An Argumentation-Based Framework for Evaluating Spanish Medical ASR

Leire Villarroya-Martinez[1,*], Stella Heras[1], Javier Palanca[1] and Vicente Botti[1,2]

[1]Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, Camí de Vera s/n, 46022 Valencia, Spain
[2]Valencian Graduate School and Research Network of Artificial Intelligence (VALGRAI), Spain

Abstract

Automatic Speech Recognition (ASR) is increasingly used in clinical documentation, but metrics such as Word Error Rate (WER) and Character Error Rate (CER) treat benign variations and clinically dangerous errors alike. We present an argumentation-based framework for evaluating Spanish medical ASR outputs. Specialized clinical evaluators detect medication, dosage, and consistency discrepancies between a reference transcription and an ASR hypothesis, producing severity labels and explanations. These outputs are formalized as arguments that support or attack the claim that the hypothesis is clinically acceptable. The safety-oriented defeat procedure is non-compensatory: one accepted major medication, dosage, or semantic error is sufficient to require review, while harmless formatting and recoverable orthographic variants may be defeated by equivalence arguments. On 90 clinically validated Spanish medical dialogues, results show that better lexical scores may coexist with more clinically critical errors, especially in medication and consistency dimensions.

Keywords

Computational Argumentation, Clinical ASR Evaluation, Medical Speech Recognition, Explainable AI, Healthcare AI

1. Introduction

Automatic Speech Recognition (ASR) is increasingly used in clinical settings to support medical documentation and reduce administrative burden. However, the evaluation of ASR systems in healthcare remains misaligned with the requirements of the domain. Standard metrics, such as Word Error Rate (WER) and Character Error Rate (CER), quantify transcription quality in terms of lexical differences, treating all substitutions, deletions, and insertions uniformly [1, 2]. Embedding-based metrics such as BERTScore and clinical adaptations of BERTScore capture semantic similarity more effectively than string matching, but they do not explicitly represent clinical risk [3, 4]. In clinical contexts, errors are not homogeneous: a transcript may contain multiple harmless formatting variations, or a single error that changes a medication, dose, negation, symptom, or instruction.

This mismatch is especially important in medical ASR. A transcription that changes a drug name, administration frequency, or negated symptom can be unsafe even if its overall WER is low. Conversely, a transcription may obtain a worse lexical score because of paraphrasing, spelling variation, or formatting differences while preserving the clinically relevant information. Evaluation therefore requires reasoning about discrepancies, their clinical dimension, their severity, and their interaction with the rest of the transcript.

Large Language Models (LLMs) can identify and explain complex inconsistencies [5, 6], but LLM-as-judge evaluation raises concerns about opacity, bias, prompt sensitivity, and auditability [7, 8]. In safety-critical domains, it should be possible to inspect which discrepancies were detected, why they matter, and how they determine the final classification.

CMNA26 – The 26th Computational Models of Natural Argument, September 14–15, 2026
*Corresponding author.
lvilmar1@epsg.upv.es (L. Villarroya-Martinez); stehebar@upv.es (S. Heras); jpalanca@dsic.upv.es (J. Palanca); vbotti@dsic.upv.es (V. Botti)
0009-0003-5300-8209 (L. Villarroya-Martinez); 0000-0001-6212-9377 (S. Heras); 0000-0002-6209-9603 (J. Palanca); 0000-0002-6507-2756 (V. Botti)

We address this need by reframing clinical ASR evaluation as an argumentation problem. Given a reference transcription R and an ASR hypothesis H, specialized evaluators identify medication, dosage, and consistency discrepancies. Their outputs are mapped into structured arguments that support or attack the claim that H is clinically acceptable with respect to R. Our contributions are: (i) a formalization of clinical ASR evaluation over reference–hypothesis pairs; (ii) a mapping from evaluator outputs to arguments, attacks, defeats, and decisions; and (iii) an empirical analysis showing that lexical metrics and clinical-risk-oriented evaluation may rank ASR systems differently.

2. Background and Related Work

2.1. Lexical and Semantic Metrics for Clinical ASR

ASR evaluation has traditionally relied on lexical distance measures such as WER and CER. These metrics are useful for general benchmarking, but they treat all token differences as equivalent. In medical contexts, this assumption is problematic because the clinical consequences of transcription differences vary substantially. Substituting a common function word is not comparable to substituting a medication name or altering a dose.

