MedicaLLM: LLM-Driven Speech and Language Solutions for Healthcare

Ronghao Pan[2], **Pedro José Vivancos-Vicente**[1], **Juan Salvador Castejón-Garrido**[1], **Tomás Bernal-Beltrán**[2], **Rafael Valencia-García**[2]
[1] VÓCALI SISTEMAS INTELIGENTES S.L. Parque Científico de Murcia,
Carretera de Madrid km 388. Complejo de Espinardo, 30100 Murcia, Spain
[2] Departamento de Informática y Sistemas, Universidad de Murcia, Campus de Espinardo 30100 Murcia
ronghao.pan@um.es, pedro.vivancos@vocali.net,
juans.castejon@vocali.net, tomas.bernalb@um.es, valencia@um.es

Abstract

Although healthcare documentation is increasingly dependent on speech-based clinical interactions, general-purpose Automatic Speech Recognition (ASR) and Large Language Models (LLMs) lack the domain adaptation, structured control and interoperability guarantees required in regulated medical environments. These limitations often result in transcription errors, hallucinated content, and limited alignment with standardized coding systems. This paper introduces MedicaLLM, a multilingual, end-to-end framework integrating domain-adapted ASR, LLM-based structured report generation, and ontology-driven semantic enrichment within a modular architecture for clinical documentation. MedicaLLM combines medical interview transcription with structured report generation, summarization, and error correction; Named Entity Recognition (NER); and Medical Entity Linking (MEL) to align with standards such as SNOMED-CT and ICD-10. Deployed as a secure software as a service (SaaS) platform with REST API integration, MedicaLLM aims to reduce the administrative burden, improve the quality of documentation, and enhance semantic interoperability across healthcare systems, all while maintaining computational efficiency and clinical reliability.

Keywords: Automatic Speech Recognition, Large Language Models, Clinical Documentation, Medical Report Generation, Named Entity Recognition, Medical Entity Linking, Multilingual Healthcare AI

1. Introduction

Healthcare systems are currently undergoing an accelerated digital transformation, driven by the increasing demand for efficiency, accuracy, and interoperability in the clinical documentation processes (Holmgren et al., 2024). The growing administrative burden faced by healthcare professionals, together with the need for precise and standardized medical records, has highlighted the limitations of traditional documentation workflows. Automatic Speech Recognition (ASR) (Blackley et al., 2019; Ng et al., 2025) and Large Language Models (LLMs) (Shool et al., 2025; Thirunavukarasu et al., 2023), represents a strategic opportunity to modernize clinical practice while maintaining high standards of safety, compliance and linguistic precision.

Despite recent advances in Transformer-based architectures such as BERT (Devlin et al., 2019), GPT-based models (Brown et al., 2020), LLaMa-3 (Grattafiori et al., 2024), Qwen-3 (Yang et al., 2025), Whisper (Radford et al., 2023) and related variants, directly adopting these technologies in the healthcare domain remains challenging. Clinical language is characterized by highly specialized terminology, domain-specific abbreviations, multilingual variability (notably Spanish and Catalan), and strict regulatory requirements (Gu et al., 2021; Carrino et al., 2022). Generic ASR and LLM systems often struggle to accurately transcribe medical consultations, correctly interpret technical terminology, or generate structured medical reports aligned with standardized ontologies such as SNOMED-CT and ICD-10 (Hu et al., 2024). Furthermore, concerns about data privacy, computational costs, robustness in noisy clinical environment, and model hallucinations further complicate their deployment in real-world healthcare settings (Kim et al., 2025).

The MedicalLLM project addresses these challenges by developing an end-to-end Artificial Intelligence (AI) pipeline. This pipeline integrates domain-specific ASR systems with LLM-based modules that are specialized for generating structured reports, recognizing named entities, and linking medical entities. Specifically designed for a multilingual healthcare environment in Spanish and Catalan, the system incorporates fine-tuned Transformer-based models, medical-domain datasets, and normalization mechanisms aligned with international clinical standards. The project aims to automate and optimize key stages of clinical documentation by combining speech processing, semantic understanding, summarization, and ontology mapping within a unified architecture.

