TEME: A Multi-Agent Evaluation Framework for Spanish Medical Speech Recognition

Extended Abstract

Leire Villarroya-Martinez
Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València
Camí de Vera s/n 46022, Valencia
Spain
lvilmar1@epsg.upv.es

Stella Heras
Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València
Camí de Vera s/n 46022, Valencia
Spain
stehebar@upv.es

Javier Palanca
Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València
Camí de Vera s/n 46022, Valencia
Spain
jpalanca@dsic.upv.es

Vicente Botti
Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València & Valencian Graduate School and Research Network of Artificial Intelligence (VALGRAI)
Camí de Vera s/n 46022, Valencia
Spain
vbotti@dsic.upv.es

Edwin Tadeo-Gomez
Hospital Juaneda
Camí de Vileta 30, 07011, Balearic Islands, Spain
edwintadeo@gmail.com

Enrique Alcazar Garzas
Omniloy
Avenida de Cataluña 11, 46020, Valencia, Spain
enrique.alcazar@omniloy.com

ABSTRACT

Conventional metrics like Word Error Rate (WER) fail to differentiate between minor variations and potentially life-threatening medical mistakes. We introduce TEME (Medical Accuracy Test in Spanish), a supervised multi-agent evaluation framework for Spanish medical Automatic Speech Recognition (ASR). TEME employs a two-layer architecture with specialized agents assessing transcriptions for clinical awareness, overseen by a consensus agent that applies safety rules. Testing on 90 validated clinical dialogues shows that TEME successfully captures clinically relevant error severity that conventional metrics miss, providing a safety-aware alternative for medical ASR evaluation.

KEYWORDS

Medical ASR; Multi-Agent Systems; LLM Evaluation; Clinical Error Assessment; Healthcare

ACM Reference Format:
Leire Villarroya-Martinez, Stella Heras, Javier Palanca, Vicente Botti, Edwin Tadeo-Gomez, and Enrique Alcazar Garzas. 2026. TEME: A Multi-Agent Evaluation Framework for Spanish Medical Speech Recognition: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/RHNT7438

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/RHNT7438

1 INTRODUCTION

Automatic Speech Recognition (ASR) is vital for reducing the clinical documentation burden, which currently consumes over 50% of healthcare providers' time [3]. However, safe deployment remains constrained by evaluation methods; standard metrics like WER and CER treat all errors uniformly, regardless of clinical significance [7, 9]. This uniform treatment is misaligned with medical reality, where confusing a medication name or altering a dosage can have life-threatening consequences while only counting as a single-token mismatch. While domain-aware measures like Medical Concept WER (MC-WER) [1] or Clinical BERTScore [6] improve sensitivity to terminology, they remain static scores that lack explicit mechanisms for assessing clinical severity or explaining errors. Furthermore, Spanish medical ASR faces unique challenges due to a lack of high-quality clinical corpora and diverse dialectal variations.

We argue that medical ASR evaluation should be a supervised, agent-based reasoning task. We introduce TEME (Medical Accuracy Test in Spanish), a multi-agent framework that decomposes transcription assessment into specialized agents—focused on medication, dosage, and consistency—operating under a safety-oriented consensus mechanism. TEME provides a clinically grounded, explainable evaluation layer essential for safety-critical. Recent studies demonstrate that structured coordination and deliberation among multiple agents significantly outperform single-agent baselines in safety-critical clinical tasks [2, 5, 11].

2 TEME: MULTI-AGENT EVALUATION FRAMEWORK

TEME implements a supervised two-layer architecture for clinical ASR evaluation, all models are based on GPT-4o [8].The first

layer consists of three domain-specialized evaluation agents, each responsible for a distinct clinical dimension. **Medication agent** evaluates the correctness of pharmaceutical terminology, detecting cases where a transcribed medication corresponds to a different drug or therapeutic class. Superficial spelling variants or formatting differences are ignored if clinical meaning is preserved. **Dosage agent** validates quantities, units, and frequencies associated with medications. It detects clinically relevant changes in dosage while ignoring equivalent reformulations (e.g., “200 mg/day” vs. “200 milligrams per day”). **Consistency agent** assesses overall clinical coherence, including symptoms, diagnoses, allergies, and instructions. Particular attention is paid to negations and omissions that may alter clinical meaning. **Synthesized Classification Criteria.** Agents classify results into three categories: NONE (no clinical change), MINOR (non-critical deviation), or MAJOR (safety-critical error). A MAJOR error is triggered by drug identity changes, significant dosage deviations, or inverted clinical meanings (e.g., "no allergies" → "have allergies"). The second layer is a meta-agent **consensus agent** that filters decisions to ensure agents remain in their domain . A deterministic rule ensures that any MAJOR error detected by an expert agent prevails in the final classification.