Domain-aware metrics have been proposed to address this limitation. Medical Concept WER (MC-WER) restricts error computation to medically relevant concepts and is therefore more clinically focused than WER [9]. Semantic metrics such as BERTScore [3], Clinical BERTScore [4], and SeMaScore [10] attempt to capture meaning preservation rather than surface identity. Nevertheless, these metrics still produce scalar similarity scores and do not explicitly represent why a transcription is clinically safe or unsafe. Recent studies of clinical ASR have shown that systems may obtain low WER while missing or distorting clinically important information [11].

2.2. Knowledge- and LLM-Based Clinical Error Detection

Knowledge-based approaches incorporate structured medical resources into ASR correction and evaluation. For example, MedSpeak uses medical knowledge graphs and phonetic information to improve ASR error correction for spoken medical question answering [12]. Such approaches move beyond text matching by grounding transcription hypotheses in domain knowledge.

In parallel, LLMs have been increasingly used to detect inconsistencies in clinical documentation. Recent work shows that LLMs can identify complex errors in medical notes, and that prompt optimization can substantially improve detection performance on benchmarks such as MEDEC [5, 6, 13]. However, LLM-based assessment is sensitive to prompt formulation and may provide judgments without a transparent reasoning structure. Surveys on LLM-as-a-judge emphasize the need for reliability, bias control, clear rubrics, and meta-evaluation when LLMs are used as evaluators [7, 8].

2.3. Multi-Agent Clinical Reasoning and Argumentation

Multi-agent LLM systems have recently been explored for clinical reasoning and error detection. Debate-based approaches allow multiple agents to exchange arguments before a judge agent makes a final decision [14]. Medical multi-agent benchmarks and simulations such as AgentClinic and MedAgentSim further illustrate the potential of structured delegation and supervision in healthcare-oriented reasoning tasks [15, 16]. ArgMed-Agents explicitly connects LLM-based clinical reasoning with argumentation schemes and directed conflict graphs [17].

Computational argumentation provides a formal basis for representing conflicting information and non-monotonic reasoning. Abstract argumentation frameworks define arguments and attack relations, enabling acceptability to be evaluated under different semantics [18]. Structured frameworks such as ASPIC+ incorporate premises and defeasible inference rules [19]. Preference- and value-based argumentation frameworks allow priorities to guide conflict resolution [20]. Quantitative argumentation models such as DF-QuAD offer gradual assessments of argument strength [21]. Argumentation has

also been studied in clinical decision support and evidence aggregation [22, 23], as a foundation for explainable AI [24], and as an argument-based approach to clinical outcome assessment validation [25].

Despite these developments, the use of formal argumentation for clinical ASR evaluation remains underexplored. Existing approaches typically rely either on aggregate similarity metrics or on LLM judgments without an explicit representation of attacks, defeat, and justification. We address this gap by modeling clinical ASR evaluation as an argumentation-based acceptability problem.

3. Dataset and Experimental Setup

3.1. Dataset

The evaluation was conducted on TEME-v1, a Spanish medical dialogue corpus introduced in the TEME framework for Spanish medical ASR evaluation [26] and released as a dataset [27]. The corpus consists of 90 anonymized clinical conversations across 10 medical specialties: cardiology, primary care, pulmonology, neurology, traumatology, endocrinology, oncology, geriatrics, emergency medicine, and pediatrics. It includes both structured JSON files and audio recordings. Overall, it contains 66,941 words and 2,330 conversational turns, with an average of 25.9 turns and 743.8 words per dialogue.

It was built from 23 anonymized real cardiology consultations and 67 synthetic but clinically realistic dialogues. The real consultations were stripped of personally identifiable information and linguistically refined while preserving clinical content; the synthetic dialogues broaden specialty coverage while preserving realistic terminology and doctor–patient interaction patterns.

