ReflectOR: an LLM-based Agent for Post-Operative Surgical Debriefing

Lorenzo Fumi[1], Marco Bombieri[1], Sara Allievi[2], Stefano Bonvini[2], Theodora Chaspari[3], Marco Zenati[4,5], Paolo Giorgini[1],

[1]Department of Information Engineering and Computer Science, University of Trento, Italy
[2] Azienda Provinciale per i Servizi Sanitari, Trento, Italy
[3] Institute of Cognitive Science and Department of Computer Science, University of Colorado, Boulder, USA
[4] Medical Robotics and Computer Assisted Surgery (MRCAS) Laboratory, Division of Cardiac Surgery, Veterans Affairs Boston Healthcare System, Boston, MA, USA
[5] Division of Cardiac Surgery, Brigham and Women's Hospital, Mass General Brigham, Harvard Medical School, Boston, MA, USA

***Correspondence:** marco.bombieri@unitn.it

Abstract

Ineffective teamwork and communication can generate medical errors in the high-pressure environment of surgery, making post-operative debriefings essential for enhancing team performance and patient safety. However, these sessions are frequently rushed or incomplete due to clinicians' limited time. This paper introduces ReflectOR, an Agentic-AI architecture designed to support surgical debriefings by processing audio recordings from the operating room. The system employs specialized sub-agents that perform tasks such as generating summaries, constructing timelines of intra-operative events, identifying potential errors, and counting the materials used. A qualitative evaluation indicates that the system effectively contextualizes transcripts, demonstrating its potential as a valuable tool for surgical debriefing. The paper also outlines key considerations for applying such an architecture in real-world clinical environments.

1 Introducion

In the high-stakes environment of surgery, ineffective teamwork and communication represent significant risk factors that can lead to medical errors. Debriefings are widely recognized as a critical tool for improving team performance, communication, and collaboration (Phrampus and O'Donnell, 2013; Endacott et al., 2018). In the case of surgery, post-operative debriefing involves a structured discussion among surgical team members, aimed at reviewing the procedure, identifying errors, discussing successes, and addressing any incidents. However, these sessions are often brief, informal, or incomplete, mainly due to the limited time available, physical tiredness, psychological stress experienced by clinicians immediately after surgery, and increased workload (Arriaga et al., 2021b). When these sessions are rushed or rely solely on human memory, critical information may be overlooked, and documentation errors may occur, potentially increasing the risk of future adverse events (Arriaga et al., 2021a). Other researches highlight that inadequate documentation of patient information is among the most common sources of clinical communication failures (Alder, 2025).

For this reason, it is crucial to support clinicians during this phase with automatic tools that enable them to conduct debriefings efficiently and effectively. Large Language Models (LLMs) have recently reshaped natural language processing (NLP), achieving near-human-level performance on various benchmarks (e.g., (Hendrycks et al., 2021; Chiang et al., 2024)) with little or no task-specific tuning. At the same time, they have become increasingly widespread in the surgical domain, where they are employed for analyzing surgical procedures and optimizing workflows (Bombieri et al., 2024b,a; Pressman et al., 2024). Moreover, techniques such as Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) are being integrated into LLMs to provide access to up-to-date knowledge bases—an essential feature in specialized domains such as surgery (Ke et al., 2025). While an LLM equipped with a RAG system can be helpful for a specific task, developing a fully functional automatic debriefing system (e.g., capable of summarizing the procedure, generating an event graph, listing errors and unexpected events, and estimating costs) requires an agent-based architecture composed of multiple LLMs, each specialized in a specific task. To the best of our knowledge, such an architecture is not yet available. Furthermore, existing debriefing systems are developed for English data: this linguistic bias poses additional challenges for adapting such models to multilin-

gual or non-English clinical environments, where linguistic and cultural variations can significantly impact system performance and usability.

This paper addresses this research gap by presenting the design, implementation, and qualitative evaluation of an agentic system for post-operative surgical debriefing based on LLMs, named *ReflectOR*. The proposed system is capable of transcribing dialogue recordings from the operating room and assisting surgeons during post-operative debriefings across multiple tasks. The system is qualitatively evaluated on *Italian* dialogues recorded during a simulated EVAR (Endovascular Aneurysm Repair) procedure. The evaluation focuses on the accuracy of the input processing pipeline, which converts surgical audio into a diarized and timestamped transcription, assessing both transcription quality and speaker diarization performance. Furthermore, we qualitatively examine several functionalities of the agentic system to illustrate its capabilities and to discuss its current limitations.

