

# An End-to-End Framework for Clinical Documentation from Speech to Structured Medical Records

Pedro José Vivancos-Vicente<sup>1</sup>, Juan Salvador Castejón-Garrido<sup>1</sup>, Ronghao Pan<sup>2</sup>, Tomás Bernal-Beltrán<sup>2</sup>, Jorge Gómez-Navalón<sup>2</sup>, José Antonio García-Díaz<sup>2</sup> and Rafael Valencia-García<sup>2</sup>

<sup>1</sup>VÓCALI SISTEMAS INTELIGENTES S.L. Parque Científico de Murcia, Carretera de Madrid km 388. Complejo de Espinardo, 30100 Murcia, Spain

<sup>2</sup>Departamento de Informática y Sistemas, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain

## Abstract

In this project, named MedicalLLM, we present a framework designed to enhance clinical documentation by integrating Automatic Speech Recognition (ASR) and Large Language Models (LLMs) tailored to multilingual healthcare settings. Although LLMs demonstrate considerable promise, they encounter difficulties in clinical settings, such as processing specialized terminology, managing multilingual variability (e.g. Spanish and Catalan), generating hallucinated content and meeting stringent reliability standards. These issues can result in inaccuracies and increased clinician workload. Our proposed framework tackles these issues by combining domain-specific ASR with LLM-driven structured report generation and semantic enrichment via Named Entity Recognition (NER) and Medical Entity Linking (MEL). The proposed framework is intended to transform spoken clinical interactions into structured, standardized, and interoperable records aligned with SNOMED CT and ICD-10. It is designed to improve reliability and mitigate hallucinations through structured generation, hybrid summarization, ontology grounding, and human-in-the-loop validation.

## Keywords

Automatic Speech Recognition, Large Language Models, Medical Entity Linking, Named Entity Recognition

## 1. Introduction

The burden of clinical documentation has become a significant challenge in modern healthcare. Clinicians spend a significant amount of time on administrative tasks, which reduces the time available for patient care. Meanwhile, healthcare systems require accurate, structured, and standardized records to meet regulatory and interoperability demands. These competing demands highlight the limitations of traditional documentation workflows. In this context, Automatic Speech Recognition (ASR) and Large Language Models (LLMs) offer a promising opportunity to improve the capture, organization, and communication of clinical information [1, 2, 3].

However, translating this potential into real-world clinical practice remains challenging. Although models such as BERT [4], GPT-based architectures [5], LLaMA-3 [6], Qwen-3 [7], and Whisper [8] have demonstrated strong performance in general domains, they do not fully account for the complexity of clinical language. Medical communication involves specialized terminology, frequent use of abbreviations, and multilingual variation, particularly in settings where Spanish and Catalan coexist [9]. As a result, general-purpose ASR and LLM systems often struggle to produce accurate transcriptions, correctly interpret clinical concepts, or generate structured outputs aligned with standards such as SNOMED CT and ICD-10. These limitations are further compounded by practical concerns, including

SEPLN 2026: 42<sup>nd</sup> International Conference of the Spanish Society for Natural Language Processing, León, Spain, 22-25 September 2026.

✉ pedro.vivancos@vocali.net (P. J. Vivancos-Vicente); juans.castejon@vocali.net (J. S. Castejón-Garrido); ronghao.pan@um.es (R. Pan); tomas.bernalb@um.es (T. Bernal-Beltrán); jorge.gomezn@um.es (J. Gómez-Navalón); joseantonio.garcia8@um.es (J. A. García-Díaz); valencia@um.es (R. Valencia-García)

🌐 <https://www.vocali.net/> (P. J. Vivancos-Vicente); <https://www.vocali.net/> (J. S. Castejón-Garrido);

<https://github.com/Smolky> (J. A. García-Díaz); <https://webs.um.es/valencia> (R. Valencia-García)

🆔 0009-0008-7317-7145 (R. Pan); 0009-0006-6971-1435 (T. Bernal-Beltrán); 0009-0005-7185-6054 (J. Gómez-Navalón); 0000-0002-3651-2660 (J. A. García-Díaz); 0000-0003-2457-1791 (R. Valencia-García)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

data privacy requirements, computational costs, sensitivity to noisy environments, and the risk of hallucinated outputs [10, 11, 12].