To evaluate the proposed framework, we present the first clinically validated dataset for Spanish medical dialogues, consisting of 90 conversations (67,000 words) across 10 specialties some such as cardiology, neurology and paediatrics. The corpus combines 23 real anonymized consultations with 67 realistic synthetic dialogues validated by a physician. The audio was generated using Google Gemini’s TTS engine. The audio was deliberately degraded (8-bit, 8kHz) to simulate real clinical environments before being processed by the transcribers [12].

3 RESULTS AND DISCUSSION

We evaluated TEME in a data set of 90 Spanish medical dialogues covering ten specialties, transcribed using two ASR systems: a general-purpose model and a medical ASR system adapted to the domain. Traditional metrics (WER, CER), domain metrics (MC-WER), Clinical BERTScore adaptation and SeMaScore adaptation were calculated together with the TEME evaluations. The adaptation of Clinical BERTScore is based on the original proposal by Mani et al. (2020) for English, we implemented a version for Spanish . This adaptation uses the PlanTL-GOB-ES/roberta-base-biomedical-clinical-es [4] language model and a curated clinical vocabulary from the TEME dataset. The metric combines general and medically weighted similarity using the formula $CBERTScore(x, \hat{x}) = k \cdot BERTScore_{medical}(x, \hat{x}) + (1-k) \cdot BERTScore_{all}(x, \hat{x})$, with a control factor $k = 0.4$. In the adaptation of SeMaScore we implemented the methodology of Sasindran et al. (2024) [10], which integrates penalties for error rate with contextual similarity. Our version uses the same biomedical model in Spanish and performs a segment alignment between reference and hypothesis, applying penalties based on the Match Error Rate (MER) to capture structural and semantic reliability.

Quantitative analysis (see Table 1) shown that Whisper achieved a lower WER (3.3%) than Omniloy-medical-voice (4.7%), yet TEME’s evaluation revealed that Whisper produced 21 MAJOR errors compared to Omniloy’s 12 . This discrepancy proves that traditional

Table 1: System Performance Summary: TEME Error Counts and ASR Metric Means.

Metric / Aspect	Omniloy-medical-voice	Whisper
TEME — Error Counts (N) (None/Low/Severe)		
Medication (Med)	81/5/6	62/11/17
Dosage (Dose)	84/1/5	85/1/4
Consistency (Consist)	70/18/2	75/8/7
Total Error Count	62/16/12	56/13/21
ASR Mean Values (↓ desirable)		
WER (↓ Error)	4.7%	3.3%
CER (↓ Error)	3.6%	2.5%
MC-WER (↓ Error)	4.3%	4.8%
ClinicalBERTScore (↑ Score)	96.1%	97.2%
SeMaScore (↑ Score)	86.4%	87.9%

NLP metrics systematically underestimate clinical risk, as they treat all substitutions uniformly regardless of their medical consequences . For instance, Whisper introduced severe clinical distortions by replacing "analgésico" (analgesic) with "energético" (energetic) and "férula" (splint) with "célula" (cell), which represent significant risks to patient safety despite having minimal impact on error rates. Furthermore, factual failures such as transcribing the pharmaceutical "Atorzet" as the non-existent medication "Torset" yielded a deceptively high ClinicalBERTScore of 0.9687 . This demonstrates that embedding-based metrics lack awareness of medical validity and can mask dangerous mistakes with high similarity scores . Critical dosage hazards further illustrated this limitation; a change from "35 mg midday" to "80 mg morning" was correctly flagged as a MAJOR error by TEME, even though the system maintained a relatively low MC-WER of 0.0962 in that context . These findings underscore the necessity of a safety-aware framework like TEME to assess the semantic coherence and clinical trustworthiness of medical transcriptions .