In more detail, the paper aims to investigate the following Research Question (RQ):

> Can existing transcription and diarization models achieve satisfactory performance on Italian surgical audio recordings, which may include significant background noise, when applied in a zero-shot setting without any domain-specific fine-tuning? What are the key challenges to be addressed to deploy an agent capable of performing debriefing from intraoperative dialogues in real-world surgical environments?

In addressing this RQ, this paper makes the following contributions:

C1: We implement a prototype of an LLM-based agent designed to assist clinicians during the post-operative debriefing process. The agent leverages the diarized transcripts as input to identify relevant events, summarize interactions, and support reflective discussions. By doing so, we quantitatively benchmark transcription and diarization techniques on a manually annotated dataset.

C2: We conduct a preliminary qualitative evaluation by presenting a demonstration of the prototype to a multidisciplinary team of clinicians and engineers. We collect and analyze their feedback to discuss the perceived challenges and opportunities in deploying such a system within real-world clinical environments.

2 Related Work

In recent years, intraoperative debriefing has received increasing attention, as it plays a crucial role in identifying common errors and, consequently, in reducing the incidence of adverse events (Arriaga et al., 2021b). Traditionally, these debriefings relied solely on the recollection of clinicians present in the operating room. Previous research (Loukissas et al., 2012) has emphasized the value of data-driven postoperative reviews that integrate multimodal perioperative information to support more effective reflection and learning. Similarly, there is a growing demand among practitioners for the integration of Automatic Speech Recognition (ASR) systems in surgical environments, which would enable accurate and objective documentation of intraoperative events (Schulte et al., 2020).

For this reason, artificial intelligence techniques are being increasingly applied to this task. For instance, (Hong et al., 2025) evaluates the capability of GPT-4o to summarize transcripts from simulated surgical procedures, with qualitative assessments collected through satisfaction questionnaires administered to a team of clinicians involved in the experiment, who reported a high level of approval. Similarly, (Fuchtmann et al., 2024) proposes a Convolutional Neural Network (CNN)-based pipeline for reconstructing intraoperative events.

At the same time, to achieve an effective and high-quality automatic debriefing system, (Klusty et al., 2025) emphasizes the importance of investing in speech-to-text technologies: only from accurate transcriptions and diarizations can AI-based debriefing tools yield meaningful benefits. Consequently, ongoing research focuses on improving transcription quality even in potentially noisy environments, both in general domains (Hong et al., 2025) and in the medical domain (Zhang et al., 2023; Li and Mu, 2024; Ng et al., 2025), where state-of-the-art techniques may still struggle to accurately recognize medical terminology, especially in less-represented accents or languages (Li and Mu, 2024).

Our work differs from these approaches in that it focuses explicitly on evaluating the performance of transcription and diarization tools in the surgical domain and in a language other than English. Moreover, we release a prototype agent system

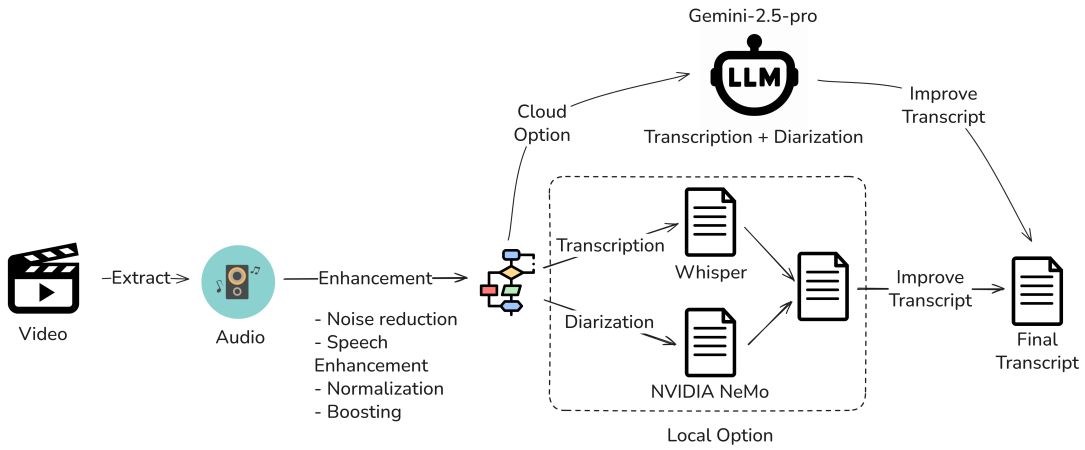

Figure 1: Audio processing pipeline.

capable of managing in an organized manner the various subtasks typical of intraoperative debriefing, extending beyond simple summarization or event detection.