To bridge the gap between technological potential and clinical applicability, we propose **MedicaLLM**, an end-to-end Artificial Intelligence (AI) framework for clinical documentation. The system combines domain-specific ASR with specialized LLM components to support tasks such as structured report generation, Named Entity Recognition (NER), and Medical Entity Linking (MEL). MedicaLLM is designed to connect these components and produce structured, semantically consistent outputs. The proposed framework aims to capture clinical information from spoken interactions and represent it in a form suitable for integration into healthcare systems. Its intended outcome is to reduce the administrative burden on healthcare professionals and improve the consistency, quality, and interoperability of clinical documentation.

The main contribution of MedicaLLM is the design of a unified multilingual framework that connects clinical speech recognition, structured report generation, semantic extraction, ontology-based normalization, and clinician validation. This integrated architecture is intended to support the generation of structured and semantically consistent clinical documentation suitable for future integration into healthcare systems.

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.

Overall, MedicaLLM aims to bridge the gap between recent advances in speech and language technologies and the practical requirements of healthcare systems. This is expected to contribute to more efficient, standardized, and scalable clinical documentation practices. To achieve this goal, the project is structured around five main objectives: (OB1) generation of multilingual resources for training and fine-tuning end-to-end ASR and LLM-based models in the medical domain; (OB2) development of end-to-end ASR models for medical interviews and clinical dictation scenarios; (OB3) design of LLM-based models for generating structured medical reports from transcriptions, incorporating summarization and error correction mechanisms; (OB4) development of LLM-based approaches for Named Entity Recognition (NER) and Medical Entity Linking (MEL); and (OB5) development of a software as a service (SaaS) platform and user interfaces for system deployment and interaction.

## 2. Background Information

In recent years, Transformer-based architectures have led to significant progress in natural language processing and speech technology. Models such as BERT [4], GPT-3 [5], LLaMA-3 [6], Qwen-3 [7], Whisper [8], HuBERT [13], and Wav2Vec 2.0 [14] have demonstrated near-human performance on various tasks, including contextual language understanding, speech transcription, and generative reasoning. These advances have enabled the widespread adoption of AI systems in applications such as virtual assistants, automated transcription, and machine translation [2, 15, 3, 16].

Despite this progress, applying these capabilities to the healthcare domain reveals a fundamental mismatch between general-purpose AI systems and the specific requirements of clinical practice. Clinical language cannot be treated as a simple extension of general language. Rather, it is a distinct and highly specialized register. It combines domain-specific terminology, dense and often ambiguous abbreviations, implicit contextual cues, and heterogeneous discourse styles that vary across institutions and even within a single clinical interaction [10, 9]. For example, clinical conversations often alternate between informal, patient-centered dialogue and highly technical medical descriptions [17], while formal reports require strict structure and terminological precision [18].

These characteristics present significant challenges for existing AI systems. Even state-of-the-art ASR models, which perform well in controlled environments, tend to degrade in realistic clinical settings where background noise, overlapping speech, and speaker variability are common [19]. Such degradation is not merely a matter of reduced accuracy: errors in transcribing medical terminology can

have direct consequences for clinical documentation, decision-making, and ultimately patient safety [20].

The problem is further compounded in multilingual contexts. While Spanish benefits from relatively strong NLP resources, Catalan remains underrepresented, particularly in the clinical domain. The scarcity of high-quality annotated corpora for medical speech and clinical text in both languages limits the effectiveness of model adaptation, fine-tuning, and evaluation. Additionally, strict regulatory requirements, including GDPR compliance, interoperability standards, and auditability constraints, impose additional barriers that generic AI solutions are not designed to address.

Beyond transcription and language understanding, generative models introduce additional risks. A well-documented limitation of LLMs is their tendency to produce hallucinated outputs—statements that sound plausible but are factually incorrect or unsupported [21]. In healthcare, where accuracy and accountability are critical, this behavior poses a significant risk. Addressing this issue requires the incorporation of structured output constraints, validation mechanisms, and human-in-the-loop oversight [12]. Although domain-specific models such as BioBERT [22] and ClinicalBERT [23] have improved in biomedical text processing, they predominantly focus on English and do not fully address multilingual deployment or system-level integration challenges.