The proposed architecture is organized into modular components that separate data acquisition, transcription, semantic processing, normalization, and presentation layers. First, an end-to-end ASR module processes medical interviews and dictations, ensuring high transcription accuracy under

Proceedings of LANLP: Bridging Ibero and Latin American NLP Communities, pages 29–37

realistic hospital acoustic conditions. Next, LLM-based modules transform raw transcripts into structured medical reports through context-aware summarization and error correction mechanisms. Then, NER and MEL components extract relevant clinical entities, such as symptoms, diagnoses, medications, and procedures, and map them to standardized terminologies to ensure interoperability and regulatory compliance. The system is deployed via a scalable software as a service (SaaS) platform with secure backend processing, user-friendly interfaces, and API-based integration with electronic health record (EHR) systems.

Beyond technological innovation, the project places strong emphasis on reliability, privacy and responsible AI deployment. Data governance mechanisms compliant with General Data Protection Regulation (GDPR) and healthcare regulations, hallucination migation strategies, structured output validation, and human-in-the-loop validation workflows are integrated into the system design. Furthermore, optimization techniques such as quantization and parameter-efficient fine-tuning are explored to enable deployment in resource-constrained healthcare environments.

The project (CPP2024-011574) is funded by the Spanish National Research Agency (AEI) through the Colaboración público-privada call. The consortium members are VOCALI SISTEMAS INTELIGENTES S.L., a company with extensive experience in medical ASR solutions, and the TECNOMOD research group at the Universidad de Murcia, specialized in NLP, LLMs and semantic technologies.

Currently, the platform is being developed and validated through structured work packages covering multilingual resource creation, ASR adaptation, LLM fine-tuning, ontology alignment, SaaS development, and integrated system validation. The expected outcome is a Technology Readiness Level (TRL) 6 prototype capable of operating in real clinical environments, reducing administrative workload, improving documentation quality, and enhancing the interoperability of medical records across healthcare systems.

In summary, MedicaLLM proposes a comprehensive, multilingual and domain-adapted AI framework that bridges the gap between state-of-the-art speech and language technologies and the practical demands of modern healthcare systems, contributing to more efficient, standardized and scalable clinical workflows.

2. Background Information

In recent years, advances in AI, particularly in Natural Language Processing (NLP) and ASR, have significantly transformed the way textual and auditory data are processed. Transformer-based architectures such as BERT (Devlin et al., 2019), LLaMa-3 (Grattafiori et al., 2024), GPT3-family models (Brown et al., 2020), Qwen-3 (Yang et al., 2025), Whisper (Radford et al., 2023), HuBERT (Hsu et al., 2021) and Wav2Vec 2.0 (Baevski et al., 2020) have demonstrated remarkable capabilities in contextual language modeling, speech-to-text transcription, and generative reasoning. These technologies have reached near-human performance in general-domain tasks and have been widely adopted in applications such as virtual assistants, automated transcription services, machine translation, and conversational systems (Blackley et al., 2019; Ng et al., 2025; Shool et al., 2025; Thirunavukarasu et al., 2023).

However, the healthcare domain presents unique linguistic, operational and regulatory challenges that limit the direct applicability of generic AI models. Clinical language is highly specialized and characterized by domain-specific terminology, acronyms, abbreviations, implicit contextual references, and heterogeneous discourse styles (Gu et al., 2021; Carrino et al., 2022). Medical interviews involve spontaneous dialogue between healthcare professionals and patients that often combining colloquial expressions with technical vocabulary (Kuligowska et al., 2023). In contrast, medical dictations and clinical reports are more structured and formal, with a greater emphasis on terminology. This variability poses significant challenges for speech recognition and language understanding systems (Hodgson and Coiera, 2016).