3 Methodology and System Architecture

Figures 1 and 2 illustrate the ReflectOR system, including its input-processing pipeline and agentic architecture for surgical debriefing, respectively.

In the input processing architecture shown in Figure 1, the audio track is extracted from the surgical video recorded in the operating room and preprocessed to reduce background noise. The audio is then automatically transcribed and diarized using a speech-to-text model. We tested several approaches for this step, including local models such as Nvidia NeMo and Whisper, as well as a cloud-based service (i.e., GEMINI-2.5-PRO). The resulting diarized transcription serves as input to the *Coordinator* component of the agentic architecture depicted in Figure 2. The Coordinator is a high-level autonomous agent responsible for interpreting clinicians' requests, expressed in natural language through the graphical user interface (GUI)[1], and delegating them to the appropriate specialized sub-agent. Upon receiving input from the clinician, the Coordinator leverages natural language understanding capabilities to determine the intent of the request and orchestrates the execution of downstream analytical tasks accordingly. Depending on the clinician's needs, it can, for instance, invoke a sub-agent to generate concise summaries of the transcriptions, another to build a detailed timeline of intraoperative events, one to identify potential errors discussed by the surgical staff, or another to track the materials and instruments used during the procedure. Finally, the clinician can request the generation of a structured PDF report summarizing the entire surgical intervention, which the Coordinator produces by collecting the output from relevant sub-agents.

Section 3.1 details the audio-processing pipeline, while Section 3.2 will provide more details about the agentic architecture.

3.1 Audio processing pipeline

Audio pre-processing techniques In surgical environments, audio recordings are often affected by suboptimal microphone quality and overlapping speech among team members, particularly during critical moments. Consequently, the recorded audio required extensive preprocessing to reduce background noise and improve overall speech intelligibility. To address this issue, we used the following preprocessing pipeline. The process began with amplifying the entire audio track to ensure that low-volume speech segments were adequately captured. DeepFilterNet (Schröter et al., 2022) (v0.5.6), a deep learning–based speech enhancement system, was then applied to suppress background noise. Next, volume normalization was performed to balance loudness levels across speakers, preventing quieter voices from being masked. SpeechBrain (Ravanelli et al., 2021) (v1.0.3) was subsequently used to enhance speech quality further. Finally, Demucs (Rouard et al., 2023) was employed to isolate the vocal components from any remaining background noise, ensuring that the resulting audio contained only the spoken content.

[1]Demonstration videos and figures of the GUI are provided in the external repository.

Transcription and diarization techniques For the automatic transcription of recorded audio, we tested WHISPER-LARGE-V3[2]. For the automatic diarization, we tested NVIDIA NEMO[3] and PYANNOTE-AUDIO[4]. We also tested Google Gemini's models to perform transcription and diarization in an end-to-end fashion. In particular, we used GEMINI-2.5-FLASH and GEMINI-2.5-PRO[5] with the prompt template reported in Table 1, which yielded the highest performance among the tested ones, with the temperature set to the default.

3.2 Agentic-AI architecture

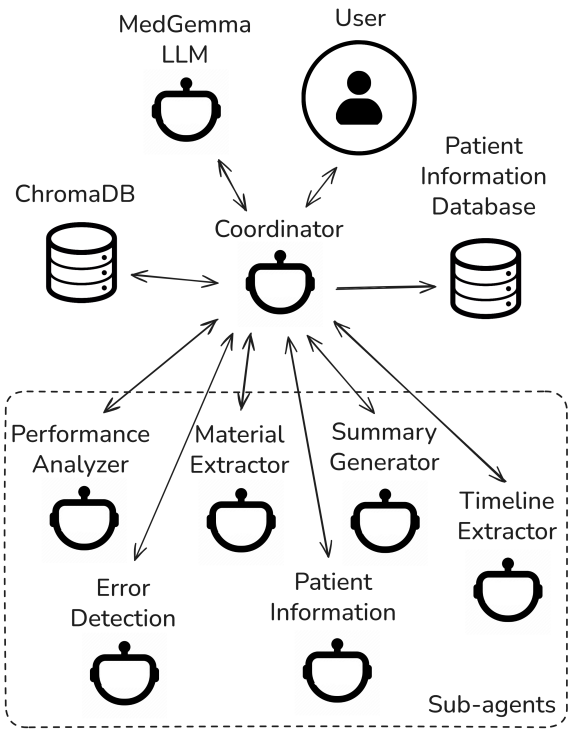

Figure 2: Agentic-AI architecture: Coordinator and its interactions with other system components.