Clinical validation was performed by a qualified cardiologist with broad clinical experience. It covered dialogue plausibility, medical terminology, patient-safety relevance, evaluator outputs, and the criteria used to distinguish NONE, MINOR, and MAJOR errors. While a sample of 90 dialogues might seem modest for training large language models, it yields a dense corpus of 2,330 turns requiring exhaustive, turn-by-turn manual validation by a clinical specialist. For the purpose of introducing and validating the logical architecture of a qualitative, argumentation-based framework, this scale is highly appropriate. It prioritizes deep clinical rigor and explicit reasoning over purely statistical scaling. We acknowledge as a limitation that validation was performed by a single specialist; future work should include multiple clinicians and inter-annotator agreement.

3.2. ASR Systems and Preprocessing

Each dialogue includes a reference transcription R and an ASR hypothesis H. The reference R is treated as the clinically correct version of the dialogue, while H is evaluated with respect to whether it preserves the clinically relevant information in R. We compare two ASR systems: Whisper-large-v3, a general-purpose multilingual ASR system, and a proprietary ASR model specialized in Spanish medical transcription. The second system is referred to as the specialized Spanish medical ASR model.

Both systems were evaluated using the same pipeline. The ASR outputs were normalized before metric computation by lowercasing, reducing punctuation inconsistencies, and standardizing numerical expressions. The goal of preprocessing was to prevent clinically irrelevant formatting variation from dominating the evaluation.

3.3. Specialized Clinical Evaluators

Clinical discrepancy detection is performed by three specialized evaluators, all implemented using GPT-4o with default generation settings [28]. Each evaluator receives the pair (R, H) and focuses on a specific clinical dimension. Each returns a severity label—NONE, MINOR, or MAJOR—together with a textual explanation grounded in the comparison between R and H.

The *Medication evaluator* assesses whether pharmaceutical terminology is preserved. It detects errors when drug identity changes, when a medication is hallucinated, or when a transcribed term could be

interpreted as a different medication. Formatting differences, abbreviations, capitalization, and spelling variants are ignored when the drug identity remains recoverable and clinically unchanged.

The *Dosage evaluator* assesses numerical values, units, and administration frequency. It ignores medication-name errors, which belong to the Medication evaluator, and focuses on posology. Equivalent expressions such as "200 mg/day" and "200 milligrams per day", or "0.5 mg" and "half a milligram", are treated as clinically equivalent.

The *Consistency evaluator* assesses whether the overall clinical meaning of the consultation is preserved. It focuses on symptoms, diagnoses, allergies, clinical instructions, negation, temporal information, and contextual coherence. Medication and dosage errors are excluded from its scope unless they also affect broader coherence.

The final decision is not obtained by allowing errors to compensate for one another. Instead, a safety-first consensus rule is applied: if any evaluator identifies an accepted MAJOR discrepancy, the transcription is classified as clinically problematic and requires review. Minor discrepancies lead to an "acceptable with minor issues" classification only when no accepted major discrepancy is present.

4. Argumentation-Based Framework

4.1. Task Definition

Let R denote the reference transcription of a clinical dialogue and let H denote the ASR hypothesis produced for the same audio. We define a clinical discrepancy as a difference between R and H that may affect medication identity, dosage, administration frequency, symptoms, diagnoses, allergies, negation, temporal information, or clinical instructions.

The main claim evaluated by the framework is:

$$C_0 : \text{The ASR hypothesis } H \text{ is clinically acceptable with respect to } R.$$

A transcription is clinically acceptable if it preserves the information required for safe interpretation of the consultation. It may contain paraphrases, formatting differences, or minor orthographic variation, provided that these do not alter clinically relevant meaning.

4.2. Arguments

Each evaluator output is mapped into an argument:

$$A = \langle p, c, s, d \rangle,$$

where p is a set of premises extracted from the comparison between R and H, c is a conclusion concerning the acceptability of H, $s \in \{\text{NONE}, \text{MINOR}, \text{MAJOR}\}$ is the severity level, and $d \in \{\text{medication}, \text{dosage}, \text{consistency}\}$ is the clinical dimension.

Arguments may support or attack the main claim C_0. An argument supports C_0 when it concludes that clinically relevant information is preserved. An argument attacks C_0 when it identifies a discrepancy that affects clinical meaning or patient safety.

For example, if R contains "Lexiana 35 mg, one at midday" and H contains "Lexiana 80 mg, one in the morning", the Dosage evaluator generates an argument whose premise is the mismatch between amount and administration time, whose conclusion is that the dosage information is not preserved, and whose severity is MAJOR. This argument attacks C_0.