Finally, the challenge extends beyond individual models to the level of system design. Building a practical clinical documentation solution requires the integration of multiple components—ASR, LLM-based summarization, NER, and MEL—into a unified and interoperable pipeline. This integration poses both research and engineering challenges, especially considering the computational demands of large Transformer models. Without optimization strategies such as quantization, parameter-efficient fine-tuning, and resource-aware deployment, these systems remain inaccessible to many healthcare institutions.

Taken together, these limitations highlight the need to shift from improving isolated models to designing comprehensive systems. Instead of relying on general-purpose models evaluated on medical benchmarks, end-to-end frameworks must be designed specifically for multilingual clinical environments. These frameworks must incorporate domain-specific speech recognition, structured language generation, and semantic normalization aligned with standards such as SNOMED CT and ICD-10, as well as robust mechanisms for privacy, reliability, and deployment. The central motivation behind the MedicalLLM project is addressing this gap.

### 3. System Architecture

Designing an intelligent clinical documentation system requires balancing two often conflicting demands: the high level of precision required in medical contexts and the flexibility needed for real-world deployment. The proposed framework addresses this challenge with a sequence of four interconnected processing stages, each of which is responsible for a specific transformation—from raw audio input to validated, standards-compliant clinical documentation. Rather than treating the system as a monolithic pipeline, the architecture maintains clear functional boundaries between stages. This allows each component to evolve independently as underlying models improve. As illustrated in Figure 1, the framework is organized into four main modules: (i) an ASR module for medical interviews and dictations; (ii) an LLM-based module for generating medical reports; (iii) a semantic enrichment module based on NER and MEL with normalization to clinical standards, and (iv) a SaaS platform providing user interfaces, service orchestration, and system integration.

In the proposed workflow, the system will process audio from either a doctor–patient interaction or a clinician dictation. The audio will be securely transmitted to the ASR module, which will generate a transcript while preserving domain-specific terminology. In consultation scenarios, speaker diarization will distinguish between patient and clinician contributions. The transcript will then be processed by the report generation module, where an LLM will transform the unstructured text into structured clinical documentation. The semantic enrichment stage will subsequently extract relevant clinical entities and link them to standardized terminologies. Finally, clinicians will review, validate, and edit

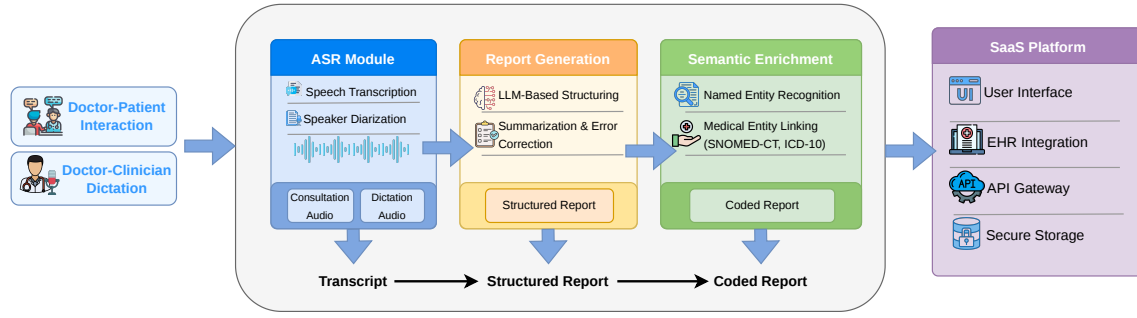


**Figure 1:** Overview of the modules comprising the MedicaLLM system architecture.

the generated outputs before their potential integration into EHR systems.

### 3.1. ASR Module

The ASR module will constitute the perceptual backbone of the MedicaLLM framework. Its objective will be to convert spoken clinical interactions into accurate textual representations while preserving medical terminology, speaker attribution, and contextual coherence.

