PubMed](#)]
62. Al Meslamani, A.Z. Beyond implementation: The long-term economic impact of AI in healthcare. *J. Med. Econ.* **2023**, *26*, 1566–1569. [[CrossRef](#)] [[PubMed](#)]
63. Lin, S.; Strong, P.; Singh, H.; Gottlieb, M. *Clinicians' Experiences with EHR Documentation and Attitudes Toward AI-Assisted Documentation*; Stanford Medicine White Paper; Stanford Medicine: Stanford, CA, USA; Google Health: Mountain View, CA, USA, 2024.
64. Li, Y.; Wang, H.; Yerebakan, H.Z.; Shinagawa, Y.; Luo, Y. FHIR-GPT Enhances Health Interoperability with Large Language Models. *NEJM AI* **2024**, *1*, A1cs2300301. [[CrossRef](#)] [[PubMed](#)]
65. Schmiedmayer, P.; Rao, A.; Zagar, P.; Ravi, V.; Zahedivash, A.; Fereydooni, A.; Aalami, O. LLM on FHIR—Demystifying Health Records. *arXiv* **2024**, arXiv:2402.01711.
66. Hong, N.; Wen, A.; Shen, F.; Sohn, S.; Liu, S.; Liu, H.; Jiang, G. Integrating Structured and Unstructured EHR Data Using an FHIR-based Type System: A Case Study with Medication Data. *AMIA Jt. Summits Transl. Sci. Proc.* **2018**, *2018*, 74–83.
67. Fagherazzi, G.; Fischer, A.; Ismael, M.; Despotovic, V. Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice. *Digit. Biomark.* **2021**, *5*, 78–88. [[CrossRef](#)]
68. De Silva, U.; Subramaniam, P.; Doan, T.N.; Babar, Z.U.D.; Nandasena, M.; Li, C.; Masek, M.; Lee, A. Clinical Decision Support Using Speech Signal Analysis: Systematic Scoping Review of Neurological Disorders. *J. Med. Internet Res.* **2025**, *27*, e63004. [[CrossRef](#)] [[PubMed](#)]
69. Vilenchik, D.; Cwikel, J.; Ezra, Y.; Hausdorff, T.; Lazarov, M.; Sergienko, R.; Abramovitz, R.; Schmidt, I.; Perez, A.S. Method Matters: Enhancing Voice-Based Depression Detection with a New Data Collection Framework. *Depress. Anxiety* **2025**, *2025*, 4839334. [[CrossRef](#)] [[PubMed](#)]
70. Montenegro, L.; Gomes, L.M.; Machado, J.M. What We Know About the Role of Large Language Models for Medical Synthetic Dataset Generation. *AI* **2025**, *6*, 109. [[CrossRef](#)]
71. Li, S.; Yerebakan, M.; Luo, Y.; Amaba, B.; Swope, W.; Hu, B. The Effect of Different Occupational Background Noises on Voice Recognition Accuracy. *J. Comput. Inf. Sci. Eng.* **2022**, *22*, 050905. [[CrossRef](#)]
72. Chiu, E.K.Y.; Chung, T.W.H. Protocol for Human Evaluation of Artificial Intelligence Chatbots in Clinical Consultations. *medRxiv* **2024**. [[CrossRef](#)]

73. Saadi, N.; Raha, T.; Christophe, C.; Pimentel, M.A.; Rajan, R.; Kanithi, P.K. Bridging Language Barriers in Healthcare: A Study on Arabic LLMs. *arXiv* **2025**, arXiv:2501.09825.
74. Wang, X.; Zhang, N.; He, H.; Nguyen, T.; Yu, K.H.; Deng, H.; Brandt, C.; Bitterman, D.; Pan, L.; Cheng, C.y.; et al. Safety challenges of AI in medicine in the era of large language models. *arXiv* **2024**, arXiv:2409.18968. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.