The agentic-AI architecture is composed of a central *Coordinator* and specialized sub-agents, each designed to perform a specific analytical or reasoning task within the surgical analysis pipeline. The *Coordinator* receives high-level instructions from clinicians through a GUI. Based on the user's request, it determines which sub-agent(s) to invoke and whether additional patient information needs to be retrieved from the institutional database (for example, when a clinician requests access to a patient's electronic health record). Specifically, the architecture includes the following sub-agents:

- *Patient Information sub-agent:* reads from an external database and returns the information about the patient, together with the planned interventions and corresponding surgical team.
- *Material Extractor sub-agent:* identifies and categorizes the surgical materials and instruments named (and thus used) during the operation, estimating total operational costs based on standardized pricing data saved on an external database.
- *Timeline Extractor sub-agent:* detects and chronologically orders from the transcriptions the key intra-operative events, generating a temporal map of the surgical process.
- *Error Detection sub-agent:* identifies deviations from established protocols, teamwork inefficiencies, and other potential sources of human or procedural error.
- *Summary Generator sub-agent:* produces a concise textual summary of the surgery, the main actions, incidents, and the overall outcome.
- *Performance Analyzer sub-agent:* evaluates the performance of the surgical team over multiple recorded operations, identifying areas of improvement and strengths.

Furthermore, the *Coordinator* can query an external *Medical LLM* (MedGemma LLM in the Figure) to retrieve general medical information relevant to the surgical procedure when requested by the clinician.

After the debriefing, a PDF is automatically generated with a structured agenda based on answers from the other sub-agents, supporting post-operative reviews.

Both the *Coordinator* and the sub-agents are implemented as LLMs, each configured with a role-specific *system prompt* defining its goal and expected behavior. The *Coordinator* and the sub-agents are standard GPT-4O models, while the external Medical LLM is MEDGEMMA (Sellergren et al., 2025). The *Coordinator* and the Medical LLM are pre-configured with a temperature of 0.7. Conversely, the other sub-agents operate with a temperature of 0.0, reflecting their deterministic and task-oriented behavior. Detailed descriptions of the system prompts used for each LLM-based agent are available as supplementary material in the paper's repository. The agentic framework is

[2]openai/whisper-large-v3 with WhisperX back-end
[3]Nvidia NeMo: nvidia/diar_sortformer_4spk-v2
[4]Pyannote: pyannote/speaker-diarization-3.1
[5]We used the models updated on June 17, 2025

Generate a transcription of the surgical operation received via audio in Italian. Include timestamps and identify the speakers. Do not invent information — use only what is present in the audio. The speakers are: [List of speakers]
It is important to include the correct names of the speakers. Do not use any markdown formatting, such as bold or italics. Use only characters from the Italian alphabet, unless you truly believe that foreign characters are correct. It is important to use the correct words and ensure proper spelling throughout.
Important:
- Do not loop or repeat the same sentence multiple times.
- Write only in Italian. Do not use English, except for universally accepted words such as "software".
- If there is silence or background noise, do not write anything.
An example with the desired template follows:
`[00:00] Lorenzo: Hi.`
`[00:02] Alessandro: Hi Lorenzo.`

Table 1: Prompt used for transcription and diarization with GEMINI-2.5-PRO.

implemented within LangGraph.[6]

3.3 Preliminary evaluation

Data gathering and annotation. To benchmark transcription and diarization methods discussed in Section 3.1 in the surgical domain and to implement a preliminary demo of the architecture described in Section 3.2, we conducted a simulated endovascular aneurysm repair (EVAR) procedure involving one male surgery professor acting as the surgeon and two first-year medical students as surgical assistants, one male and one female. The team communicated in Italian. The simulation took place in the Laboratory of Augmented Health Environments at the University of Trento, Italy, and was recorded using a camera. In total, 52 minutes of video were collected. The audio was extracted from the recordings in `.wav` format, and the first 5 minutes were manually transcribed and diarized.

Evaluation of the audio processing pipeline. To assess the performance and reliability of both the transcription and diarization processes and thus to answer the first part of our RQ, we conducted a comparative evaluation between the manually annotated ground truth and the automatically generated outputs obtained using the models described in Section 3.1. The evaluation was carried out by considering the following standard quantitative metrics commonly employed in speech processing:

- Word Error Rate (WER) measures the proportion of errors made by a transcription system compared to a reference (ground truth) transcription. It is computed using the Levenshtein distance (Levenshtein, 1966), which counts the minimum number of substitutions (S), insertions (I), and deletions (D) required to transform the system's output into the reference text of length N words:

$$WER = \frac{S + D + I}{N}$$

- Diarization Error Rate (DER) quantifies how accurately the automatic system can determine *who spoke when* in an audio recording. It represents the fraction of time that is incorrectly attributed to a speaker and is defined as:

$$DER = \frac{T_{error}}{T_{total}}$$

where T_{error} is the total duration of speaker-attributed errors, and T_{total} is the total reference speech time.