Clinical speech presents significant challenges due to acoustic variability, including background noise, overlapping speech, and speaker diversity, as well as linguistic complexity arising from frequent shifts between colloquial and technical language [19, 20]. To address these challenges, the module will distinguish between two scenarios: (i) multi-speaker doctor-patient interviews and (ii) clinician dictations. Each scenario will require a specific adaptation strategy.

The module will be based on Transformer-based ASR architectures such as Whisper, Wav2Vec 2.0, and HuBERT [8, 14, 13]. These models will be adapted to the clinical domain through fine-tuning on multilingual Spanish and Catalan datasets developed within the project. The datasets will include manually transcribed real and simulated consultations, acoustically augmented recordings reproducing hospital conditions, and diverse speaker profiles to improve robustness.

For interview scenarios, speaker diarization mechanisms will be incorporated to distinguish between clinician and patient turns, which will be essential for downstream interpretation. For dictation scenarios, preprocessing mechanisms will be explored to normalize acronyms and expand abbreviations. Different adaptation strategies, such as encoder freezing, gradual unfreezing, and full fine-tuning, will be investigated to balance domain specialization and generalization.

Performance will be evaluated using Word Error Rate (WER), medical terminology accuracy, speaker diarization performance, and Real-Time Factor (RTF). The expected output will be a timestamped and optionally diarized transcript that will serve as input for the subsequent processing stages.

### 3.2. LLM-Based Clinical Report Generation

The report generation module will transform raw ASR transcripts into structured and clinically meaningful documentation, bridging the gap between conversational input and formal medical records.

The module will explore multilingual LLMs (e.g., Gemma-3 [24], Phi-4 [25] and Qwen-3 [7]), which will be adapted to the clinical domain through prompt engineering, supervised fine-tuning, and instruction tuning on curated datasets containing transcripts and expert-authored reports. The development process will focus on alignment with clinical documentation conventions and standardized reporting formats.

The module will perform three main functions. First, structured report generation will organize transcript content into predefined sections such as chief complaint, history of present illness, examination findings, diagnosis, and treatment plan. Prompt conditioning, instruction tuning, and template-guided decoding will be investigated to support this task.

Second, summarization methods will be developed to condense long transcripts into coherent clinical narratives. A hybrid extractive–abstractive approach will be explored to preserve critical information, including medications, dosages, laboratory values, symptoms, and diagnoses.

Third, context-aware error correction mechanisms will be investigated to mitigate transcription errors by using clinical context to detect terminology inconsistencies and potentially incorrect ASR outputs. These mechanisms will be designed to reduce the amount of manual correction required while avoiding unsupported modifications [12].

Optimization strategies such as quantization and parameter-efficient fine-tuning, including QLoRA, will also be assessed to support deployment in resource-constrained environments. The expected output of this module will be a structured and linguistically refined clinical report ready for semantic enrichment.

### **3.3. Semantic Enrichment**

The semantic enrichment module will be designed to make generated clinical reports interoperable and machine-readable by extracting and normalizing relevant clinical entities.

The Named Entity Recognition component will identify entities such as symptoms, diagnoses, medications, procedures, laboratory tests, dates, and dosage expressions. Two complementary approaches will be evaluated: (i) LLM-based extraction using zero-shot or few-shot prompting, particularly for low-resource settings, and (ii) fine-tuned encoder-based models (e.g., BERT [4], RoBERTa [26], XLM-R [27]) trained on annotated corpora for efficient and accurate extraction. Performance will be evaluated using precision, recall, and F1-score.

The Medical Entity Linking component will map extracted entities to standardized terminologies such as SNOMED CT and ICD-10. Contextual embeddings and similarity-based retrieval methods will be investigated to generate candidate concepts, followed by re-ranking mechanisms to resolve ambiguity and distinguish between closely related clinical terms.

MEL performance will be evaluated using linking accuracy and ranking-based metrics, including top-k accuracy. The expected output will be an enriched clinical report annotated with standardized medical codes, enabling interoperability with electronic health record systems.