Although state-of-the-art ASR models have achieved strong performance in controlled environments, their accuracy tends to degrade in domain-specific contexts such as hospitals, where background noise, overlapping speech, and speaker variability are common (Lamy et al., 2018). Moreover, generic ASR systems often misrecognize specialized medical terminology, resulting in transcription errors that can have critical implications for diagnosis, treatment planning, and clinical documentation (Zuchowski and Göller, 2022). There is a significant unmet need for domain-adapted ASR systems that are fine-tuned using multilingual medical datasets, particularly in Spanish and Catalan.

Concurrent with the development of ASR, LLMs have precipitated a paradigm shift in NLP by facilitating sophisticated functionalities, including text summarization, information extraction, reasoning, and structured generation (Brown et al., 2020). In the medical domain, LLMs exhibit considerable promise in automating clinical documentation, generating structured reports from raw transcripts, extracting relevant medical entities, and mapping them to standardized ontologies such as SNOMED-CT and ICD-10 (Thirunavukarasu et al., 2023). Spe-

cialized variants such as BioBERT (Lee et al., 2020) and ClinicalBERT (Huang et al., 2019) have exhibited enhancements in biomedical text mining tasks. However, the majority of high-performing generative models are predominantly trained on English data, thereby constraining their robustness in other languages.

The presence of multiple languages in a given context introduces a degree of complexity that must be addressed. While Spanish is extensively incorporated in NLP resources, Catalan is underrepresented in substantial medical datasets. The paucity of high-quality annotated corpora for medical speech and clinical text in these languages imposes constraints on the fine-tuning and evaluation of domain-adapted transformer models. Additionally, healthcare systems mandate strict adherence to data protection regulations (e.g., GDPR), interoperability standards, and traceability requirements, which generic AI deployments frequently neglect to address adequately.

Another critical challenge pertains to reliability and robustness. LLMs have been observed to generate hallucinations, defined as outputs that are either factually inaccurate or fabricated, particularly when operating outside the parameters of their training distribution or when tasked with generating structured medical content, as evidenced by recent (Huang et al., 2025). In healthcare environments, where precision and accountability are essential, such risks must be mitigated through controlled generation strategies, structured output constraints, validation layers, and human-in-the-loop supervision mechanisms (Kim et al., 2025).

From a technological maturity perspective, many AI-based healthcare documentation solutions are still in the experimental or early prototype stage. Integrating ASR, LLM-based summarization, NER, and MEL into a unified, scalable, and interoperable pipeline poses significant research and engineering challenges. Additionally, the computational requirements for deploying large transformer models can be prohibitive for small and medium-sized healthcare institutions. This makes model optimization techniques, such as quantization, parameter-efficient fine-tuning, and resource-aware deployment, essential for real-world adoption.

In this context, there is a clear need for an end-to-end framework tailored to multilingual healthcare environments. This framework should combine domain-specific speech recognition, structured language generation, semantic normalization, interoperability with clinical coding systems, secure data governance, and scalable deployment via SaaS architectures. Addressing these gaps will not only reduce administrative burden and improve documentation efficiency but also enhance data quality, interoperability, and clinical decision support in modern healthcare systems.

3. System Architecture

The MedicaLLM platform is designed as a modular end-to-end pipeline that integrates domain-adapted ASR and LLMs to support multilingual (Spanish and Catalan) clinical documentation workflows. As described in the Figure 1, the global SaaS system is organized into four main modules: (1) ASR for medical interviews and dictation; (2) medical report generation with summarization and error correction; (3) NER and MEL with normalization to clinical standards; and (4) user interfaces and service communication. The architecture separates perception (speech-to-text), clinical reasoning and structuring (LLM-based generation), semantic enrichment (NER/MEL), and interaction layers (SaaS UI and APIs), enabling independent optimization and scalable deployment in real healthcare environments.

