SofIA: AI Clinical Companion for Real-Time Documentation and Decision Support

Demonstration Track

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

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

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

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

ABSTRACT

We present SofIA, a hospital-ready assistant that helps doctors with three everyday tasks: (1) summarising a patient's history, (2) answering clinical lookups from hospital guidelines, and (3) drafting structured clinical notes during or after the visit. The demo focuses on the user experience rather than algorithms. Users can try SofIA on a set of synthetic patients and see how summaries, answers with cited sources, and draft notes are produced in seconds. We will also show how SofIA connects to hospital systems and how safety checks keep humans in control.

KEYWORDS

Clinical Documentation; Decision Support; MAS; Healthcare

ACM Reference Format:
Leire Villarroya-Martínez, Enrique Alcazar Garzas, Stella Heras, Javier Palanca, and Vicent Botti. 2026. SofIA: AI Clinical Companion for Real-Time Documentation and Decision Support: Demonstration Track. 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/KJGN4402

1 INTRODUCTION

Doctors spend a significant portion of their time (over 50% of working hours) on documentation and interacting with Electronic Health Records (EHRs) [1]. We present SofIA, a hospital-ready assistant that helps with three everyday tasks: summarising patient histories, answering clinical lookups from guidelines, and drafting structured clinical notes. This demonstration illustrates SofIA's workflow as deployed within the hospital's EHR. The clinician selects a patient and can interact with SofIA in two modes: (1) Chat Interface for research queries (e.g., medication interactions and structured summary), which returns concise, sourced explanations; and (2) Transcription Mode (ambient scribing) for real-time audio capture, identifying clinical concepts and generating an editable, structured draft note (e.g., SOAP (Subjective, Objective, Assessment, Plan) or APSO (Assessment, Plan, Subjective, Objective)) [5, 8]. SofIA's design aligns with current agentic models for medical decision support, utilizing modular multi-agent architectures and Retrieval-Augmented Generation (RAG) frameworks to ensure provenance and auditability [3, 4, 6, 7]. Crucially, the system proposes, but never autonomously commits changes to the EHR, preserving human oversight at all stages. Users can try SofIA live on synthetic patients to see its evidence-linked reasoning and reviewable draft note generation.

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/KJGN4402

2 SOFIA TOOL

The core of SofIA's intelligence lies in its modular multi-agent architecture, designed to balance autonomy with clinical safety (see Figure 1). Its demo video is available here: https://youtu.be/6s80M2loYAw. Once a request is received via the Web Component, the Cognitive Framework acts as an orchestrator, dispatching sub-tasks to specialized agents: (1) the Notes Agent utilizes a structured pipeline to map clinical findings into standard SOAP/APSO templates; (2) the Coding Agent interfaces with the Coding REST API to ensure clinical terms align with hospital nomenclature; and (3) the Reviewer Agent performs a final consistency check, flagging unsupported claims (missing citations), cross-section inconsistencies, and stale evidence before presentation. This separation of concerns enables bounded agent objectives, iterative draft–review–refinement loops, and easier auditing of the decision-making process.

To ensure seamless integration, SofIA is built on top of standard healthcare APIs (e.g., HL7 FHIR) [2], allowing it to synchronize

in real-time with the Centar EHR/HIS. The data flow is strictly unidirectional for automated tasks: the system fetches patient resources—such as Observation for labs, Medication Statement for current drugs, and Condition for the problem list—to construct a context-aware prompt for an LLM-based Retrieval-Augmented Generation (RAG) engine. Write-back operations to the EHR are protected by a "Human-in-the-Loop" gatekeeper; only after the clinician's manual review and digital sign-off are the generated JSON artifacts committed to the permanent record. SofIA follows a simple pipeline designed for clarity and easy oversight: (1) Input selection: the clinician selects a patient (synthetic in the demo) and optionally a task (summary, Q&A, draft note). (2) Retrieval: relevant items are fetched from the record snapshot (problems, meds, allergies, labs, notes) and from guideline sources if needed. (3) Generation with guardrails: answers and drafts are produced with citations, safety checks, and style constraints. (4) Human review: the clinician edits/accepts; nothing is written back without explicit confirmation.

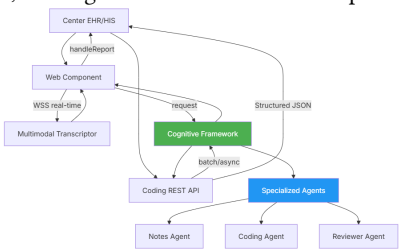

Figure 1: System overview for SofIA.