4.3. Attack Relations

We distinguish two types of attack.

First, discrepancy arguments may attack C_0. A MAJOR medication, dosage, or consistency argument directly attacks C_0, because a single clinically significant error is sufficient to require human review. A

MINOR argument does not defeat C_0 by itself, but supports a weaker conclusion: H is acceptable with minor issues.

Second, arguments may attack other arguments when they provide competing interpretations of the same discrepancy. This is important for avoiding false positives. For example, a surface mismatch such as "10 mg" versus "10 milligrams" may initially appear as a dosage discrepancy. However, an equivalence argument based on identical numerical value and unit normalization attacks that discrepancy interpretation. In this case, the discrepancy is defeated because the clinical meaning is preserved.

4.4. Preference and Defeat

Severity is used as a safety-oriented priority relation, not as a compensatory score:

$$\text{MAJOR} \succ \text{MINOR} \succ \text{NONE}.$$

However, this ordering does not mean that several minor confirmations can compensate for a major error. Major arguments are non-compensatory: once accepted, they defeat the claim that the transcription is clinically acceptable. This reflects clinical practice, where one critical medication, dosage, or negation error is sufficient to require review regardless of how accurate the rest of the transcript may be.

We say that an argument A_i defeats an argument A_j if A_i attacks A_j and A_j is not strictly preferred to A_i. Equivalence arguments may defeat superficial discrepancy arguments when they establish preservation of clinical meaning. In contrast, an argument identifying a clinically significant dosage change cannot be defeated by lexical similarity or by unrelated correctly transcribed information.

4.5. Decision Procedure

The final classification is obtained as follows:

$$\text{Final}(H) = \begin{cases} \text{PROBLEMATIC}, & \text{if an accepted MAJOR argument attacks } C_0; \\ \text{MINOR-ISSUES}, & \text{if no MAJOR argument attacks } C_0 \text{ and at least one MINOR argument is present;} \\ \text{ACCEPTABLE}, & \text{if no clinical discrepancy is detected.} \end{cases}$$

The procedure below summarizes the construction and decision process.

1. Run the Medication, Dosage, and Consistency evaluators on (R, H) and map each output to $A = \langle p, c, s, d \rangle$.
2. Add attacks from discrepancy arguments to C_0 when clinical acceptability is weakened or rejected.
3. Add equivalence arguments for formatting, abbreviation, unit, number, or recoverable orthographic variation, and let them attack superficial discrepancies.
4. Accept undefeated equivalence arguments and undefeated clinically relevant discrepancy arguments.
5. Return CLINICALLY-PROBLEMATIC if any accepted MAJOR argument attacks C_0; otherwise return MINOR-ISSUES if any accepted MINOR argument remains; otherwise return ACCEPTABLE, together with accepted and defeated arguments as justification.

Thus, critical errors cannot be compensated by otherwise correct information, while superficial mismatches can be defeated when clinical meaning is preserved. In terms of formal argumentation semantics [18], this safety-first deterministic procedure closely aligns with a cautious, grounded semantics approach. By demanding that the acceptability claim (C_0) must be defended against all undefeated clinical discrepancies, the framework avoids multiple conflicting extensions. It yields a single, unambiguous skeptical extension that is highly suitable for safety-critical medical decision-making.

Table 1
Empirical evaluation results on the clinical ASR dataset.

Metric / Aspect	Whisper-large-v3	Specialized Spanish medical ASR
WER ↓	14.8%	16.9%
CER ↓	10.1%	14.2%
MC-WER ↓	4.8%	4.3%
Clinical BERTScore ↑	97.2%	96.1%
SeMaScore ↑	87.9%	86.4%
Medication Errors (None/Minor/Major)	62/11/17	79/5/6
Dosage Errors (None/Minor/Major)	85/1/4	84/1/5
Consistency Errors (None/Minor/Major)	75/8/7	70/18/2

5. Results

Table 1 reports the empirical results obtained on the 90-dialogue Spanish medical dataset. The table includes conventional lexical metrics, semantic metrics, and the distribution of clinical severity labels produced by the specialized evaluators. Lower WER, CER, and MC-WER values are better; higher Clinical BERTScore and SeMaScore values are better. The error-count rows report the number of dialogues classified as NONE/MINOR/MAJOR for each clinical dimension.