At runtime, a healthcare professional records or uploads audio from either a doctor–patient interview or a clinician dictation. Audio is securely transmitted to the ASR module, which produces a transcript while preserving domain terminology and, in interview scenarios, applying speaker diarization to distinguish patient and clinician turns. The resulting text is routed to the report generation module, where an LLM transforms unstructured transcripts into structured clinical documentation (e.g., symptoms, diagnosis, treatment plan), incorporating summarization and context-aware error correction to mitigate transcription artifacts. Next, the NER/MEL module extracts relevant clinical entities and links them to standardized terminologies (e.g., SNOMED-CT, ICD-10), enabling interoperability and downstream integration with EHR systems. Finally, the outputs are presented to the user through the SaaS interface, where clinicians can validate and edit results, and optionally export them through secure API integrations.

All core services are conceived to operate under a SaaS deployment model, supporting both real-time and offline workflows, and enabling integration with external hospital systems via RESTful APIs. The following subsections describe each module in more detail.

3.1. ASR Module for Medical Interviews and Dictations

The ASR module constitutes the perceptual backbone of the MedicaLLM architecture. It objective is to convert spoken medical interactions into accurate textual representations while preserving clinical terminology, speaker structure and contextual coherence.

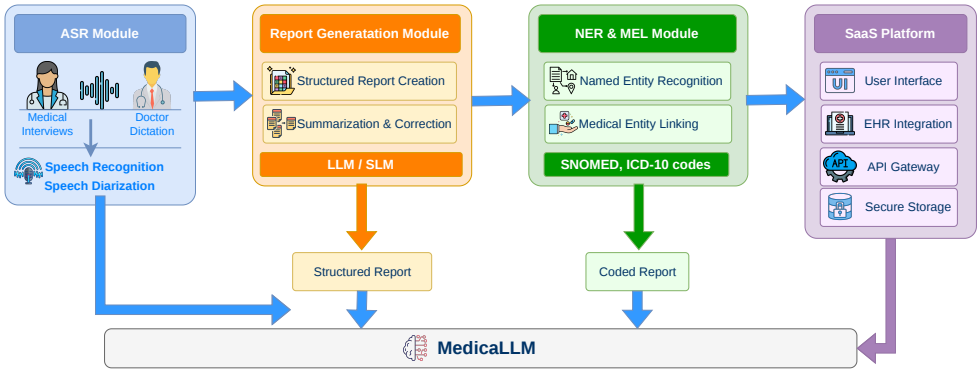

Figure 1: Overview of the modules that conform the MedicaLLM system architecture.

Unlike generic speech recognition systems, the MedicaLLM ASR component is specifically adapted to the medical domain. It addresses two distinct but complementary use cases:

1. **Medical interviews:** spontaneous, multi-speaker interactions between clinician and patient, characterized by turn-taking, colloquial expressions, incomplete sentences, interruptions and embedded technical terminology.

2. **Medical dictation:** structured, terminology-dense speech produced by healthcare professionals, often including acronyms, abbreviations and highly domain-specific expressions.

The module will be built upon transformer-based end-to-end ASR architectures (e.g., Whisper, Wav2Vec 2.0, or HuBERT variants), which will be fine-tuned using multilingual domain-specific datasets generated within the project. These datasets will include manually transcribed real and simulated consultations, augmented audio reflecting hospital noise conditions, and diverse accents to ensure robustness across clinical settings.

For interview scenarios, speaker diarization mechanisms will be integrated to distinguish between clinician and patient turns. This separation is essential for downstream structured report generation, where the attribution of symptoms and statements must be correctly contextualized. For dictation scenarios, additional preprocessing steps will be applied, including phonetic normalization of acronyms and abbreviation expansion modeling, enabling the system to handle expressions such as letter-by-letter pronunciations and abbreviated clinical terminology.

The fine-tuning strategy will evaluate multiple adaptation configurations, including encoder freezing, full fine-tuning, and gradual unfreezing, optimizing the trade-off between domain adaptation and the preservation of general acoustic representations. Performance will be evaluated using Word Error Rate (WER), domain-specific terminology accuracy, and Real-Time Factor (RTF), ensuring both transcription precision and operational feasibility.

The output of this module will be a timestamped, optionally diarized transcript that will serve as structured input for the LLM-based report generation module.