Preliminary evaluation of the architecture. To address the second part of the RQ and to foster discussion regarding the feasibility of employing an LLM-based agent for debriefing dialogues recorded in the operating room, we implemented the architecture described in Section 3.2. Using the transcriptions and diarization outputs obtained from the 52 minutes of audio extracted from surgical videos, we developed an interactive demonstration of the system. This prototype was presented to a multidisciplinary group of subject matter experts (SMEs), comprising three clinicians and three engineers, as both clinical and technical perspectives are essential to ensure that the system's functionalities align with real-world surgical workflows and technological feasibility. The demonstration aimed to explore the system's capabilities and limitations in interpreting and summarizing intraoperative communication, as well as to assess its potential integration into clinical workflows.

4 Results and discussions

Regarding the audio processing pipeline. Table 2 reports the performance of the models evaluated on Italian surgical audio recordings in a zero-shot setting, i.e., without any domain-specific fine-tuning.

For transcription, GEMINI-2.5-PRO achieved the lowest WER (10.72%), outperforming all other models, including GEMINI-2.5-FLASH (20.58%),

[6] LangGraph: https://www.langchain.com/langgraph [Last access: 2025-10-10]

Transcription	
Model	**WER (%)**
GEMINI-2.5-PRO	**10.72**
GEMINI-2.5-FLASH	20.58
WHISPER-LARGE-V3	41.16
Diarization	
Model	**DER (%)**
GEMINI-2.5-PRO	**10.89**
GEMINI-2.5-FLASH	11.53
NVIDIA NEMO	26.42
PYANNOTE-AUDIO	14.84

Table 2: Performance of automatic transcription and diarization methods. Transcription results are evaluated using WER, and diarization results using DER.

and WHISPER-LARGE-V3 (41.16%). These results show that GEMINI-2.5-PRO seems to be more robust to domain-specific terminology typical of surgical environments, even without task-specific adaptation.

Regarding diarization, a similar trend was observed. GEMINI-2.5-PRO again achieved the best performance, with a DER of 10.89%, followed by GEMINI-2.5-FLASH (11.53%) and PYANNOTE-AUDIO (14.84%). In contrast, NVIDIA NEMO performed worse in our recordings, with a DER value of 26.42%. These results suggest that large multimodal LLM-based architectures, such as Gemini, can effectively leverage contextual understanding and cross-modal representations to improve speaker segmentation and attribution, even in acoustically challenging clinical settings.

Overall, the experimental findings suggest that GEMINI-2.5-PRO can achieve satisfactory performance on Italian surgical recordings in a zero-shot scenario. Some degradation in accuracy remains inevitable due to environmental noise, overlapping speech, and domain-specific language.

Regarding the agentic architecture We implemented the architecture described in Section 3.2, which was fed with the transcriptions and diarization files discussed in Section 3.3.

Tables 3, 4, 5, and 6 contain examples of the output of the implemented architecture when the *Timeline Extractor*, *Error Detection*, *Material Extractor*, and *Summary Generator* sub-agents are called by the *Coordinator*, respectively. The complete output generated by the individual sub-agents is available in the repository together with some illustrative videos. Table 3 illustrates an extract of the *Timeline Extractor* output, which timestamps and organizes individual events during the surgical discussion. This component enables easy tracking of procedural steps and the flow of conversation, showing the model's capacity to segment and contextualize temporal information from the transcription. The SMEs reviewed the full output from this component and found it generally clear, while some steps may appear simplified or slightly misaligned, likely due to transcription limitations. Table 4 presents examples from the *Error Detection* sub-agent, which identifies potential mistakes or misconceptions by participants, such as incorrect measurements or procedural misunderstandings. The SMEs found the output effective, although they noted that the model occasionally produces false positives, which could trigger unnecessary concern. As a potential improvement, they suggested correlating the output more directly with past reports to provide a historical record of common errors for each procedural phase. Table 5 reports an extract of the *Material Extractor* output, showing how the agent estimates usage times and costs of surgical materials. This demonstrates the sub-agent's ability to automatically quantify procedural resources, potentially aiding in inventory management, cost tracking, and the preparation of materials in advance for future procedures. The SMEs found the output useful and clear. As a suggested improvement, they proposed dividing costs between consumable materials and drugs, and adding functionality to provide recommendations for optimizing the use of materials and

Time	Event
00:20	Identification of the clinical case: infrarenal abdominal aortic aneurysm to be treated with EVAR.
00:34	Analysis of the aneurysm neck: confirmation that it is long and standard for the procedure.
01:10	Confirmation that the iliac arteries are not involved in the aneurysm.
01:36	Discussion of the neck measurements: 52 mm, compatible with multiple stent grafts.
02:19	Identification of the preoperative projection: 15° LAO and 8° caudal.
...	...