The results show a mismatch between lexical similarity and clinical safety. Whisper-large-v3 achieves better WER and CER, but produces more MAJOR medication and consistency errors. The specialized Spanish medical ASR model obtains worse WER and CER, but fewer MAJOR errors in those dimensions. Dosage errors are similar, with four MAJOR cases for Whisper-large-v3 and five for the specialized model.

Thus, lexical metrics and clinical-risk-oriented evaluation may rank systems differently. A transcript can be lexically close to the reference while still requiring review, whereas another can contain more paraphrases or formatting differences but preserve clinically relevant information in specific dimensions.

6. Argumentation-Based Case Analysis

6.1. Case 1: Lower Lexical Error but Unsafe Clinical Meaning

One illustrative case is a traumatology consultation involving a distal radius fracture. The reference states that the patient is prescribed tramadol as a strong analgesic and that a splint should be kept dry. Whisper-large-v3 preserves much of the lexical structure of the consultation, but introduces clinically relevant substitutions: “analgésico” is transcribed as “energético”, “férula” as “célula”, and “fractura de Colles” as “fractura de coles”.

The framework maps these discrepancies to arguments. A^1_{med} states that “analgesic” is transcribed as “energetic”, distorting the treatment function; it is a MAJOR attack on C_0. A^1_{cons} states that “splint” is repeatedly transcribed as “cell”, weakening or attacking C_0 because care instructions become ambiguous. A grammatical change in “do not exceed four per day” is only MINOR because the numerical limit is preserved. Since A^1_{med} is accepted, C_0 is defeated and the final decision is CLINICALLY-PROBLEMATIC, despite good surface similarity.

6.2. Case 2: Multiple Medication and Dosage Attacks

A second case involves a cardiology consultation where the patient’s medication list is reviewed. In the reference, the patient is taking Valsartán 80 mg in the morning, Lexiana 35 mg at midday, Furosemida 40 mg in the morning, Rivotril 0.5 mg, and other medication. In the ASR hypothesis, several clinically relevant substitutions occur: “Valsartán” becomes “Balsartan”, “Rivotril” becomes “Ribotril”, and “Seguril” becomes “Seuril”. More importantly, the dose of Lexiana changes from 35 mg at midday to 80 mg in the morning.

The framework represents these discrepancies as independent but convergent attacks: A^2_{med} attacks medication preservation because "Valsartán" becomes "Balsartan"; A^3_{med} and A^4_{med} weaken acceptability because "Rivotril" and "Seguril" are distorted; and A^2_{dose} is a MAJOR attack because "Lexiana 35 mg at midday" becomes "Lexiana 80 mg in the morning". The transcription is therefore CLINICALLY-PROBLEMATIC: at least one medication argument and one dosage argument attack C_0, and correctly transcribed information elsewhere cannot compensate for them.

6.3. Case 3: False Positive Avoided Through Equivalence

The framework also handles cases where surface differences should not become clinical errors. For example, "10 mg" versus "10 milligrams" or "6–8 hours" versus "six or eight hours" may produce lexical differences. However, equivalence arguments based on unit normalization and numerical preservation defeat the initial discrepancy interpretation. The final output is therefore ACCEPTABLE or MINOR-ISSUES, rather than CLINICALLY-PROBLEMATIC.

7. Discussion and Limitations

The proposed formulation separates harmless variation from clinical risk. Formatting differences, spelling variants, and equivalent dosage expressions can be defeated by equivalence arguments, whereas medication substitutions, dosage changes, missing negations, and semantic inversions attack acceptability.

The framework should be interpreted as an argumentation-based formalization of a clinical ASR evaluation pipeline, not as a fully implemented general-purpose argumentation solver. It makes evaluator reasoning explicit by representing outputs as premises, conclusions, severity levels, and clinical dimensions; future work should connect this layer to existing argumentation engines and compare semantics.

Limitations remain: the dataset has 90 dialogues; validation was performed by one qualified cardiologist; GPT-4o evaluators should be tested across prompts, models, temperatures, and local clinical language models; and the framework assesses transcription safety with respect to a reference transcript, not diagnosis or replacement of expert review.