### **3.4. SaaS Platform**

The MedicaLLM framework will be deployed through a scalable SaaS architecture designed for real-world clinical environments. The platform will follow a modular structure in which the ASR, report generation, and semantic enrichment components will operate as independent but interconnected services.

The backend will be developed to manage service orchestration, resource allocation, model lifecycle management, secure data processing, and communication between components. It will support both real-time and batch execution workflows. Data protection mechanisms will include encryption at rest and in transit, role-based access control, audit logging, anonymization or pseudonymization, and GDPR-compliant data management procedures.

The user interface will be designed to allow clinicians to record or upload audio, review generated transcripts, validate structured reports, inspect extracted entities and their corresponding standardized codes, and introduce corrections when necessary. Human validation will remain mandatory before any output is approved for integration into an electronic health record system.

RESTful APIs will be developed to support interoperability with EHR systems and other external clinical services. A microservices-based architecture will be adopted to enable scalability, maintainability, and the independent updating of the individual modules.

The platform will also be evaluated in terms of inference latency, processing throughput, memory consumption, hardware requirements, usability, system reliability, and successful integration with external healthcare systems.

### 3.5. Implementation Status, Evaluation and Validation Strategy

The system will rely on domain-specific datasets being developed within the project to address the limited availability of medical resources in Spanish and Catalan. These datasets are expected to include transcriptions of real and simulated doctor–patient interactions, clinical dictations, structured reports, synthetic audio generated through text-to-speech, and manually annotated corpora for NER and MEL. These resources will support the training and evaluation of the different components under diverse clinical and acoustic conditions. All data-handling procedures will follow GDPR requirements, including anonymization or pseudonymization, secure storage, restricted access, and expert-driven validation.

At the current stage, MedicaLLM is in the initial phase of the research and development project and should not yet be considered a fully implemented or validated production-ready system. The project is currently focused on the collection, preparation, and annotation of multilingual clinical resources, together with the initial design and development of the individual modules. Baseline ASR architectures and candidate language models have been identified, while preliminary pipelines for structured report generation, summarization, NER, MEL, and context-aware error correction are being designed and explored. The integration of these components into the SaaS platform will be carried out in subsequent phases of the project. Consequently, this paper primarily presents the motivation, architecture, methodological design, and planned validation strategy of the framework.

The evaluation will be performed at both component and end-to-end levels. The ASR module will be assessed using WER, medical terminology accuracy, speaker diarization performance, and RTF. Results will be analyzed separately for doctor–patient interviews and clinical dictations, since these scenarios differ in terms of speaker interaction, acoustic variability, and terminology density.

The clinical report generation module will be evaluated in terms of structural completeness, factual consistency, preservation of clinically relevant information, omission rate, hallucination rate, and the amount of clinician correction required. Particular attention will be paid to the correct preservation of symptoms, diagnoses, medications, dosages, laboratory values, temporal expressions, and negation. Automatic evaluation will be complemented by expert assessment, since conventional text-similarity metrics may not adequately reflect the clinical accuracy and safety of generated reports.

NER performance will be evaluated using entity-level precision, recall, and F1-score, while MEL will be assessed through linking accuracy and ranking-based metrics. Results will also be analyzed by entity type and by the degree of ambiguity of the clinical concepts. At the system level, the evaluation will consider documentation-time reduction, clinician validation time, correction rate, report completeness, usability, clinician satisfaction, and successful interoperability with EHR systems.

Considering the risks associated with generative models in healthcare, MedicaLLM will incorporate several complementary hallucination-mitigation strategies. These include constrained generation based on predefined clinical templates, hybrid extractive–abstractive summarization, context-aware correction, ontology grounding through MEL, and human-in-the-loop validation. Clinicians will be able to review transcripts, inspect structured reports, verify extracted entities and standardized mappings, introduce corrections, and formally approve the final output before its transfer to an EHR system.

A systematic error analysis will be conducted at both the component and pipeline levels. ASR errors will include general lexical errors, medical terminology substitutions, abbreviation-related errors, speaker-attribution errors, and errors caused by background noise or overlapping speech. Report-generation errors will include omitted clinical facts, unsupported additions, incorrect negation, altered numerical values, and inconsistencies between the source transcript and the generated report. NER and MEL errors will be analyzed according to incorrect entity boundaries, entity-type confusion, missed concepts, ambiguous mappings, and incorrect ontology codes. Particular attention will be paid to error propagation across modules.