3.2. LLM-Based Medical Report Generation, summarization and Error Correction

The report generation module transforms raw ASR transcripts into structured, standardized, and clinically meaningful documentation. This process is expected to significantly reduce the administrative workload for healthcare professionals and improve consistency in clinical reporting.

The module will leverage multilingual LLMs and SLMs (e.g., Gemma-3, Mistral, Phi-4, and Qwen-3) adapted to the medical domain through supervised fine-tuning and instruction tuning on curated clinical datasets. This process will incorporate transcripts, structured reports, and annotated data to ensure alignment with clinical writing conventions.

This component performs three tightly integrated functions:

3.2.1. Structured Report Generation

The system transforms unstructured conversational transcripts into organized reports divided into predefined sections such as, chief complaint, history of present illness, physical examination findings, diagnosis and treatment plan.

This structuring will be achieved through prompt conditioning, instruction tuning, and template-guided decoding strategies, ensuring compliance with standardized medical documentation formats.

3.2.2. Abstractive and Extractive summarization

The module uses hybrid summarization techniques. Abstractive summarization will condense long transcripts into coherent, concise clinical narratives, while extractive mechanisms will ensure that critical factual elements, such as drug names, dosages, and lab results, are explicitly preserved. This dual strategy is expected to reduce information loss while maintaining readability.

Different advanced approaches will be progressively evaluated to support this objective. These approaches include encoder–decoder architectures, such as mT5 and mBART, for multilingual abstractive summarization. Subsequently, LLMs will be assessed to enhance contextual coherence, domain adaptation, and structured generation capabilities.

3.2.3. Context-Aware Error Correction

Given that ASR systems may introduce transcription errors, especially in terminology-dense contexts, the LLM incorporates contextual correction mechanisms. By leveraging domain knowledge encoded during fine-tuning, the model identifies probable inconsistencies and corrects medical terminology when contextual evidence supports such correction. This step enhances reliability and reduces post-editing workload.

Prompt engineering strategies (zero-shot, few-shot, and chain-of-thought prompting) will be evaluated to improve structured generation under computational constraints. Additionally, model optimization techniques such as quantization and QLoRA will be explored to enable deployment in resource-constrained healthcare infrastructures.

The output of this module is a structured, linguistically refined clinical report ready for semantic enrichment.

3.3. NER and Medical Entity Linking for Normalization

The semantic enrichment layer enhances interoperability and machine-readability of the generated reports by extracting and normalizing clinically relevant entities.

This module consists of two sequential components:

3.3.1. Named Entity Recognition (NER)

The NER subsystem identifies entities such as symptoms, diagnoses, medications, procedures, date, dose of the medication, organization, laboratory tests, and among others.

Two complementary strategies will be evaluated:

- **LLM-based extraction via prompt engineering**, enabling flexible zero-shot or few-shot recognition in low-resource contexts.
- **Fine-tuned encoder-based models** (e.g., BERT, RoBERTa, XLM-R variants) trained on manually annotated corpora, optimized for computational efficiency.

Performance is measured using precision, recall and F1-score, ensuring clinical-grade entity extraction accuracy.

3.3.2. Medical Entity Linking (MEL)

Following entity extraction, the MEL subsystem will map surface-level terms to standardized ontology entries such as SNOMED-CT and ICD-10 codes. Transformer-based embedding approaches (e.g., SapBERT-style representations) will be used to generate contextual embeddings for candidate concepts, which will be retrieved through similarity search mechanisms.

To address ambiguity and near-synonymous terminology, a re-ranking layer will be incorporated to refine candidate selection using context-sensitive scoring. This step is critical in clinical scenarios where subtle semantic differences may correspond to distinct diagnostic codes.

The output will be an enriched clinical report annotated with standardized codes, enabling seamless integration into EHR systems and supporting downstream analytics.

3.4. SaaS Platform, User Interfaces and Service Communication

The final layer of the architecture corresponds to the SaaS deployment infrastructure and user interaction components. The platform is designed to operate as a scalable cloud-based service while supporting integration with on-premise hospital systems when required.