Table 3: Example of the output from the *Timeline Extractor* sub-agent (translation from Italian).

Time	Error
01:36	Student 1 reports that the neck is 52 millimeters, apparently without having carefully verified the measurement on the CT scan or the available data, leading to a potential inaccuracy in the case preparation.
03:28	Student 1 states that the best side for inserting the stent graft is the right one due to lower tortuosity, but this statement is contradicted by the professor, who emphasizes the permissive anatomy.
05:03	Student 1 initially fails to recognize that heparin administration should be performed before proceeding with further interventions. He must be corrected by the professor in order to continue properly.
...	...

Table 4: Example of the output from the *Error Detection* sub-agent (translation from Italian).

Name	Used at	Cost (€)
Soft guide	05:03	123
Rigid guide	08:00	246
Pigtail catheter	12:05	321
...	...	...

Table 5: Example of the output from the *Material Extractor* sub-agent (translation from Italian).

medications. Table 6 illustrates the *Summary Generator* output, which condenses the transcript into a structured summary capturing key clinical observations, planning decisions, and measurements. The example highlights the agent's ability to preserve critical clinical details while producing a concise output. Nevertheless, discussions with SMEs highlighted that overall system performance is highly dependent on audio quality, which can be challenging in real-world, high-stakes operating room environments. For instance, background noise between 0:44 and 0:50 rendered the dialogue nearly imperceptible, resulting in missing segments in the generated reports.

Collectively, these examples indicate that the sub-agents (instances of GPT-4o in our demo) demonstrate strong capabilities in understanding and contextualizing surgical transcriptions and diarizations, suggesting potential utility in supporting post-operative debriefings. Remaining inaccuracies stem from transcription limitations and occasional oversimplifications by the agents. The former could be alleviated by exploring more effective audio-processing techniques, while the latter could be mitigated by refining sub-agent prompts, providing few-shot examples curated by clinical experts, or incorporating chain-of-thought reasoning to handle complex clinical discussions better.

Discussions with the SMEs also highlighted several ethical and practical challenges that need to be carefully addressed before clinical deployment. First, excessive reliance on the system could lead to complacency, with clinicians paying less attention during procedures in the expectation that the system will handle the documentation and data processing tasks (Goddard et al., 2011; Quinn et al., 2020). Second, the use of real patient data requires robust data protection measures, including advanced anonymization, encryption both at rest and in transit, and strong authentication and authorization mechanisms. Furthermore, dependence on external cloud-based LLM services raises concerns related to privacy, security, and service continuity (Dennstädt et al., 2025). While deploying local models and servers could mitigate these risks, such an approach entails high computational costs and often results in lower performance and accuracy compared to state-of-the-art cloud solutions. Third, systems of this nature are susceptible to jailbreaking attacks (Han et al., 2024), which in some cases can be executed with minimal effort or deceptively simple prompts (Bombieri et al., 2025), highlighting the critical need for robust and comprehensive security measures.

5 Conclusions

This paper presents ReflectOR, an Agentic-AI architecture designed to support debriefing dialogues captured in operating rooms. Qualitative analyses suggest that the system has the potential to effectively summarize, contextualize, and chronologically extract intraoperative events—capabilities strongly desired by clinicians (Loukissas et al., 2012; Schulte et al., 2020). However, these results are dependent on the quality of the audio recordings, and for some sub-agents, further fine-tuning is needed. Finally, this paper summarizes some ethical and practical challenges that must be addressed prior to clinical deployment, including the risk of excessive reliance on the system, the need

Summary
The transcript documents the discussion and simulation of an endovascular (EVAR) procedure for the repair of a 44 mm infrarenal abdominal aortic aneurysm. The case is analyzed in detail by participants, including a professor and several students, through CT evaluation and preoperative surgical planning. Below is the full summary: • **Aneurysm:** Located at the level of the infrarenal abdominal aorta, with a long neck and no involvement of the iliac arteries. No significant circumferential calcifications are present. • **Planning:** An aorto-bisiliac endovascular intervention is decided. The most appropriate operative projection indicated is 15° LAO and 8° caudal. The renal arteries are at different levels, with the left lower, but do not require an aggressive approach. • **Measurements:** Neck of 52 mm; stent graft to be applied with a diameter of 26 mm (15% oversizing) and planned lengths of about 138 mm on the right and 134 mm on the left. • ...