8. Conclusion

This paper presented an argumentation-based framework for evaluating Spanish medical ASR outputs. Instead of treating ASR evaluation as a scalar similarity measurement, the framework models detected clinical discrepancies as arguments that support or attack the claim that an ASR hypothesis is clinically acceptable. The proposed safety-oriented defeat procedure makes major medication, dosage, and consistency errors non-compensatory, while allowing harmless formatting and recoverable orthographic variation to be defeated by equivalence arguments.

The empirical analysis shows that conventional metrics may fail to reflect clinical risk: Whisper-large-v3 achieves better WER and CER, but produces more major medication and consistency errors than a specialized Spanish medical ASR model. These results support the need for evaluation methods that explicitly represent clinical severity and provide auditable justifications. Future work will extend the clinical validation protocol, evaluate multiple argumentation semantics, and study whether structured argumentation-based explanations improve the efficiency and reliability of expert review.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (OpenAI) for grammar and spelling checking, and for paraphrasing and rewording to improve the clarity of the manuscript. After using

this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Morris, V. Maier, P. Green, From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition, in: Proceedings of Interspeech 2004, 2004. doi:10.21437/Interspeech.2004-668.

[2] K. Kuhn, V. Kersken, B. Reuter, N. Egger, G. Zimmermann, Measuring the accuracy of automatic speech recognition solutions, ACM Transactions on Accessible Computing 16 (2024) 1–23. doi:10.1145/3636513.

[3] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, 2019. doi:10.48550/arXiv.1904.09675. arXiv:1904.09675.

[4] J. Shor, R. A. Bi, S. Venugopalan, S. Ibara, R. Goldenberg, E. Rivlin, Clinical BERTScore: An improved measure of automatic speech recognition performance in clinical settings, in: Proceedings of the 5th Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1–7. URL: https://aclanthology.org/2023.clinicalnlp-1.1/. doi:10.18653/v1/2023.clinicalnlp-1.1.

[5] P. May, S. Nokodian, C. Nuernbergk, M. Knauer, M. Hefter, A. Becker von Rose, F. Bassermann, J. Jung, Artificial intelligence-assisted error detection in complex clinical documentation: Leveraging large language models to enhance patient safety in oncology, JCO Clinical Cancer Informatics 10 (2026) e2500194. doi:10.1200/CCI-25-00194.

[6] C. Myles, P. Schrempf, D. Harris-Birtill, Importance of prompt optimisation for error detection in medical notes using language models, in: Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026), Association for Computational Linguistics, Rabat, Morocco, 2026, pp. 236–248. URL: https://aclanthology.org/2026.healing-1.20/. doi:10.18653/v1/2026.healing-1.20.

[7] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, S. Wang, K. Zhang, Y. Wang, W. Gao, L. Ni, J. Guo, A survey on LLM-as-a-judge, 2024. doi:10.48550/arXiv.2411.15594. arXiv:2411.15594.

[8] D. Li, B. Jiang, L. Huang, A. Beigi, C. Zhao, Z. Tan, A. Bhattacharjee, Y. Jiang, C. Chen, T. Wu, K. Shu, L. Cheng, H. Liu, From generation to judgment: Opportunities and challenges of LLM-as-a-judge, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2025, pp. 2757–2791. URL: https://aclanthology.org/2025.emnlp-main.138/.

[9] A. Adedeji, S. Joshi, B. Doohan, The sound of healthcare: Improving medical transcription ASR accuracy with large language models, 2024. doi:10.48550/arXiv.2402.07658. arXiv:2402.07658.

[10] Z. Sasindran, H. Yelchuri, T. V. Prabhakar, SeMaScore: A new evaluation metric for automatic speech recognition tasks, in: Proceedings of Interspeech 2024, 2024, pp. 4558–4562. doi:10.21437/Interspeech.2024-2033.

[11] H. Hwang, E. Jordan, D. H. Kim-Dufor, C. Lemey, M. Alrahabi, Evaluating ASR in a clinical context: What Whisper misses, in: Proceedings of the 8th International Conference on Natural Language and Speech Processing, 2025, pp. 429–435. URL: https://aclanthology.org/2025.icnlsp-1.36/.

