ORCA: AI Documentation for the Operating Room

Team #26 ORAI SOLUTIONS®

Team Members: Luisalberto Acosta, Kaden Kobashigawa, Andrew Opstad, Andres Valdes

Mentors: Bradley Greger, PhD & Trejas Shah, MD

School of Biological and Health Systems Engineering, Arizona State University

Ira A. Fulton Schools of Engineering
Arizona State University
School of Biological and Health Systems Engineering

Background

Electronic health record (EHR) documentation occupies a significant portion of physicians' workload, often leading to burnout [1][2]. Automating intraoperative and postoperative documentation offers an opportunity to reduce this burden and return valuable clinical time to surgeons. ORAI Solutions™ proposes the Operating Room Charting Assistant (ORCA), an AI-powered system that isolates a healthcare professional's speech and uses integrated speech recognition and large language models (LLMs) to automatically generate surgical documentation.

Mission Statement

At ORAI Solutions™ we empower surgeons with reliable voice-powered documentation, so they can focus on patients, not paperwork.

Product Specifications

Customer Need	Metric
Report Accuracy: ASR Accuracy	**96.5%** @ 6% White Noise
Report Accuracy: LLM Accuracy	**91.5%** Extraction Accuracy
HIPPA Compliance	0 Violations
Documentation Time Reduction	~70% Reduction of Total Time
Cost-Efficiency	~$6,000 Annual Cost: 4x ROI

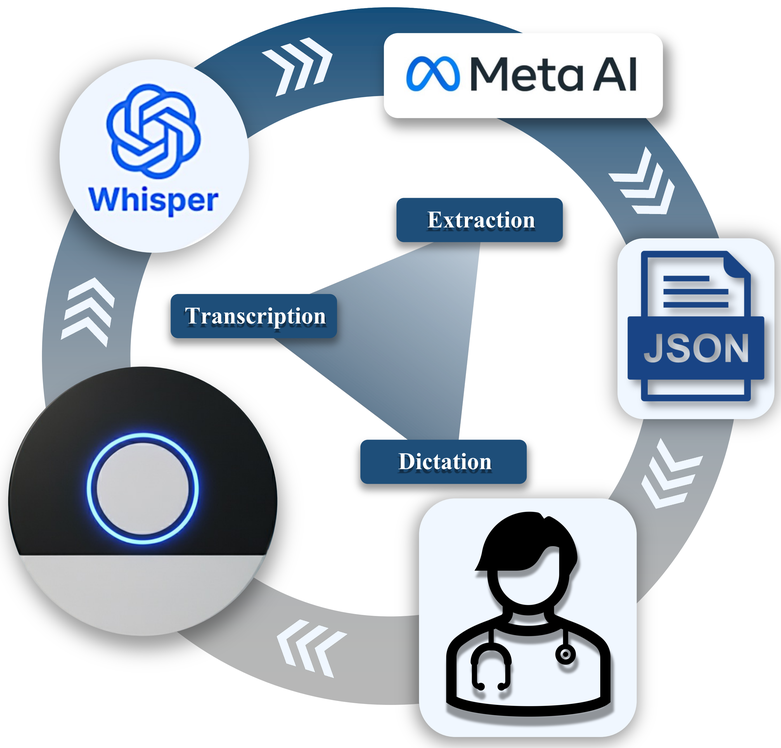
Table 1. **Product Specifications.**

Predicted Savings

Assumption	Saved
1 device per surgeon	-
~ 7 mins of documentation time saved per operation	7 mins/op.
~ 4 general surgical operations daily [3]	28 mins/day
~ 240 workdays per year	112 hours/year
~ $210 hourly wage [4]	**$23,520/year**

Table 2. **Predicted Savings.**

Prototype



Final Technical Model

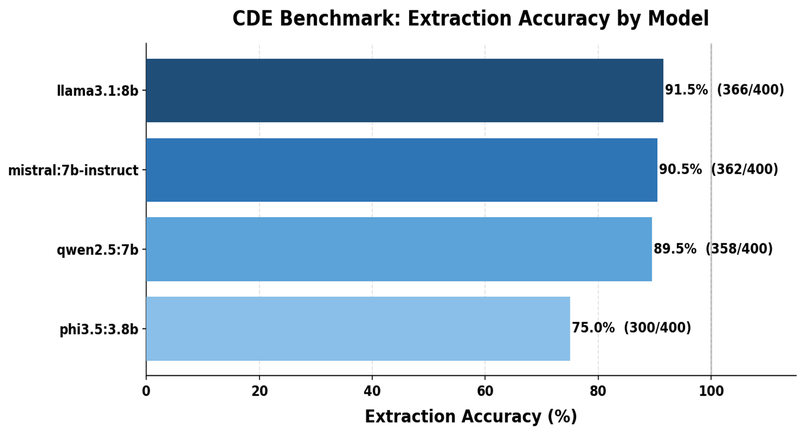


Figure 1. **Extraction Accuracy (%) vs. LLM Model.** Report accuracy tested across multiple open-weight models. The top performing model was Llama 3.1 (8B) at 91.5% and the lowest performing model being Phi 3.5 (3.8B) at 75%.

ASR Verification Results

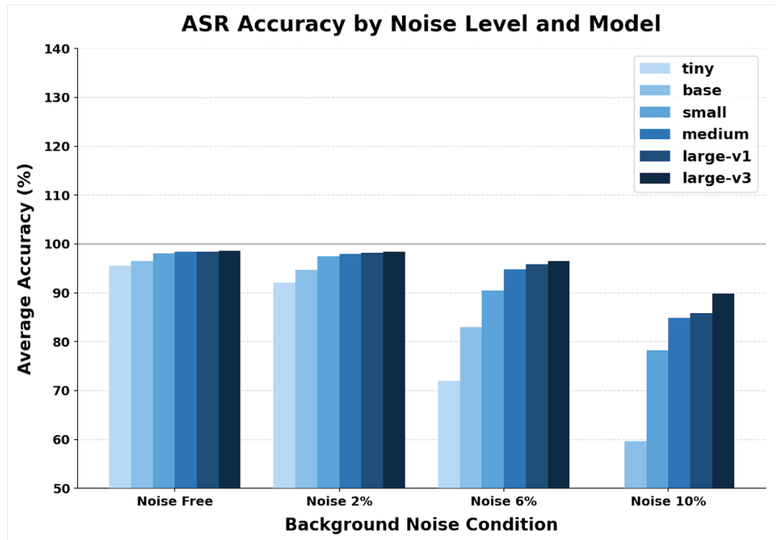


Figure 2. **Average Accuracy (%) vs. ASR Model & Noise Levels.** Transcription accuracy of 6 model sizes of Faster Whisper, a quantized version of OpenAI's Whisper, were tested across multiple levels of white noise. Larger noise percentages decreased overall accuracy of all models, but smaller models suffered proportionally more compared with larger models.

Future Steps

Future work would focus on improving CDE accuracy under realistic OR conditions: surgical background noise, robust Pi and microphone mounting, and integration with existing intraoperative workflows. We would also revisit the Pi/laptop hardware split for cost efficiency while adding compute headroom for higher-fidelity Whisper variants and larger models than the current Llama 3.1 (8B) baseline.

Acknowledgements

We would like to thank Dr. Bradley Greger, Dr. Brent Vernon, Professor Michael Sobrado, and Dr. Trejas Shah for their mentorship and guidance throughout the design process.

QR Code

Scan here for extra resources, including references and a code architecture diagram!