Table 6: Example of the output from the *Summary Generator* sub-agent (translation from Italian).

for data protection measures, the costs associated with using local models to preserve privacy, and the risk of jailbreaking attacks.

6 Limitations and Future Works

This paper represents a first step towards the implementation of an agentic AI system for post-operative surgical debriefing that can be effectively adopted in Italian operating rooms. Nevertheless, some limitations should be acknowledged and addressed in future work. First, our benchmarking experiments for transcription and diarization were conducted on a 5-minute portion of surgical dialogue corresponding to the initial phase of the procedure. While this sample provides an initial and controlled setup to compare transcription and diarization tools, it may introduce a selection bias, as the linguistic and acoustic characteristics of the early stage might not fully represent the entire surgical workflow. A more extensive evaluation on longer and more heterogeneous recordings would therefore be necessary to obtain generalizable results. Second, the audio recordings used to evaluate the agentic AI architecture were collected from a single operating room in a controlled setting. This inevitably reduces the diversity in acoustic conditions, surgical teams, and procedural variability. Future research will therefore focus on data collection from multiple hospitals and surgical specialties to assess the scalability, robustness, and adaptability of the proposed approach across different clinical environments. Third, the present study provides a qualitative assessment limited to the speech-to-text transcription and speaker diarization components of the proposed system, while the AI architecture as a whole is discussed only at a conceptual level. A comprehensive evaluation of the architecture would require more diverse and extensively annotated datasets—resources that are currently difficult to obtain—as well as dedicated user studies. Future work will therefore focus on expanding the dataset to enable a more detailed and quantitative evaluation of all system components. Fourth, for computational reasons, the agents in this study were implemented using API-based LLMs. While this approach is acceptable for demos and experimental setups, it is not suitable for real clinical contexts because APIs require that data be uploaded to the cloud, potentially leading to privacy and security concerns. Future implementations in realistic clinical environments should therefore consider using locally hosted LLMs, accepting the associated computational and performance limitations.

Nevertheless, we emphasize that the primary aim of this paper is to present the idea and discuss its potential limitations and future research directions.

Data availability

The code used to run these experiments, and some videos and images illustrating the system functionalities, are available at: https://github.com/DeeJack/ReflectOR-Data

Acknowledgments

Paolo Giorgini and Marco Bombieri are supported by the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by NextGenerationEU, and AI@TN2.0 project funded by the autonomous province of Trento.

References

Steve Alder. 2025. Effects of poor communication in healthcare. The HIPAA Journal.

AF Arriaga, YK Chen, MPT Pimentel, AM Bader, and D Szyld. 2021a. Critical event debriefing: a checklist for the aftermath. *Current Opinion in Anesthesiology*, 34(6):744–751.

Alexander Arriaga, Yun-Yun Chen, Marc Pimentel, Angela Bader, and Demian Szyld. 2021b. Critical event debriefing: a checklist for the aftermath. *Current Opinion in Anaesthesiology*, Publish Ahead of Print.

Marco Bombieri, Simone Paolo Ponzetto, and Marco Rospocher. 2025. The dangerous effects of a frustratingly easy llms jailbreak attack. *IEEE Access*, 13:126418–126431.

Marco Bombieri, Marco Rospocher, Simone Paolo Ponzetto, and Paolo Fiorini. 2024a. The robotic-surgery propositional bank. *Lang. Resour. Evaluation*, 58(3):1043–1071.

Marco Bombieri, Marco Rospocher, Simone Paolo Ponzetto, and Paolo Fiorini. 2024b. Surgicberta: a pre-trained language model for procedural surgical language. *Int. J. Data Sci. Anal.*, 18(1):69–81.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Fabio Dennstädt, Janna Hastings, Paul Martin Putora, Max Schmerder, and Nikola Cihoric. 2025. Implementing large language models in healthcare while balancing control, collaboration, costs and security. *npj Digital Medicine*, 8:143.

Ruth Endacott, Tracey Gale, Anne O'Connor, and Sarah Dix. 2018. Frameworks and quality measures used for debriefing in team-based simulation: a systematic review. *BMJ Simulation and Technology Enhanced Learning*, 5(2):61–72.

Jonas Fuchtmann, Thomas Riedel, Maximilian Berlet, Alissa Jell, Luca Wegener, Lars Wagner, Simone Graf, Dirk Wilhelm, and Daniel Ostler-Mildner. 2024. Audio-based event detection in the operating room. *Int. J. Comput. Assist. Radiol. Surg.*, 19(12):2381–2387.

Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2011. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127.

Tianyu Han, Sven Nebelung, Firas Khader, Tianci Wang, Gustav Müller-Franzes, Christiane Kuhl, Sebastian Försch, Jens Kleesiek, Christoph Haarburger, Keno K. Bressem, Jakob Nikolas Kather, and Daniel Truhn. 2024. Medical large language models are susceptible to targeted misinformation attacks. *npj Digital Medicine*, 7:288.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Eury M. A. Hong, Sundes Kazmir, Benjamin Dylik, Marc Auerbach, Matteo Rosati, Sofia Athanasopoulou, Russell Himmelstein, Travis M. Whitfill, Lindsay Johnston, Traci A. Wolbrink, Arielle Shibi Rosen, and Isabel T. Gross. 2025. Exploring the use of a large language model in simulation debriefing: An observational simulation-based pilot study. *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare*. Published online May 13, 2025.

Yuhe Ke, Liyuan Jin, Kabilan Elangovan, Hairil Rizal Abdullah, Nan Liu, Alex Tiong Heng Sia, Chai Rick Soh, Joshua Yi Min Tung, Jasmine Chiat Ling Ong, Chang Fu Kuo, Shao-Chun Wu, Vesela P. Kovacheva, and Daniel Shu Wei Ting. 2025. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *npj Digit. Medicine*, 8(1).

Mitchell A Klusty, W Vaiden Logan, Samuel E Armstrong, Aaron D Mullen, Caroline N Leach, Ken Calvert, Jeff Talbert, and V K Cody Bumgardner. 2025. Toward automated clinical transcriptions. *AMIA J. Summits Transl. Sci. Proc.*, 2025:235–241.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jiafeng Li and Yanda Mu. 2024. Searching for best practices in medical transcription with large language model. *Preprint*, arXiv:2410.03797.

Yanni A. Loukissas, Jason K. Maron, Marco A. Zenati, and David Mindell. 2012. Redesigning postoperative review. In *Proceedings of the 1st Annual IEEE Healthcare Innovation Conference (IEEE EMBS)*, Houston, Texas, USA. IEEE.

Joel Jia Wei Ng, Eugene Wang, Xinyan Zhou, Kevin Xiang Zhou, Charlene Xing Le Goh, Gabriel

Zheng Ning Sim, Hiang Khoon Tan, Serene Si Ning Goh, and Qin Xiang Ng. 2025. Evaluating the performance of artificial intelligence-based speech recognition for clinical documentation: a systematic review. *BMC Medical Informatics and Decision Making*, 25(236).

Paul E. Phrampus and John M. O'Donnell. 2013. *Debriefing Using a Structured and Supported Approach*, pages 73–84. Springer New York, New York, NY.

Sophia M Pressman, Sahar Borna, Cesar A Gomez-Cabello, Syed Ali Haider, Clifton R Haider, and Antonio Jorge Forte. 2024. Clinical and surgical applications of large language models: A systematic review. *Journal of Clinical Medicine*, 13(11).

Thomas P Quinn, Manisha Senadeera, Stephan Jacobs, Simon Coghlan, and Vuong Le. 2020. Trust and medical ai: the challenges we face and the expertise needed to overcome them. *Journal of the American Medical Informatics Association*, 28(4):890–894.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A general-purpose speech toolkit. *Preprint*, arXiv:2106.04624. ArXiv:2106.04624.

Simon Rouard, Francisco Massa, and Alexandre Défossez. 2023. Hybrid transformers for music source separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Hendrik Schröter, Alberto N. Escalante-B., Tobias Rosenkranz, and Andreas Maier. 2022. DeepFilterNet: A low complexity speech enhancement framework for full-band audio based on deep filtering. In *ICASSP 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Antonia Schulte, Rodrigo Suarez-Ibarrola, Daniel Wegen, Philippe-Fabian Pohlmann, Elina Petersen, and Arkadiusz Miernik. 2020. Automatic speech recognition in the operating room – an essential contemporary tool or a redundant gadget? a survey evaluation among physicians in form of a qualitative study. *Annals of Medicine and Surgery*, 59:81–85.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. 2025. Medgemma technical report. *Preprint*, arXiv:2507.05201.

Jun Zhang, Jingyue Wu, Yiyi Qiu, Aiguo Song, Weifeng Li, Xin Li, and Yecheng Liu. 2023. Intelligent speech technologies for transcription, disease diagnosis, and medical equipment interactive control in smart hospitals: A review. *Computers in Biology and Medicine*, 153:106517.