The evaluation will also consider the computational and operational requirements of the framework under realistic deployment conditions. Measurements will include inference latency, processing throughput, memory consumption, hardware requirements, and RTF for both the individual modules and the integrated pipeline. Different optimization strategies, including quantization and parameter-efficient fine-tuning, will be assessed to analyze the trade-off between clinical performance, response time, and

computational cost.

Finally, results will be reported separately for Spanish and Catalan. Since Catalan has fewer available clinical resources, its limited data coverage may affect robustness and generalization. To mitigate this limitation, the project will investigate multilingual transfer learning, acoustic and textual data augmentation, synthetic speech generation, and cross-lingual adaptation. Nevertheless, the scarcity of Catalan clinical resources will be explicitly considered when interpreting the experimental results.

## 4. Conclusions and Further Work

This project represents an initial step toward the development of reliable, multilingual, and clinically grounded AI systems for medical documentation. MedicaLLM proposes the integration of domain-adapted ASR, LLM-based report generation, and semantic enrichment to support the future conversion of spoken clinical interactions into structured, standardized, and interoperable medical records. The proposed framework is designed to address multilingual variability, terminology normalization, and alignment with clinical standards. Its intended outcome is to reduce the administrative burden on healthcare professionals and improve the consistency, quality, and interoperability of clinical documentation.

As MedicaLLM is currently in the initial phase of the project, the present work focuses on establishing the methodological and architectural foundations of the proposed framework. The current activities are centered on collecting and annotating domain-specific resources for ASR, NER, MEL, and clinical report generation, as well as defining the initial configurations of the individual modules. In subsequent project phases, these resources will support model adaptation, training, and systematic evaluation. The assessment will benchmark each component under realistic clinical conditions and evaluate the complete system through metrics such as documentation-time reduction, correction rate, inference latency, and clinician satisfaction. Comparative analyses with existing clinical documentation approaches will also be conducted once the integrated prototype becomes available.

Future project phases will focus on enhancing the system’s robustness and adaptability in real-world clinical environments. Specifically, we plan to optimize ASR performance in challenging acoustic conditions, incorporate advanced domain-adaptive techniques for LLMs, and enhance disambiguation capabilities in medical entity linking. Additionally, we aim to extend the system to support a broader range of clinical specialties and documentation formats. We also plan to explore multimodal inputs that combine speech and structured clinical data.

Finally, future developments will emphasize ensuring compliance with regulatory and ethical requirements. This includes reinforcing GDPR-compliant data handling procedures, improving anonymization and security mechanisms, and implementing transparent validation and auditing processes. Particular attention will be given to safely deploying generative models in healthcare by incorporating stricter hallucination control strategies and expanding human-in-the-loop validation to guarantee reliability and accountability in clinical practice.

## Acknowledgments

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.

## Declaration on Generative AI

During the preparation of this work, the authors used DeepL in order to: Grammar and spelling check. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