3.4.1. Backend Layer

The backend orchestrates the full operational lifecycle of the system, coordinating ASR processing, LLM inference, semantic enrichment workflows, dataset management, and model versioning within a unified service layer. It manages task scheduling, resource allocation, logging, and inter service communication, ensuring that each module, including speech recognition, report generation, and entity normalization, operates cohesively within the overall pipeline.

Secure storage mechanisms are integrated at multiple levels, including encrypted data at rest and in transit, role based access control, audit logging, and traceability of model outputs and edits. The

infrastructure is designed to comply with GDPR requirements and healthcare data protection standards, supporting anonymization or pseudonymization strategies where appropriate, as well as controlled data retention policies.

The architecture supports both synchronous real time and asynchronous batch processing workflows. In real time scenarios, such as live dictation or consultation transcription, low latency inference pipelines ensure timely feedback to clinicians. In asynchronous mode, larger audio files or bulk documentation tasks can be processed through batch execution, optimizing computational resource utilization. This dual capability enables flexible deployment across diverse healthcare environments, from small clinics to large hospital networks.

3.4.2. Frontend Interfaces

User interfaces provide healthcare professionals with intuitive and user-centered tools that support the full clinical documentation workflow. Through the interface, users can securely record consultations in real time or upload previously recorded audio files. They are able to review automatically generated transcripts with timestamp navigation, compare diarized speaker segments when applicable, and detect potential transcription inconsistencies. The interface also enables validation of structured clinical reports, allowing clinicians to examine how information has been organized into predefined sections such as symptoms, diagnosis and treatment plan. In addition, users can inspect extracted clinical entities along with their associated standardized codes, verify their correctness within context, and perform manual corrections when necessary. Before final submission, professionals can edit, refine and formally approve the documentation to ensure clinical accuracy and completeness.

Visualization dashboards complement these tools by providing operational transparency. They offer real-time and aggregated metrics on transcription quality (e.g., error rates and processing latency), report generation performance, entity extraction statistics, and dataset usage indicators. These monitoring capabilities support quality assurance, facilitate model evaluation and version comparison, and enable data-driven optimization of the overall system.

3.4.3. API and Integration Layer

RESTful APIs enable integration with third-party hospital systems, EHRs, and external services. An API gateway manages authentication, access control and secure communication between frontend and backend services.

The SaaS architecture is designed following microservices principles, enabling containerized deployment, horizontal scalability and independent updating of ASR, LLM and NER components. This modular design ensures maintainability, resilience and adaptability to evolving clinical requirements.

3.5. Implementation Status and Validation

To clarify the distinction between contributions and ongoing work, this section details the implementation status of each component, the datasets used, and the validation strategy adopted in the MedicaLLM project.

MedicaLLM is currently under development within a structured work plan and is expected to reach a TRL 6 prototype, validated in realistic clinical environments. The system follows a modular design, where different components are at different stages of maturity:

- **Data resources (In progress)**: Multilingual datasets in Spanish and Catalan have been created, including real and simulated doctor–patient conversations, medical dictations, and structured clinical reports. These datasets are manually transcribed and annotated by experts, forming gold-standard corpora for ASR, NER, and MEL tasks.
- **ASR module (under development)**: Baseline transformer-based ASR systems (e.g., Whisper, Wav2Vec 2.0) have been selected and are currently being adapted through domain-specific fine-tuning using the generated datasets. Optimization strategies such as encoder freezing and gradual fine-tuning are under evaluation.
- **LLM-based report generation (under development)**: Initial pipelines for structured report generation, summarization, and error correction have been implemented using multilingual LLMs. Ongoing work focuses on instruction tuning, prompt optimization, and alignment with clinical documentation standards.
- **NER and MEL modules (under development)**: Annotated datasets for entity recognition and linking have been completed. Model development is ongoing, combining fine-tuned encoder-based models and LLM-based extraction approaches, along with ontology alignment to SNOMED-CT and ICD-10.
- **Integrated SaaS platform (planned integration phase)**:The full end-to-end integration of all modules into a scalable SaaS platform, including API-based interoperability with EHR systems, is currently under development and will be validated in later project stages.

