Application of Artificial Intelligence for the Retrieval, Processing and Generation of Knowledge from Clinical Data

Doctoral Consortium

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

ABSTRACT

Current systems based on large language models (LLMs) provide useful but hardly auditable answers for drawing conclusions about clinical problems based on patient histories: it is difficult to know which evidence supports each conclusion and how alternatives were discarded. In this thesis, I propose a theoretical approach to computational argumentation orchestrated by a multiagent system in the medical domain: specialized agents capable of retrieving medical evidence, which build and evaluate argument graphs with support/attack relations and deliver traceable justifications alongside the answers. The goal is to improve verifiability, robustness and usefulness in transversal clinical tasks (e.g., medication reconciliation, treatment selection, indication of tests).

KEYWORDS

Computational Argumentation; Clinical Decision Support; Explainable AI; Healthcare AI

ACM Reference Format:
Leire Villarroya-Martínez. 2026. Application of Artificial Intelligence for the Retrieval, Processing and Generation of Knowledge from Clinical Data: Doctoral Consortium. 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/CFJJ1446

1 INTRODUCTION

Modern clinical decision making requires integrating and assessing evidence scattered across electronic health records, guidelines, reports and scientific literature. In this context, multiagent systems (MAS) have emerged as a solution to orchestrate perception, reasoning and communication capabilities among specialized components, with applications ranging from diagnostic support to therapeutic recommendations and monitoring [19]. However, beyond answering fluently, decision support systems must justify why one clinical hypothesis prevails over its alternatives, leaving a clear trail of the evidence considered and the counterevidence discarded.

Despite great advances in clinical information extraction [9] and automatic report generation [11], a gap remains: the absence of end-to-end systems that systematically integrate retrieval, normalization and computational argumentation within coordinated multiagent architectures, with formal acceptability semantics and metrics that assess argument quality in addition to textual quality (ROUGE, BLEU, BERTScore) [11, 15, 24]. In the field of explainability (XAI), attribution methods such as LIME, SHAP or CAM/Grad-CAM [13, 16, 18, 25], make it possible to visualize which features influence a model's output, but they focus on individual models and not on the process of deliberation or justification across clinical evidence. In contrast to such post-hoc explainability, computational argumentation aims to provide explicit reasoning that connects premises, counterarguments and conclusions in a verifiable and traceable way. There are proposals of linked argumentation graphs for multidisciplinary decision support [22], but their systematic integration with MAS, comparative evaluation and fine-grained traceability (claim→evidence→conclusion) remain poorly standardized. What is missing, in particular, is (i) a functional decomposition into specialized agent roles that makes the inner steps of deliberation—including support, attack, and exceptions,fully observable, (ii) formal criteria to evaluate and select sets of arguments by weighting clinical evidence levels, allowing the system to resolve conflicts through principled, transparent preference relations, and (iii) a standard set of metrics and a reproducible evaluation protocol to compare systems in transversal clinical tasks. In parallel, classical clinical MAS have emerged [19], argumentation-based XAI frameworks [2] and MAS proposals with ethical governance [5], as well as LLM-based agents such as MedAgents [7] or Agent Hospital [10]. This thesis proposes a multiagent architecture centered on argument graphs, with specialized agents for retrieval/normalization, construction, evaluation and verification, coordinated by an orchestrator, can improve the auditability, robustness and clinical usefulness of recommendations. The planned doctoral contribution is mainly theoretical and is structured in three objectives: (i) to formally define an argumentation-centred multiagent architecture and its associated formalisms; (ii) to develop a minimal set of argument-quality metrics and an evaluation protocol that go beyond purely textual aspects; and (iii) to theoretically and empirically validate the proposed system on representative clinical scenarios.

2 CONTEXT AND GAP

Our work lies at the intersection of three lines of research, focusing on multi-agent systems in which agents are driven by large language models (LLMs): multiagent systems (MAS) for supporting clinical decision making, explainability and justification techniques in AI, and computational argumentation applied to medicine. As hinted in the introduction, partial solutions already exist in each area; this section provides a more structured synthesis of their contributions and remaining gaps, as a step toward converging on an integrated architecture for argumentative clinical reasoning.

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/CFJJ1446

2.1 Multiagent Systems in Clinical Settings

Multiagent systems (MAS) have been used in Clinical Decision Support (CDS) to manage distributed knowledge (records, guidelines, literature) and integrate different sources of information and reasoning. Classical architectures combine supervisor agents, cognitive agents and rule-based reasoning modules for diagnosis, treatment and monitoring in constrained domains [15]. However, the way in which evidence is combined and prioritized is usually encoded implicitly in internal rules or heuristics, without an explicit representation of arguments and counterarguments or formal criteria for selecting warranted conclusions.