## References

- [1] A. J. Holmgren, J. Adler-Milstein, N. C. Apathy, Electronic health record documentation burden crowds out health information exchange use by primary care physicians: Article examines electronic health record documentation burden, *Health Affairs* 43 (2024) 1538–1545.
- [2] S. V. Blackley, J. Huynh, L. Wang, Z. Korach, L. Zhou, Speech recognition for clinical documentation from 1990 to 2018: a systematic review, *Journal of the American Medical Informatics Association* 26 (2019) 324–338.
- [3] S. Shool, S. Adimi, R. Saboori Amleshi, E. Bitaraf, R. Golpira, M. Tara, A systematic review of large language model (llm) evaluations in clinical medicine, *BMC Medical Informatics and Decision Making* 25 (2025) 117.
- [4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), 2019, pp. 4171–4186.
- [5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, *Advances in neural information processing systems* 33 (2020) 1877–1901.
- [6] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, *arXiv preprint arXiv:2407.21783* (2024).
- [7] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 technical report, *arXiv preprint arXiv:2505.09388* (2025).
- [8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: *International conference on machine learning*, PMLR, 2023, pp. 28492–28518.
- [9] C. P. Carrino, J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, M. Villegas, Pretrained biomedical language models for clinical NLP in Spanish, in: D. Demner-Fushman, K. B. Cohen, S. Ananiadou, J. Tsujii (Eds.), *Proceedings of the 21st Workshop on Biomedical Language Processing*, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 193–199. URL: <https://aclanthology.org/2022.bionlp-1.19/>. doi:10.18653/v1/2022.bionlp-1.19.
- [10] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, *ACM Transactions on Computing for Healthcare (HEALTH)* 3 (2021) 1–23.
- [11] Y. Hu, Q. Chen, J. Du, X. Peng, V. K. Keloth, X. Zuo, Y. Zhou, Z. Li, X. Jiang, Z. Lu, et al., Improving large language models for clinical named entity recognition via prompt engineering, *Journal of the American Medical Informatics Association* 31 (2024) 1812–1820.
- [12] Y. Kim, H. Jeong, S. Chen, S. S. Li, M. Lu, K. Alhamoud, J. Mun, C. Grau, M. Jung, R. Gameiro, L. Fan, E. Park, T. Lin, J. Yoon, W. Yoon, M. Sap, Y. Tsvetkov, P. Liang, X. Xu, X. Liu, D. McDuff, H. Lee, H. W. Park, S. Tulebaev, C. Breazeal, Medical hallucination in foundation models and their impact on healthcare, *medRxiv* (2025). doi:10.1101/2025.02.28.25323115.
- [13] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, Hubert: Self-supervised speech representation learning by masked prediction of hidden units, *IEEE/ACM transactions on audio, speech, and language processing* 29 (2021) 3451–3460.
- [14] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, *Advances in neural information processing systems* 33 (2020) 12449–12460.
- [15] J. J. W. Ng, E. Wang, X. Zhou, K. X. Zhou, C. X. L. Goh, G. Z. N. Sim, H. K. Tan, S. S. N. Goh, Q. X. Ng, Evaluating the performance of artificial intelligence-based speech recognition for clinical documentation: a systematic review, *BMC medical informatics and decision making* 25 (2025) 236.
- [16] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, D. S. W. Ting, Large language models in medicine, *Nature medicine* 29 (2023) 1930–1940.

- [17] K. Kuligowska, M. Stanusch, M. Koniew, Challenges of automatic speech recognition for medical interviews-research for polish language, *Procedia Computer Science* 225 (2023) 1134–1141.
- [18] T. Hodgson, E. Coiera, Risks and benefits of speech recognition for clinical documentation: a systematic review, *Journal of the american medical informatics association* 23 (2016) e169–e179.
- [19] M. Lamy, R. Pereira, J. C. Ferreira, J. B. Vasconcelos, F. Melo, I. Velez, Extracting clinical information from electronic medical records, in: *International Symposium on Ambient Intelligence*, Springer, 2018, pp. 113–120.
- [20] M. Zuchowski, A. Göller, Speech recognition for medical documentation: an analysis of time, cost efficiency and acceptance in a clinical setting, *British Journal of Healthcare Management* 28 (2022) 30–36.
- [21] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et al., A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, *ACM Transactions on Information Systems* 43 (2025) 1–55.
- [22] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics* 36 (2020) 1234–1240.
- [23] K. Huang, J. Altosaar, R. Ranganath, Clinicalbert: Modeling clinical notes and predicting hospital readmission, *arXiv preprint arXiv:1904.05342* (2019).
- [24] G. Team, Gemma 3 (2025). URL: <https://arxiv.org/abs/2503.19786>.
- [25] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, et al., Phi-4 technical report, *arXiv preprint arXiv:2412.08905* (2024).
- [26] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, *arXiv preprint arXiv:1907.11692* (2019).
- [27] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, *CoRR abs/1911.02116* (2019). URL: <http://arxiv.org/abs/1911.02116>. *arXiv:1911.02116*.