[12] Y. Song, S. Shrestha, C. Lyu, E. Khatibi, P. Zhang, H. Xu, N. Dutt, A. Rahmani, MedSpeak: A knowledge graph-aided ASR error correction framework for spoken medical QA, 2026. doi:10.48550/arXiv.2602.00981. arXiv:2602.00981.

[13] A. Ben Abacha, W.-w. Yim, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, T. Lin, MEDEC: A benchmark for medical error detection and correction in clinical notes, 2024. doi:10.48550/arXiv.2412.19260. arXiv:2412.19260.

[14] A. Maiga, A. Shah, E. Yilmaz, Error detection in medical notes through multi-agent debate, in:

Proceedings of the 24th Workshop on Biomedical Language Processing, 2025, pp. 124–135. URL: https://aclanthology.org/2025.bionlp-1.12/. doi:10.18653/v1/2025.bionlp-1.12.

[15] S. Schmidgall, R. Ding, L. Jiang, S. Mishra, C. K. Reddy, AgentClinic: A multi-agent multimodal benchmark for clinical LLMs, 2025. doi:10.48550/arXiv.2405.07960. arXiv:2405.07960.

[16] W. Almansoori, K. Kumar, H. Cholakkal, Self-evolving multi-agent simulations for realistic clinical interactions, 2025. doi:10.48550/arXiv.2503.22678. arXiv:2503.22678.

[17] S. Hong, L. Xiao, X. Zhang, J. Chen, ArgMed-Agents: Explainable clinical decision reasoning with LLM discussion via argumentation schemes, in: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2024, pp. 5486–5493. doi:10.1109/BIBM62325.2024.10822109.

[18] P. M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, Artificial Intelligence 77 (1995) 321–357. doi:10.1016/0004-3702(94)00041-X.

[19] H. Prakken, An abstract framework for argumentation with structured arguments, Argument and Computation 1 (2010) 93–124. doi:10.1080/19462166.2010.486479.

[20] T. J. M. Bench-Capon, Persuasion in practical argument using value-based argumentation frameworks, Journal of Logic and Computation 13 (2003) 429–448. doi:10.1093/logcom/13.3.429.

[21] A. Rago, F. Toni, M. Aurisicchio, P. Baroni, Discontinuity-free decision support with quantitative argumentation debates, in: Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning, 2016, pp. 63–73. URL: https://proceedings.kr.org/2016/6/.

[22] J. Fox, D. Glasspool, D. Grecu, Argumentation-based clinical decision support, in: Artificial Intelligence in Medicine, 2007, pp. 34–43. doi:10.1007/978-3-540-74126-8_3.

[23] A. Hunter, M. Williams, Aggregation of evidence using argumentation: A case study in medicine, The Knowledge Engineering Review 27 (2012) 1–27. doi:10.1017/S0269888912000106.

[24] K. Cyras, A. Rago, E. Albini, P. Baroni, F. Toni, Argumentative XAI: A survey, in: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, 2021, pp. 4392–4399. URL: https://www.ijcai.org/proceedings/2021/600. doi:10.24963/ijcai.2021/600.

[25] K. P. Weinfurt, R. J. Wirth, M. C. Edwards, B. B. Reeve, Applying the argument-based approach to validation with clinical outcome assessments: Strategies for constructing a rationale, Value in Health 29 (2026) 811–816. doi:10.1016/j.jval.2026.01.007.

[26] L. Villarroya-Martinez, S. Heras, J. Palanca, V. Botti, E. Tadeo-Gomez, E. Alcazar Garzas, Teme: A multi-agent evaluation framework for spanish medical speech recognition, AAMAS '26, 2026, p. 3365–3367.

[27] L. Villarroya Martínez, S. Heras, J. Palanca, V. Botti, E. Tadeo-Gomez, E. Alcazar Garzas, Spanish medical dialogue dataset (TEME-v1), 2026. URL: https://doi.org/10.5281/zenodo.17280661. doi:10.5281/zenodo.17280661.

[28] OpenAI, Hello GPT-4o, 2024. URL: https://openai.com/index/hello-gpt-4o/, accessed 2026-06-28.