2.2 Computational Argumentation and Explainable AI

Computational argumentation provides the formal machinery to model defeasible reasoning and decisions under conflicting evidence. Dung's abstract frameworks and structured extensions such as ASPIC+ allow defining acceptability of arguments under different semantics and managing relations of support and attack [1, 6, 14]. In medicine, these frameworks have been applied to the justification of diagnoses, the evaluation of treatments and the reconciliation of guidelines, and they have been specifically adapted to explain medical decisions [3, 12]. XAI has developed attribution techniques (LIME, SHAP, CAM/Grad-CAM) and intrinsically interpretable models, useful for identifying which features influence a prediction, but focused on individual models and without capturing deliberation processes across clinical evidence. To improve traceability and structure of reasoning, combinations of information extraction, ontologies and knowledge graphs have been combined with argumentative reasoning [4, 17, 20], as well as hybrid architectures that integrate retrieval-augmented generation (RAG), knowledge graphs and systematic evaluation of clinical QA systems [8, 21, 23]. Despite these advances, three key gaps persist: (i) there is no end-to-end architecture that, within a MAS, connects in a distributed way evidence retrieval, semantic normalization (e.g., via knowledge graphs) and the construction and evaluation of argument graphs; (ii) there are no shared metrics and benchmarks to systematically measure argument quality. This is especially important in the medical domain, where "good" arguments can be judged from different perspectives, such as strength of evidence, alignment with clinical guidelines, safety/risk, and patient preferences; and (iii) even LLM-based agents that offer textual explanations rarely make explicit the relations of attack and support or the reasons why certain clinical hypotheses prevail over others.

3 DOCTORAL PROPOSAL

The proposed doctoral thesis addresses the need for clinical decision support systems with traceable, intrinsic justification by designing, implementing, and evaluating a MAS architecture with LLM-based agents, centered on computational argumentation, in which LLMs are used to interpret clinical evidence and produce explicit arguments and counterarguments.

3.1 Multiagent Architecture for Clinical Argumentation

We propose a MAS architecture that decomposes the end-to-end process of generating justifications into specialized roles. Evidence will be formalized through a Clinical Knowledge Graph (CKG), which acts as the repository of structured knowledge. **Orchestrator Agent.** Coordinates the interaction, receives the initial clinical query and launches the deliberation. It manages the life cycle of the argumentative justification until a stable state of acceptability is reached. **Retrieval Agent.** Specializes in retrieving relevant evidence from clinical records and literature, and in its semantic standardization by building or mapping the extracted information to the CKG, which is essential to reduce ambiguity and ensure consistent, interoperable representations that support reliable reasoning and traceable links from conclusions back to sources. **Argument Construction Agent.** Generates claims, supporting arguments, and counterarguments by transforming structured facts from the knowledge graph into a first-pass argument graph (argumentation framework). **Evaluation and Verification Agent.** Applies the formal acceptability semantics of computational argumentation to determine which arguments are supported, and computes the proposed argument quality metrics.

3.2 Research Objectives and Plan

The thesis is structured around the following three objectives, which define the theoretical and practical contributions. **Objective 1: Definition of the Architecture and Formalisms.** Formally define the MAS architecture and the interaction specifications between agents (i.e., roles, message types, and coordination rules). This includes the design of an argument representation scheme adapted to the clinical domain, which will be compatible with and derivable from the Clinical Knowledge Graph (e.g., a simplified Toulmin or ASPIC$^+$ type scheme), and the selection of an acceptability semantics (for example, grounded or preferred) that guarantees stability and traceability of the justification. **Objective 2: Development of Metrics and Evaluation Protocol.** Propose a core set of metrics to evaluate reasoning quality (beyond textual quality), including acceptability, traceability and the system's ability to identify and address counterarguments, and and robustness to small changes in evidence. **Objective 3: Experimental and Comparative Validation.** Investigate and demonstrate the fundamental properties of the proposed system: that it terminates in a finite number of steps; the conditions under which it returns a recommendation with a warranted justification; and the conditions under which it cannot reach an acceptable conclusion (e.g., strongly contradictory evidence or lack of data).

ACKNOWLEDGMENTS

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

REFERENCES

[1] Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Simari, Matthias Thimm, and Serena Villata. 2017. Towards artificial argumentation. *AI magazine* 38, 3 (2017), 25–36.

