Large Language Models for the Extraction of Structured Information from Spoken Text

Developing a fully functional prototype for nursing documentation to improve efficiency in patient care.

Student

Gioia Mosciatti

Introduction: In long-term care, accurate and structured documentation of patient wellbeing and care processes plays a critical role. Modern nursing documentation software supports caregivers in recording essential information across modules such as vital signs, care reports, and many more. However, enabling a seamless, voice-based input process could offer an additional flexible way to enter information, significantly streamlining documentation in cases where speaking is more efficient than manually entering data. This project explores the use of Large Language Models (LLMs) to extract and structure spoken caregiver observations into documentation fields, in collaboration with Switzerland's leading provider of nursing documentation software.

Approach: Building on this vision, the project aimed to develop a functional prototype for voice-based documentation, focusing on three modules: vital signs, elimination, and care reports. These modules include various field types, such as numerical inputs, free-text entries, and enumerated selection fields.

To evaluate the ability of Large Language Models to transform spoken texts into a given structure, a custom dataset of 28 example texts was created and manually annotated with ground truth labels. Diverse models, prompts, and configurations were tested. Their performance was assessed using precision, recall, and F1-score for field detection, as well as accuracy for value correctness. Additional metrics tracked hallucinated field types and invalid values. Special attention was given to dynamically handling selection fields, which may vary between care institutions.

Further it was examined how such a system could be integrated into existing nursing documentation workflows. OpenAI's Whisper model was tested informally for its transcription capabilities, providing insights for end-to-end integration.

Result: The evaluation revealed that GPT-4.0 performed remarkably well on the task, achieving consistently high results across different output formats, with F1-scores ranging from 0.96 to 0.98 and accuracy scores between 0.92 and 0.95 (Figure 1). Structural aspects of the target schema, such as hierarchical JSON formatting, had minimal impact on results. Allowing null values in enumerated fields led to a notable improvement in model precision and a reduction in hallucinated field entries. Additionally, more detailed prompt instructions led to measurable gains of up to 10% in F1 and 6% in accuracy.

A selection of current open-source models were tested and shown to achieve comparable performance. LLaMa Maverick and Deepseek V2 in particular demonstrated results close to those of GPT-4.0, both in terms of F1 and accuracy scores (see Figure 2). These findings suggest that competitive performance is possible even with non-proprietary models, expanding the options for practical deployment.

Figure 1: Performance of GPT-4.0 across different output formats.
Own presentment

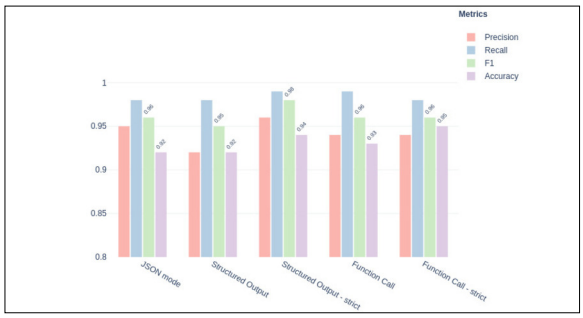
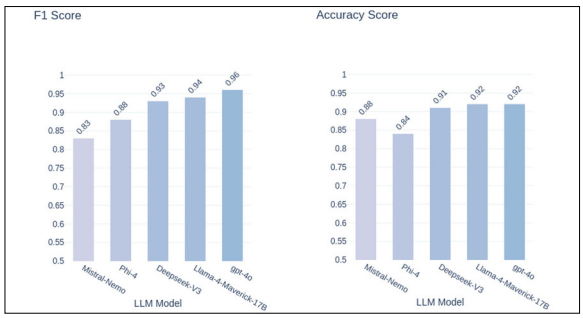


Figure 2: Comparison of performance between alternative open-source language models and GPT-4.0.
Own presentment



Advisor
Prof. Dr. Lin Himmelmann

Subject Area
Data Science

Project Partner
easyDOK AG, Wilen