4 CONCLUSION AND FUTURE WORK

This work introduced a supervised multi-agent framework for evaluating medical recognition of Spanish-speaking patients that addresses the limitations of conventional metrics that do not capture clinical severity. By combining specific domain agents for medication, dosage and consistency under a deterministic consensus layer, TEME provides an explainable, safety-sensitive and clinically sound assessment. Future work includes extending the agent set to additional clinical dimensions, refining inter-agent calibration, and exploring the integration of TEME into real-world ASR pipelines and other languages.

ACKNOWLEDGMENTS

This work was partially supported by the Spanish Government through projects PID2021-123673OB-C31 and PID2024-158227NB-C33, funded by MICIU/AEI/10.13039/501100011033/ ERDF/EU, and by the Valencian Government through grant CIPROM/2021/077.

REFERENCES

[1] Ayo Adedeji, Sarita Joshi, and Brendan Doohan. 2024. The Sound of Healthcare: Improving Medical Transcription ASR Accuracy with Large Language Models. *arXiv preprint arXiv:2402.07658* (2024). https://doi.org/10.48550/arXiv.2402.07658

[2] Waleed Almansoori, Mohamed Elhoseiny, and Shuaib Shah. 2025. MedAgentSim: Self-evolving Multi-Agent Simulations for Realistic Clinical Interactions. *arXiv preprint arXiv:2503.22678* (2025). https://doi.org/10.48550/arXiv.2503.22678

[3] Brian G. Arndt, John W. Beasley, Michael D. Watkinson, Jonathan L. Temte, Wen-Jan Tuan, Christine A. Sinsky, and Valerie Gilchrist. 2017. Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. *Annals of Family Medicine* 15, 5 (2017), 419–426. https://doi.org/10.1370/afm.2121

[4] Casimiro Pio Carrino, Jordi Armengol-Estapé, Asier Gutiérrez-Fandiño, Joan Llop-Palao, Marc Pàmies, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. Biomedical and Clinical Language Models for Spanish: On the Benefits of Domain-Specific Pretraining in a Mid-Resource Scenario. arXiv:2109.03570 [cs.CL] https://arxiv.org/abs/2109.03570

[5] Abdine Maiga, Anoop Shah, and Emine Yilmaz. 2025. Error Detection in Medical Notes through Multi-Agent Debate. In *Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP 2025)*. Association for Computational Linguistics, 124–135. https://doi.org/10.18653/v1/2025.bionlp-1.12

[6] Siddharth Mani, Anmol Pal, and Parminder Bhatia. 2020. Clinical BERTScore: An Improved Measure of Automatic Speech Recognition Performance in Clinical Settings. *arXiv preprint arXiv:2007.12626* (2020). https://doi.org/10.48550/arXiv.2007.12626

[7] Andrew Morris, Viktoria Maier, and Phil Green. 2004. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *Interspeech 2004*. https://doi.org/10.21437/Interspeech.2004-668

[8] OpenAI. 2024. GPT-4o. https://openai.com/research/gpt-4o. Large language model.

[9] Somnath Roy. 2021. Semantic-WER: A Unified Metric for the Evaluation of ASR Transcript for End Usability. *CoRR* abs/2106.02016 (2021). arXiv:2106.02016 https://arxiv.org/abs/2106.02016

[10] Zitha Sasindran, Harsha Yelchuri, and T. V. Prabhakar. 2024. SeMaScore: A new evaluation metric for automatic speech recognition tasks. In *Interspeech 2024*. ISCA, 4558–4562. https://doi.org/10.21437/Interspeech.2024-2033

[11] Samuel Schmidgall, Ruidong Ding, Liyan Jiang, Swaroop Mishra, and Chandan K. Reddy. 2025. AgentClinic: Multi-agent Multimodal Benchmark for Clinical LLMs. *arXiv preprint arXiv:2405.07960* (2025). https://doi.org/10.48550/arXiv.2405.07960

[12] L. Villarroya Martínez. 2026. *Spanish Medical Dialogue Dataset (TEME-v1)*. Zenodo. https://doi.org/10.5281/zenodo.17280661