[2] Cristian Cardellino, Theo Collias, Benjamin Molinet, Erwan Hain, Wei Sun, Rodrigo Agerri, Serena Villata, and Elena Cabrio. 2024. ANTIDOTE: ArgumeNtaTIon-Driven explainable artificial intelligence fOr digiTal mEdicine. In *ECAI 2024 (Frontiers in Artificial Intelligence and Applications, Vol. 392)*. 4455–4458. https://doi.org/10.3233/FAIA241028

[3] Luciano Caroprese, Eugenio Vocaturo, and Ester Zumpano. 2022. Argumentation approaches for explanaible AI in medical informatics. *Intelligent Systems with Applications* 16 (2022), 200109. https://doi.org/10.1016/j.iswa.2022.200109

[4] Paroma Chandak, Kevin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific Data* 10, 1 (2023), 67. https://doi.org/10.1038/s41597-023-01960-3

[5] Y.-J. Chen, A. Albarqawi, and C.-S. Chen. 2025. Reinforcing clinical decision support through multi-agent systems and ethical AI governance. arXiv preprint. https://doi.org/10.48550/arXiv.2504.03699

[6] Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence* 77, 2 (1995), 321–357. https://doi.org/10.1016/0004-3702(94)00041-X

[7] Shengxin Hong, Liang Xiao, Xin Zhang, and Jianxia Chen. 2024. ArgMed-Agents: Explainable Clinical Decision Reasoning with LLM Disscusion via Argumentation Schemes. arXiv:2403.06294 [cs.AI] https://arxiv.org/abs/2403.06294

[8] Pengshuai Jiang, Chao Xiao, Meng Jiang, Parth Bhatia, Taha Kass-Hout, Jimeng Sun, and Jiawei Han. 2024. Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. *arXiv preprint arXiv:2410.04585* (2024).

[9] M. Y. Landolsi, L. Hlaoua, and L. Ben Romdhane. 2023. Information Extraction from Electronic Medical Documents: State of the Art and Future Research Directions. *Knowledge and Information Systems* 65 (2023), 463–516. https://doi.org/10.1007/s10115-022-01779-1

[10] J. Li, Y. Lai, W. Li, J. Ren, M. Zhang, X. Kang, S. Wang, P. Li, Y.-Q. Zhang, W. Ma, and Y. Liu. 2025. Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents. arXiv preprint. https://arxiv.org/abs/2405.02957

[11] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:964287

[12] Luca Longo. 2013. Argumentation Theory for Decision Support in Health-Care: A Comparison with Machine Learning. In *Lecture Notes in Computer Science*. Springer. https://doi.org/10.1007/978-3-319-02753-1_17

[13] Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. arXiv:1705.07874 [cs.AI] https://arxiv.org/abs/1705.07874

[14] Sanjay Modgil and Henry Prakken. 2013. A general account of argumentation with preferences. *Artificial Intelligence* 195 (2013), 361–397. https://doi.org/10.1016/j.artint.2012.10.008

[15] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (Philadelphia, Pennsylvania) *(ACL '02)*. Association for Computational Linguistics, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv:1602.04938 [cs.LG] https://arxiv.org/abs/1602.04938

[17] Lucas Rizzo, Damiano Verda, Serena Berretta, and Luca Longo. 2024. A Novel Integration of Data-Driven Rule Generation and Computational Argumentation for Enhanced Explainable AI. *Machine Learning and Knowledge Extraction* 6, 3 (2024), 2049–2073. https://doi.org/10.3390/make6030101

[18] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2019. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* 128, 2 (Oct. 2019), 336–359. https://doi.org/10.1007/s11263-019-01228-7

[19] Y. Shen, J. Colloc, A. Jacquet-Andrieu, and K. Lei. 2015. Emerging Medical Informatics with Case-Based Reasoning for Aiding Clinical Decision in Multi-Agent System. *Journal of Biomedical Informatics* 56 (2015), 307–317. https://doi.org/10.1016/j.jbi.2015.06.012

[20] Nikolaos Stylianou and Ioannis Vlahavas. 2021. TransforMED: End-to-nd Transformers for Evidence-Based Medicine and Argument Mining in medical literature. *Journal of Biomedical Informatics* 117 (2021), 103767. https://doi.org/10.1016/j.jbi.2021.103767

[21] Jiayu Wu, Jing Zhu, Yuchen Qi, Jialong Chen, Min Xu, Filippo Menolascina, and Vicente Grau. 2024. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv preprint arXiv:2408.04187* (2024).

[22] L. Xiao and D. Greer. 2023. Linked Argumentation Graphs for Multidisciplinary Decision Support. *Healthcare (Basel)* 11, 4 (2023), 585. https://doi.org/10.3390/healthcare11040585

[23] Guangyi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*. 6233–6251.

[24] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. arXiv:1904.09675 [cs.CL] https://arxiv.org/abs/1904.09675

[25] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2015. Learning Deep Features for Discriminative Localization. arXiv:1512.04150 [cs.CV] https://arxiv.org/abs/1512.04150