The operational pipeline follows a well-defined sequence. First, audio is acquired either from medical interviews or clinician dictations. The signal is processed by the ASR module, which generates a transcript and optionally performs speaker diarization to distinguish between patient and clinician turns. The resulting text is then transformed by the LLM-based module into structured clinical reports, incorporating summarization and context-aware error correction. Subsequently, NER and MEL components extract relevant clinical entities and map them to standardized terminologies such as SNOMED-CT and ICD-10. The normalized output is then presented to clinicians through the user interface, where validation and editing take place before final export and integration into EHR systems. This design ensures a clear separation between perception, reasoning, semantic normalization, and user interaction layers.

The system relies on domain-specific datasets generated within the project, addressing the scarcity of medical resources in Spanish and Catalan. These datasets include transcribed doctor–patient conversations (both real and simulated), clinical dictations, structured medical reports, and synthetic audio generated via text-to-speech to improve robustness under diverse acoustic conditions. In addition, manually annotated corpora have been created for NER and MEL tasks, enabling supervised training and evaluation of semantic extraction and normalization components. All data collection and processing follow strict GDPR-compliant protocols, including anonymization procedures and expert validation.

Given the risks associated with generative models in clinical contexts, MedicaLLM incorporates multiple complementary hallucination mitigation strategies. These include template-guided and structured generation, which constrains outputs to predefined clinical sections, as well as hybrid extractive–abstractive summarization techniques that preserve critical factual information while ensuring readability. Furthermore, context-aware error correction mechanisms leverage domain knowledge during LLM inference to detect and resolve inconsistencies. Ontology grounding through MEL aligns generated outputs with standardized medical terminologies, and human-in-the-loop validation ensures that all outputs are reviewed before clinical use. Together, these mechanisms transform hallucination mitigation from a conceptual objective into an operational component of the system.

Human validation constitutes a central element of the workflow. Through the SaaS interface, clinicians can review transcripts and structured reports, inspect extracted entities and their associated standardized codes, and perform manual corrections where necessary. Users are able to edit, refine, and formally approve the final report, ensuring clinical accuracy and completeness. Only validated outputs are exported to EHR systems, guaranteeing reliability, traceability, and regulatory compliance.

Although large-scale experimental results are still under development, the evaluation framework is already defined. The ASR module is assessed using metrics such as WER, domain-specific terminology accuracy, and RTF. The NER component is evaluated through precision, recall, and F1-score, while MEL performance is measured in terms of linking accuracy and ranking quality. At the system level, evaluation includes clinician validation time, correction rate, and usability indicators.

4. Conclusions and Further Work

MedicaLLM integrates the core technological components required for end-to-end multilingual clinical documentation, including a domain-adapted ASR module for medical interviews and dictation, LLM-based structured report generation with summarization and error correction capabilities, a semantic enrichment layer for NER and MEL, and a secure SaaS platform enabling user interaction and EHR integration. The architecture combines speech perception, structured generation, ontology alignment and human-in-the-loop validation within a modular and scalable microservices framework tailored to healthcare environments.

Currently, the platform is being developed and validated through structured work packages covering multilingual resource creation, ASR adaptation, LLM fine-tuning, ontology alignment, SaaS development, and integrated system validation. These activities include the systematic construction and annotation of domain-specific datasets, the progressive adaptation of speech and language models to clinical terminology, and the iterative evaluation of transcription accuracy and structured report generation quality. The validation framework assesses the structural and semantic consistency of automatically generated medical reports, as well as the precision and recall of entity extraction and normalization components. Particular emphasis is placed on minimizing transcription artifacts, reducing clinically unsafe hallucinations, and ensuring full alignment with standardized medical coding systems and interoperability requirements.