The live demo has three short, hands-on steps. Each step is designed to work offline with synthetic data if needed. **Step 1: Pick a Patient and Get a One-Page Summary (Figure 2).** The presenter selects a synthetic patient case. SofIA displays a concise summary (conditions, recent labs, medications, allergies). **Step 2: Ask a Clinical Question (Figure 2).** The user types a short question (e.g., "Is there a contraindication between Drug A and Drug B?" or "What was the last creatinine and when?"). SofIA returns an answer with short rationale and links to the underlying source (guideline or note). **Step 3: Draft a Note (Figure 3).** With one click, SofIA drafts a note using a hospital template (e.g., SOAP), ready for quick edits and sign-off by the clinician. A core feature of the SofIA demonstration is the Ambient Scribing mode, which automates clinical documentation during the patient encounter. The Multimodal Transcriptor captures real-time audio and generates a time-stamped transcript with speaker diarization. The Cognitive Framework then performs Clinical Entity Recognition (CER) over the transcript and extracts relevant findings, which are mapped by the Notes Agent into a structured SOAP or APSO draft.

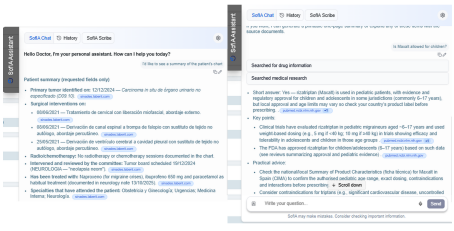

Figure 2: Patient Summary (conditions, labs...) and Q&A.

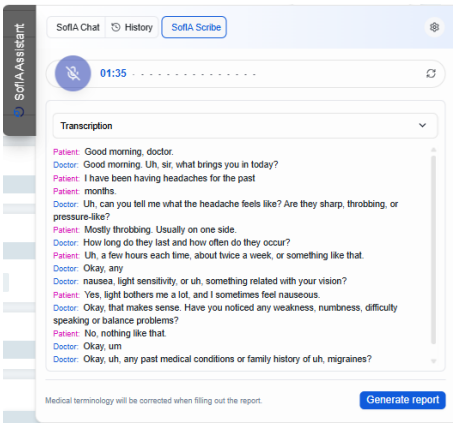

Figure 3: Illustrates real-time Multimodal Transcription.

How It Fits in a Hospital. We keep the explanation at a high level. SofIA is embedded as a web component next to the EHR screen. The data access to read patient data is performed to summarise and answer clinician questions; but SofIA writes only drafts on explicit approval. The interoperability is designed to connect via standard healthcare APIs used by common hospital systems. Regarding privacy & safety, SofIA allows for encryption in transit, strict access, and full human review before anything is saved.

3 EXPLAINABILITY AND HUMAN ALIGNMENT

SofIA aims to keep its reasoning legible to clinicians with lightweight, human-centered techniques. Explainability and transparency is achieved through explicit evidence linking and provenance. Unlike black-box models, SofIA provides Inline Citations: every claim in a patient summary or Q&A response is indexed to its source, whether it is a specific laboratory result or a paragraph from a clinical guideline. Furthermore, the system implements Uncertainty Cues; if the retrieved data is stale (e.g., lab results older than 6 months) or conflicting, the UI displays warning badges to prompt further investigation. For complex reasoning, users can trigger Contrastive Explanations, asking "Why not X?" to understand which clinical evidence would be required to support an alternative diagnosis or treatment plan.

4 FUTURE WORK

Future steps include prospective usability studies with timing and error-rate baselines as part of the ongoing hospital deployment, multi-lingual support (ES/EN), explainability UX A/B tests with clinicians, and options for on-prem/edge deployment. Systematic metric collection is currently in progress.

Ethics and Data. The demo uses only synthetic or anonymised patient data created for demonstration. No real patient data is processed.

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] 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

[2] Duane Bender and Kamran Sartipi. 2013. HL7 FHIR: An Agile and RESTful Approach to Healthcare Information Exchange. In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems (CBMS)*. 326–331. https://doi.org/10.1109/CBMS.2013.6627810

[3] Shuai Hong, Lin Xiao, Xuanjing Zhang, and Jialong Chen. 2024. ArgMed-Agents: Explainable clinical decision reasoning with LLM discussion via argumentation schemes. arXiv:2403.06294 [cs.AI] arXiv preprint.

[4] Jing Li et al. 2025. Agent Hospital: A simulacrum of hospital with evolvable medical agents. arXiv:2405.02957 [cs.AI] arXiv preprint.

[5] Chen-Tan Lin, Marlene McKenzie, Jonathan Pell, and Liron Caplan. 2013. Health Care Provider Satisfaction With a New Electronic Progress Note Format: SOAP vs APSO Format. *JAMA Internal Medicine* 173, 2 (01 2013), 160–162.

[6] Nikhil Mehandru et al. 2024. Evaluating large language models as agents in the clinic. *npj Digital Medicine* 7 (2024), 84.

[7] Yabin Shen, Joëlle Colloc, Anne Jacquet-Andrieu, and Keqiu 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.2015.06.005

[8] Lawrence L. Weed. 1968. Medical Records That Guide and Teach. *New England Journal of Medicine* 278, 11 (1968), 593–600. https://doi.org/10.1056/NEJM196803142781105