Future work will concentrate on several complementary research and development directions. First, we plan to enhance domain robustness through adaptive fine-tuning strategies that incorporate continual learning mechanisms, enabling the ASR and LLM components to evolve as new clinical terminology, interaction styles and specialty-specific vocabularies emerge. Second, we aim to explore tighter integration between report genera-

tion and ontology normalization, investigating joint modeling approaches where structured generation is directly constrained by standardized terminologies during decoding. Finally, we will investigate automated feedback loops between clinician corrections and model refinement pipelines. By incorporating validated post-edits into incremental training cycles, the system will progressively improve its structured generation accuracy and semantic consistency.

Acknowledgements

This research is part of the research project MedicaLLM (CPP2024-011574) funded by MICIU/AEI /10.13039/501100011033 and the European Regional Development Fund (ERDF EU/FEDER UE)- a way of making Europe.

5. Ethical Considerations and Limitations

The data used for system validation and adaptation consist of enterprise-provided incident records, policy documents and internally generated test cases. All materials are handled within controlled corporate environments in accordance with applicable data protection regulations. No personal or sensitive information is publicly released as part of this research.

The system is designed to operate within enterprise infrastructures, ensuring that multimodal inputs (including text, images and audio) are processed under organizational data governance policies. When required, anonymization and access control mechanisms are applied to prevent unauthorized exposure of customer information.

6. Bibliographical References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Suzanne V Blackley, Jessica Huynh, Liqin Wang, Zfania Korach, and Li Zhou. 2019. Speech recognition for clinical documentation from 1990 to 2018: a systematic review. *Journal of the american medical informatics association*, 26(4):324–338.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Casimiro Pio Carrino, Joan Llop, Marc Pàmies, Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Joaquín Silveira-Ocampo, Alfonso Valencia, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Pretrained biomedical language models for clinical NLP in Spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 193–199, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Tobias Hodgson and Enrico Coiera. 2016. Risks and benefits of speech recognition for clinical documentation: a systematic review. *Journal of the american medical informatics association*, 23(e1):e169–e179.

A Jay Holmgren, Julia Adler-Milstein, and Nate C Apathy. 2024. Electronic health record documentation burden crowds out health information exchange use by primary care physicians: Article examines electrnoic health record documentation burden. *Health Affairs*, 43(11):1538–1545.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou,

Zehan Li, Xiaoqian Jiang, Zhiyong Lu, et al. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1812–1820.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, Lizhou Fan, Eugene Park, Tristan Lin, Joonsik Yoon, Wonjin Yoon, Maarten Sap, Yulia Tsvetkov, Paul Liang, Xuhai Xu, Xin Liu, Daniel McDuff, Hyeonhoon Lee, Hae Won Park, Samir Tulebaev, and Cynthia Breazeal. 2025. Medical hallucination in foundation models and their impact on healthcare. *medRxiv*.

Karolina Kuligowska, Maciej Stanusch, and Marek Koniew. 2023. Challenges of automatic speech recognition for medical interviews-research for polish language. *Procedia Computer Science*, 225:1134–1141.

Manuel Lamy, Rúben Pereira, João C Ferreira, José Braga Vasconcelos, Fernando Melo, and Iria Velez. 2018. Extracting clinical information from electronic medical records. In *International Symposium on Ambient Intelligence*, pages 113–120. Springer.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Joel Jia Wei Ng, Eugene Wang, Xinyan Zhou, Kevin Xiang Zhou, Charlene Xing Le Goh, Gabriel Zheng Ning Sim, Hiang Khoon Tan, Serene Si Ning Goh, and Qin Xiang Ng. 2025. Evaluating the performance of artificial intelligence-based speech recognition for clinical documentation: a systematic review. *BMC medical informatics and decision making*, 25(1):236.

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

Sina Shool, Sara Adimi, Reza Saboori Amleshi, Ehsan Bitaraf, Reza Golpira, and Mahmood Tara. 2025. A systematic review of large language model (llm) evaluations in clinical medicine. *BMC Medical Informatics and Decision Making*, 25(1):117.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Matthias Zuchowski and Aydan Göller. 2022. Speech recognition for medical documentation: an analysis of time, cost efficiency and acceptance in a clinical setting. *British Journal of Healthcare Management*, 28(1):30–36.